Foundations of Machine Learning (ST510)

Zoltán Szabó

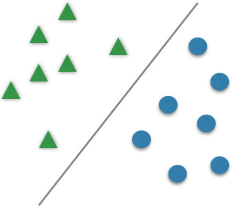
LSE,

Feb. 5, 2024

Motivating examples

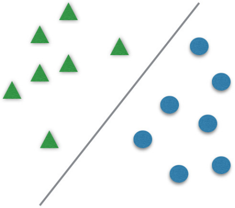
Example-1: non-linear (large-margin) classification

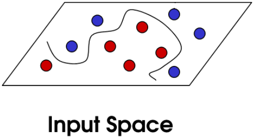
- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.



Example-1: non-linear (large-margin) classification

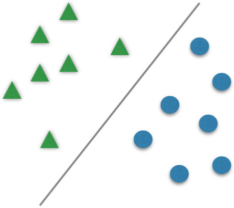
- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.

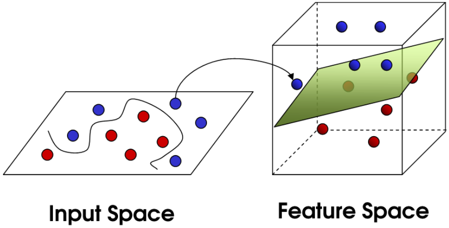




Example-1: non-linear (large-margin) classification

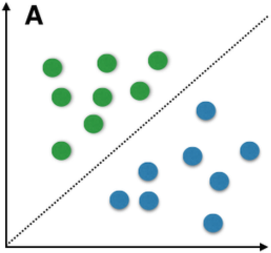
- Given: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $y_i \in \{-1, 1\}$.
- Goal: find an f classifier such that $f(\mathbf{x}) \approx y$.





Example-1: continued – linear separability

Idealized situation

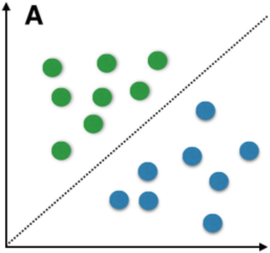


Decision surface:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\}$$

Example-1: continued – linear separability

Idealized situation



Decision surface:

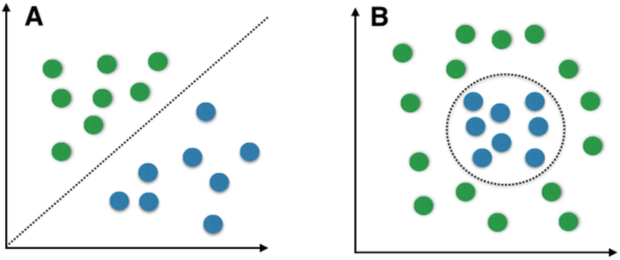
$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\} \;\Rightarrow$$

classes:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle \geq 0\} \qquad \{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle < 0\}$$

Example-1: continued – non-linear separability

Idealized situation   Real world



Decision surface (left):

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle = 0\} \Rightarrow$$

classes:

$$\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle \geq 0\} \qquad \{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle < 0\}\,.$$

Example-1: non-linear separability – continued

On the ellipse

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1 \right\}$$

Example-1: non-linear separability – continued

On the ellipse, outside

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1 \right\},$$

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1 \right\}$$

Example-1: non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1 \right\},$$

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1 \right\},$$

$$\left\{ \mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1 \right\}.$$

Example-1: non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1\right\}.$$

With polynomial feature: $\varphi(\mathbf{x}) = (x_1^2, x_1, 1, x_2^2, x_2)$:

- Decision surface: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle = 0\}$.

Example-1: non-linear separability – continued

On the ellipse, outside, inside:

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} = 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} > 1\right\},$$

$$\left\{\mathbf{x} : \frac{(x_1 - c_1)^2}{a^2} + \frac{(x_2 - c_2)^2}{b^2} < 1\right\}.$$

With polynomial feature: $\varphi(\mathbf{x}) = (x_1^2, x_1, 1, x_2^2, x_2)$:

- Decision surface: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle = 0\}$.
- Classes: $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle > 0\}$, $\{\mathbf{x} : \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle < 0\}$.

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = ?$$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$
\begin{aligned}
\varphi(\mathbf{x}) &= \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right), \\
\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle \\
&= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2 \\
&= \left(x_1x_1' + x_2x_2'\right)^2
\end{aligned}
$$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= \left(x_1x_1' + x_2x_2'\right)^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle \mathbf{x}, \mathbf{x}' \rangle^2 =: k(\mathbf{x}, \mathbf{x}').$$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$
\begin{aligned}
\varphi(\mathbf{x}) &= \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),\\
\langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle\\
&= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2}x_1x_2(x_1')(x_2') + x_2^2(x_2')^2\\
&= \left(x_1x_1' + x_2x_2'\right)^2\\
&= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle\mathbf{x}, \mathbf{x}'\rangle^2 =: k(\mathbf{x}, \mathbf{x}').
\end{aligned}
$$

$\langle\mathbf{x}, \mathbf{x}'\rangle^d = \langle\varphi(\mathbf{x}), \varphi(\mathbf{x}')\rangle$: $\varphi(\mathbf{x})$ = d-order polynomial. $\Rightarrow$

Example-1: quadratic & polynomial features

Still in $\mathbb{R}^2$:

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= (x_1x_1' + x_2x_2')^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \langle \mathbf{x}, \mathbf{x}' \rangle^2 =: k(\mathbf{x}, \mathbf{x}').$$

$\langle \mathbf{x}, \mathbf{x}' \rangle^d = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$: $\varphi(\mathbf{x}) =$ d-order polynomial. $\Rightarrow$ Explicit computation would be heavy!

Example-2: characterizing distributions / independence

- Given: random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim \mathbb{P}_{XY}$.
- Goal: measure the dependence of X and Y.

Example-2: characterizing distributions / independence

- Given: random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim \mathbb{P}_{XY}$.
- Goal: measure the dependence of X and Y.
- Desiderata for a $Q(\mathbb{P}_{XY})$ independence measure [Rényi, 1959]:

  1. $Q(\mathbb{P}_{XY})$ is well-defined,
  2. $Q(\mathbb{P}_{XY}) \in [0, 1]$,
  3. $Q(\mathbb{P}_{XY}) = 0$ iff. $X \perp Y$.
  4. $Q(\mathbb{P}_{XY}) = 1$ iff. $Y = f(X)$ or $X = g(Y)$.

Example-2: continued

- He showed:

$$Q(\mathbb{P}_{XY}) = \sup_{f,g:\ \text{measurable}} \text{corr}(f(X), g(Y)),$$

  satisfies 1-4.

- Too ambitious:
  - computationally intractable.
  - many measurable functions.

Example-2: continued; measurable $\to$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \to \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!

Example-2: continued; measurable $\rightarrow$ continuous

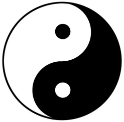
- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \rightarrow \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!
- Idea: to take function spaces
  - dense in $C_b(\mathcal{X})$,
  - computionally tractable.

Example-2: continued; measurable $\to$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \to \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!
- Idea: to take function spaces
  - dense in $C_b(\mathcal{X})$,
  - computionally tractable.

Key: balance

denseness $\to$ universality, computation $\to$ RKHS.



Motivation: kernels = generalized inner product

1. Various data types.
2. RKHS: flexible ($\stackrel{1:1}{\longleftrightarrow}$ probability measures).
3. Still computationally tractable: $k(x_i, x_j) \in \mathbb{R}$.
4. RKHS: Hilbert $\Rightarrow$ statistical analysis.
5. v-RKHS $[k(x, x') \in \mathcal{L}(Y)]$: dependency among output coordinates.

Kernel, RKHS: definition, kernel factory

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

Kernel, RKHS: generalized inner product, -linear methods

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

Constructively, $\mathcal{H}_k = \overline{\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\}}$.

Kernel, RKHS: generalized inner product, -linear methods

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

  Constructively, $\mathcal{H}_k = \overline{\{\sum_{i=1}^n \alpha_i k(\cdot, x_i)\}}$.
- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

Constructively, $\mathcal{H}_k = \overline{\{\sum_{i=1}^n \alpha_i k(\cdot, x_i)\}}$.

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Kernel, RKHS: generalized inner product, -linear methods

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

Constructively, $\mathcal{H}_k = \overline{\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i)\}}$.

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.
- Examples on $\mathbb{R}^d$ ($\gamma > 0$, $p \in \mathbb{Z}^+$): $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$, $k_e(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$.

Some kernel-enriched domains: $(\mathcal{X}, k)$

- Strings [Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- time series [Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- groups and specifically rankings [Cuturi et al., 2005, Jiao and Vert, 2016],
- sets [Haussler, 1999, Gärtner et al., 2002, Balanca and Herbin, 2012, Fellmann et al., 2023], probability distributions [Berlinet and Thomas-Agnan, 2004, Hein and Bousquet, 2005, Smola et al., 2007, Sriperumbudur et al., 2010a],
- various generative models [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- fuzzy domains [Guevara et al., 2017], or
- graphs [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Bai et al., 2020, Borgwardt et al., 2020, Schulz et al., 2022, Nikolentzos and Vazirgiannis, 2023].

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why?

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.

Kernel factory

- We know: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k+\gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geq 0$ ($m = 1, \ldots, M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \ ?$$

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geq 0$ ($m = 1, \ldots, M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y) \rangle_{\mathcal{H}_m}$$

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geq 0$ ($m = 1, \ldots, M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y) \rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}},$$

Kernel factory

- We know: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k + \gamma$: kernel ($\gamma \in \mathbb{R}^{\geq 0}$). Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geq 0$ ($m = 1, \ldots, M$), then

$$\sum_{m=1}^{M} \alpha_m k_m(x, y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y) \rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}},$$
$$\varphi(x) = (\sqrt{\alpha_1}\varphi_1(x), \ldots, \sqrt{\alpha_M}\varphi_M(x)) \in \mathcal{H} := \oplus_{m=1}^{M} \mathcal{H}_m.$$

Kernel factory

- We know: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ is a kernel.
- A few useful rules:
  1. Non-negative shift. k: kernel $\Rightarrow k+\gamma$: kernel $(\gamma \in \mathbb{R}^{\geq 0})$. Why? $\Leftarrow$ Gram.
  2. Cone. If $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\alpha_m \geq 0$ $(m = 1, \ldots, M)$, then

$$\sum_{m=1}^{M} \alpha_m k_m(x,y) = \sum_m \alpha_m \langle \varphi_m(x), \varphi_m(y) \rangle_{\mathcal{H}_m}$$
$$= \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}},$$
$$\varphi(x) = (\sqrt{\alpha_1}\varphi_1(x), \ldots, \sqrt{\alpha_M}\varphi_M(x)) \in \mathcal{H} := \oplus_{m=1}^{M} \mathcal{H}_m.$$

Example: $\oplus_{m=1}^{M} \mathbb{R} = \mathbb{R}^M$.

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^{M}$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^{M} k_m\right)\left(\left(x_1, \ldots, x_M\right), \left(x_1', \ldots, x_M'\right)\right) = \prod_{m=1}^{M} k_m\left(x_m, x_m'\right).$$

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

   - Thus, $(k_m)_{m=1}^M : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ kernels $\Rightarrow \prod_{m=1}^M k_m(x,x')$: kernel on $\mathcal{X}$.

Kernel factory – continued

4. Product. If $(k_m)_{m=1}^M$ are kernels on $\mathcal{X}_m$, then

$$\left(\otimes_{m=1}^M k_m\right)\left(\left(x_1,\ldots,x_M\right),\left(x_1',\ldots,x_M'\right)\right)=\prod_{m=1}^M k_m\left(x_m,x_m'\right).$$

- Thus, $(k_m)_{m=1}^M : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ kernels $\Rightarrow \prod_{m=1}^M k_m(x,x')$: kernel on $\mathcal{X}$.
- Consequence ($\gamma\ge 0$, $p\in\mathbb{Z}^+$):

$$k(\mathbf{x},\mathbf{y})=\left(\langle\mathbf{x},\mathbf{y}\rangle_2+\gamma\right)^p$$

is a kernel.

Kernel factory: product indeed

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$(k_1 \otimes k_2)\left((x,y),(x',y')\right) = k_1(x,x')k_2(y,y')$$

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y') \\ &= \langle \varphi_1(x), \varphi_1(x') \rangle_{\mathcal{H}_1} \langle \varphi_2(y), \varphi_2(y') \rangle_{\mathcal{H}_2}\end{aligned}$$

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}(k_1 \otimes k_2)\left((x, y), (x', y')\right) &= k_1(x, x')k_2(y, y') \\ &= \langle \varphi_1(x), \varphi_1(x') \rangle_{\mathcal{H}_1} \langle \varphi_2(y), \varphi_2(y') \rangle_{\mathcal{H}_2} \\ &= \varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y)\end{aligned}$$

Kernel factory: product indeed

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$\begin{aligned}
(k_1 \otimes k_2)\left((x,y),(x',y')\right) &= k_1(x,x')k_2(y,y') \\
&= \langle \varphi_1(x), \varphi_1(x') \rangle_{\mathcal{H}_1} \langle \varphi_2(y), \varphi_2(y') \rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y) \\
&= \operatorname{tr}\left(\varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y)\right)
\end{aligned}$$

Kernel factory: product indeed

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$
\begin{aligned}
(k_1 \otimes k_2)\left((x, y), (x', y')\right) &= k_1(x, x') k_2(y, y') \\
&= \langle \varphi_1(x), \varphi_1(x') \rangle_{\mathcal{H}_1} \langle \varphi_2(y), \varphi_2(y') \rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y) \\
&= \mathrm{tr}\left(\varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y)\right) \\
&= \mathrm{tr}\left(\varphi_2(y) \varphi_1(x)^\top \varphi_1(x') \varphi_2(y')^\top\right)
\end{aligned}
$$

Kernel factory: product indeed

Let $M = 2$ and assume that $\varphi_m(x) \in \mathbb{R}^{d_m}$:

$$
\begin{aligned}
(k_1 \otimes k_2)\left((x, y), (x', y')\right) &= k_1(x, x')k_2(y, y') \\
&= \langle \varphi_1(x), \varphi_1(x') \rangle_{\mathcal{H}_1} \langle \varphi_2(y), \varphi_2(y') \rangle_{\mathcal{H}_2} \\
&= \varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y) \\
&= \operatorname{tr}\left(\varphi_1(x)^\top \varphi_1(x') \varphi_2\left(y'\right)^\top \varphi_2(y)\right) \\
&= \operatorname{tr}\left(\varphi_2(y)\varphi_1(x)^\top \varphi_1(x') \varphi_2(y')^\top\right) \\
&= \left\langle \underbrace{\varphi_1(x)\varphi_2(y)^\top}_{\in \mathbb{R}^{d_1 \times d_2}}, \underbrace{\varphi_1(x')\varphi_2(y')^\top}_{\in \mathbb{R}^{d_1 \times d_2}} \right\rangle_{\mathsf{F}} ,
\end{aligned}
$$

where $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathsf{F}} = \operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right) = \sqrt{\sum_{ij} A_{ij} B_{ij}}$ is the Frobenius inner product.

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x, x') := \lim_{n\to\infty} k_n(x, x')$$

is a kernel. Why?

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.
Example ($\gamma > 0$):

$$k(\mathbf{x},\mathbf{y}) = e^{\gamma\langle \mathbf{x},\mathbf{y}\rangle_2} = \sum_{n\in\mathbb{N}} \frac{\left(\gamma\langle \mathbf{x},\mathbf{y}\rangle_2\right)^n}{n!}$$

Kernel factory

6. Limit. If $(k_n)_{n\in\mathbb{N}}$ are kernels on $\mathcal{X}$, then

$$k(x,x') := \lim_{n\to\infty} k_n(x,x')$$

is a kernel. Why? $\Leftarrow$ Gram.
Example ($\gamma > 0$):

$$k(\mathbf{x},\mathbf{y}) = e^{\gamma\langle\mathbf{x},\mathbf{y}\rangle_2} = \sum_{n\in\mathbb{N}} \frac{\left(\gamma\langle\mathbf{x},\mathbf{y}\rangle_2\right)^n}{n!}$$

Reason: polynomial kernel & limit rule.

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x, y) = f(x)k(x, y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x, y) = f(x)k(x, y)f(y) = f(x)\langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}} f(y)$$

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\begin{aligned}\tilde{k}(x,y) &= f(x)k(x,y)f(y) = f(x)\langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}} f(y)\\ &= \Big\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y)\Big\rangle_{\mathcal{H}}.\end{aligned}$$

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x,y) = f(x)k(x,y)f(y) = f(x)\langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}} f(y)$$
$$= \Big\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y)\Big\rangle_{\mathcal{H}}.$$

Example (Gaussian kernel, $\gamma > 0$): previous example & new rule

$$k(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}$$

by using $\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle$.

Kernel factory – continued

7. Pre-post multiplication. k kernel on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, then

$$\tilde{k}(x, y) = f(x)k(x, y)f(y)$$

is a kernel. Check (feature view):

$$\tilde{k}(x, y) = f(x)k(x, y)f(y) = f(x)\langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} f(y)$$
$$= \Big\langle \underbrace{f(x)\varphi(x)}_{=:\tilde{\varphi}(x)}, f(y)\varphi(y) \Big\rangle_{\mathcal{H}}.$$

Example (Gaussian kernel, $\gamma > 0$): previous example & new rule

$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2} = e^{-\gamma\|\mathbf{x}\|_2^2} e^{2\gamma\langle \mathbf{x},\mathbf{y}\rangle_2} e^{-\gamma\|\mathbf{y}\|_2^2}$$

by using $\|\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - 2\langle \mathbf{x}, \mathbf{y} \rangle$.

Properties of $\mathcal{H}_k$, computational tractability

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y \in \mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\le} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\le} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\le} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\le} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.
- $k \in C^m \Rightarrow \forall f \in \mathcal{H}_k$ is m-times continuously differentiable.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k: bounded $[\sup_{x,y\in\mathcal{X}} k(x,y) \le C] \Rightarrow \forall f \in \mathcal{H}_k$ is bounded:

$$|f(x)| = \left|\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}\right| \overset{\text{CBS}}{\le} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot,x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k: continuous $\Rightarrow \mathcal{H}_k$: separable $\left[\ell^2(\mathbb{N})\right]$.
- k: bounded and continuous $\Rightarrow \forall f \in \mathcal{H}_k$ is bounded & continuous.
- $k \in C^m \Rightarrow \forall f \in \mathcal{H}_k$ is m-times continuously differentiable.
- k: analytic $\Rightarrow \forall f \in \mathcal{H}_k$ is analytic.

Representer theorem [Schölkopf et al., 2001, Yu et al., 2013]

- Given: $\{(x_i, y_i)\}_{i=1}^n$, say classification/regression.
- Goal:

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|_{\mathcal{H}_k}^2\right) \to \min_{\mathcal{H}_k},$$

r : monotonically increasing.

Representer theorem [Schölkopf et al., 2001, Yu et al., 2013]

- Given: $\{(x_i, y_i)\}_{i=1}^n$, say classification/regression.
- Goal:

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{\mathcal{H}_k},$$

r : monotonically increasing.

- Example:

$$V(\ldots) = \frac{1}{n}\sum_{i=1}^n \max(1 - y_i f(x_i), 0) \quad \text{(soft classification)},$$

$$V(\ldots) = \frac{1}{n}\sum_{i=1}^n [f(x_i) - y_i]^2 \quad \text{(regression)}.$$

Representer theorem – continued

...then

- $\exists$ solution in the form:

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \quad \alpha_i \in \mathbb{R}.$$

- r: strictly increasing $\Rightarrow$ $\forall$ solution is of this form.
- Example: $r(z) = \lambda z$, $\lambda > 0$.

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \dots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{f \in \mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

$$\begin{aligned} S &= \operatorname{span}\left(k(\cdot, x_i), i \in [n]\right), \\ f &= f_S + f_\perp, \\ \|f\|^2_{\mathcal{H}_k} &= \|f_S\|^2_{\mathcal{H}_k} + \underbrace{\|f_\perp\|^2_{\mathcal{H}_k}}_{\geq 0} \geq \|f_S\|^2_{\mathcal{H}_k}. \end{aligned}$$

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{f \in \mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

$$\begin{aligned} S &= \mathrm{span}\left(k(\cdot, x_i), i \in [n]\right), \\ f &= f_S + f_\perp, \\ \|f\|^2_{\mathcal{H}_k} &= \|f_S\|^2_{\mathcal{H}_k} + \underbrace{\|f_\perp\|^2_{\mathcal{H}_k}}_{\geq 0} \geq \|f_S\|^2_{\mathcal{H}_k}. \end{aligned}$$

In J

- 1st term: depends on f_S only, $f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}$.

Representer theorem – proof

Objective

$$J(f) = V(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + r\left(\|f\|^2_{\mathcal{H}_k}\right) \to \min_{f \in \mathcal{H}_k}.$$

Decompose & Pythagorean theorem:

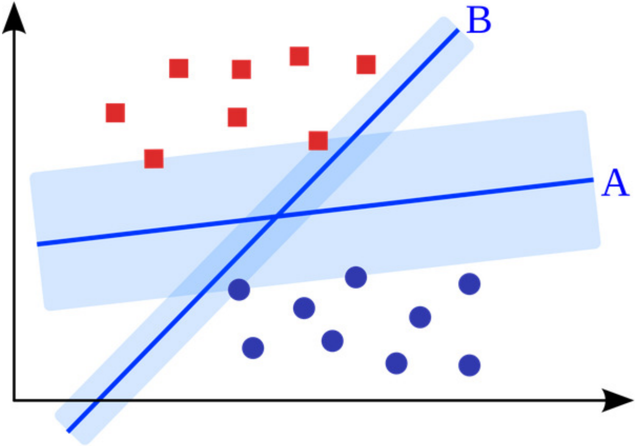
$$\begin{aligned} S &= \operatorname{span}\left(k(\cdot, x_i), i \in [n]\right), \\ f &= f_S + f_\perp, \\ \|f\|^2_{\mathcal{H}_k} &= \|f_S\|^2_{\mathcal{H}_k} + \underbrace{\|f_\perp\|^2_{\mathcal{H}_k}}_{\geq 0} \geq \|f_S\|^2_{\mathcal{H}_k}. \end{aligned}$$

In J

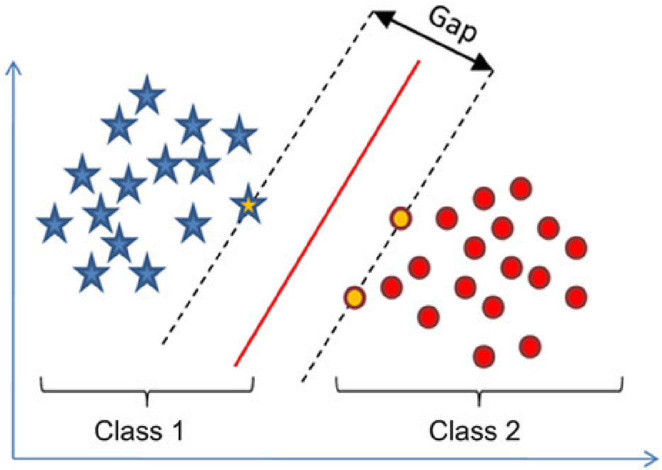
- 1st term: depends on f_S only, $f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}$.
- 2nd term: can only decrease by neglecting $f_\perp$ ($r \nearrow$).

Classification: SVMC

Which separating line is the 'best'?



Answer: the one with the largest margin.



SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$,
  - $\mathbf{w}$: normal vector, b: offset.

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$,
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$,
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min_{\mathbf{w},b} \|\mathbf{w}\|_2^2 \,, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

- Shortly,

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \forall i.$$

SVM formulation: hard classification

- Hyperplane: $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$,
  - $\mathbf{w}$: normal vector, b: offset.
- Goal:

$$\max_{\mathbf{w},b} \underbrace{\frac{2}{\|\mathbf{w}\|_2}}_{\text{margin}} \Leftrightarrow \min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } \underbrace{\begin{cases} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 & \text{if } y_i = 1, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 & \text{otherwise.} \end{cases}}_{\text{correction classification}}$$

- Shortly,

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2, \text{ s.t. } y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \forall i.$$

- Decision: $\hat{y}(\mathbf{x}) = \operatorname{sign}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b \right)$.

SVM formulation: soft classification

- Hard classification objective:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2 \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geq 1, \forall i.$$

There might not be solution! (non-linearly separable case)

SVM formulation: soft classification

- Hard classification objective:

$$\min_{\mathbf{w},b} \|\mathbf{w}\|_2^2 \ \text{ s.t. } \ y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1, \forall i.$$

  There might not be solution! (non-linearly separable case)

- Soft classification objective:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \ \text{ s.t. } \ y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i.$$

  Linear penalty on misclassification.

Note on the soft objective of SVMC

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \xi_i \ \text{ s.t. } \ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1 - \xi_i, \ \xi_i \geq 0, \ \forall i$$

Note on the soft objective of SVMC

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i. \Leftrightarrow$$

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\max\Big(1 - y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right), 0\Big)$$

Note on the soft objective of SVMC

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \text{ s.t. } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i. \Leftrightarrow$$

$$, \min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \underbrace{\max\Big(1 - y_i(\underbrace{\langle \mathbf{w}, \mathbf{x}_i\rangle + b}_{=f(\mathbf{x}_i)}), 0\Big)}_{=:h(y_i f(\mathbf{x}_i))},$$

where $h(u) = \max(1-u, 0)$ is the hinge loss.

Note on the soft objective of SVMC – continued

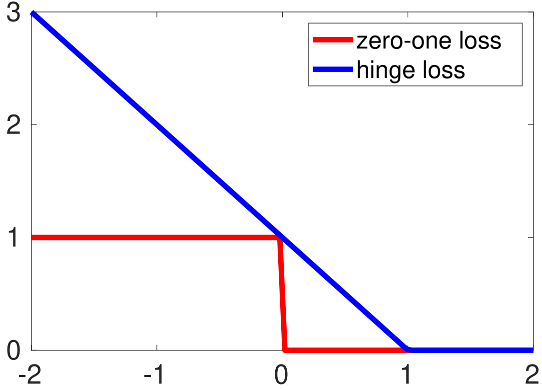
The hinge loss is the convex envelope of the zero-one loss:

$$z(u) = \mathbb{I}_{u<0}, \qquad u = y_i f(x_i),$$
$$h(u) = \max(1-u, 0).$$

Note on the soft objective of SVMC – continued

The hinge loss is the convex envelope of the zero-one loss:

$$z(u) = \mathbb{I}_{u<0}, \qquad u = y_i f(x_i),$$
$$h(u) = \max(1-u, 0).$$



Soft classification objective:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i, \text{ s.t. } y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (\forall i).$$

Lagrangian function: with $\alpha_i \geq 0, \beta_i \geq 0$ $(\forall i)$

$$L(\mathbf{w}, b, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{objective - Lagrangian multipliers} \times \text{conditions}$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i.$$

Solving for $\frac{\partial L}{\partial \text{primal}} = 0$, we get ...

SVM formulation: soft classification

$$L(\mathbf{w}, b, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) =$$

$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \left[y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^n \beta_i \xi_i.$$

Optimality equations:

$$\mathbf{0} = \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad (\mathbf{w} \leftrightarrow \boldsymbol{\alpha}),$$

$$0 = \frac{\partial L}{\partial b} = \sum_{i=1}^n \alpha_i y_i,$$

$$0 = \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i.$$

Plugging these equations back to L, we have ...

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \leq \alpha_i \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \leq \alpha_i \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

- $b \Leftarrow y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1 \Leftarrow \alpha_i > 0$ [complementary slackness].

SVM formulation: soft classification

Dual form:

$$\max_{\boldsymbol{\alpha}} \underbrace{\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{\text{quadratic in } \boldsymbol{\alpha}}, \text{ s.t. } \underbrace{0 \le \alpha_i \le C, \sum_{i=1}^{n} \alpha_i y_i = 0}_{\text{linear in } \boldsymbol{\alpha}}.$$

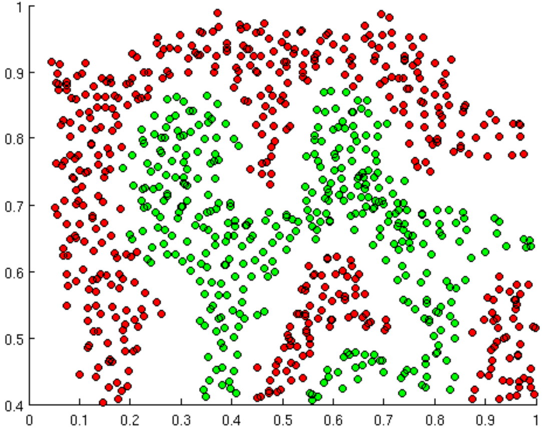
- $b \Leftarrow y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1 \Leftarrow \alpha_i > 0$ [complementary slackness].
- QP: solvers are available.

If linear separability does not hold

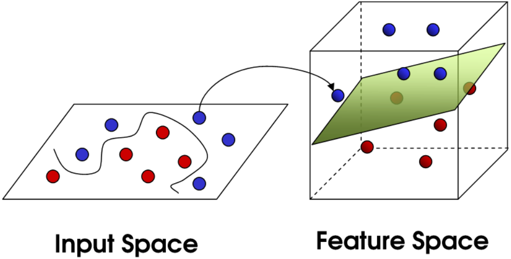
- Until this point:
  - (almost) linearly separable case.

If linear separability does not hold

- Until this point:
    - (almost) linearly separable case.
- Now:



If linear separability does not hold: kernel trick



Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C\ (\forall i).$$

Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

- Nonlinear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

Nonlinear SVM

- Linear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

- Nonlinear SVM (dual):

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

- Nonlinear SVM (primal):

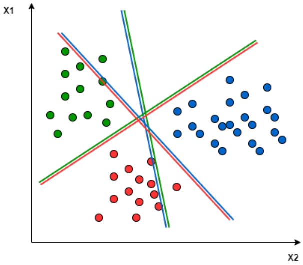
$$\min_{f \in \mathcal{H}_k, \boldsymbol{\xi}} \frac{1}{2} \|f\|_{\mathcal{H}_k}^2 + C \sum_{i=1}^{n} \xi_i, \text{ s.t. } y_i f(x_i) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i.$$

Multiclass (say M) classification with SVM

Idea

Break down the problem to multiple binary classification problems.

1. one-to-one approach:
   - $\frac{M(M-1)}{2}$ SVMC-s, i vs. j ($i \neq j$),
   - on new input x: the class with the most votes is predicted.

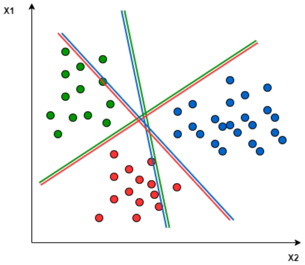


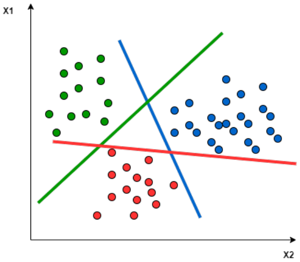
Multiclass (say M) classification with SVM

Idea

Break down the problem to multiple binary classification problems.

1. one-to-one approach:
   - $\frac{M(M-1)}{2}$ SVMC-s, i vs. j ($i \neq j$),
   - on new input x: the class with the most votes is predicted.
2. one-to-rest approach:
   - M SVMC-s, each predicts one class.
   - Classifiers give real-valued confidence scores: $f_m(x)$, $m \in [M]$.
   - Decision: $\hat{m} = \arg\max_{m \in [M]} f_m(x)$.





Parameter selection $\to$ cost: C, kernel parameter: σ

M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
   - training set $(X_{\text{tr}}, Y_{\text{tr}})$: $X_{\text{val},m}$, $Y_{\text{val},m}$, $m \in [M]$.
   - test set: X_{te}, Y_{te}.

Parameter selection $\rightarrow$ cost: C, kernel parameter: σ

M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
    - training set $(X_{\text{tr}}, Y_{\text{tr}})$: $X_{\text{val},m}$, $Y_{\text{val},m}$, $m \in [M]$.
    - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error ($m \in [M]$) while
    - trained on: $X_{\text{tr}} \backslash X_{\text{val},m}$, $Y_{\text{tr}} \backslash Y_{\text{val},m}$,
    - tested on: $X_{\text{val},m}$, $Y_{\text{val},m}$.

Parameter selection $\to$ cost: C, kernel parameter: σ

M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
   - training set $(X_{\text{tr}}, Y_{\text{tr}})$: $X_{\text{val},m}$, $Y_{\text{val},m}$, $m \in [M]$.
   - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error ($m \in [M]$) while
   - trained on: $X_{\text{tr}} \backslash X_{\text{val},m}$, $Y_{\text{tr}} \backslash Y_{\text{val},m}$,
   - tested on: $X_{\text{val},m}$, $Y_{\text{val},m}$.
3. $\theta^* :=$ minimizer of CV error.

Parameter selection $\rightarrow$ cost: C, kernel parameter: σ

M-fold cross-validation $[\theta := (C, \sigma)]$:

1. Split data:
   - training set $(X_{\text{tr}}, Y_{\text{tr}})$: $X_{\text{val},m}$, $Y_{\text{val},m}$, $m \in [M]$.
   - test set: X_{te}, Y_{te}.
2. For fixed θ: evaluate the average error $(m \in [M])$ while
   - trained on: $X_{\text{tr}} \backslash X_{\text{val},m}$, $Y_{\text{tr}} \backslash Y_{\text{val},m}$,
   - tested on: $X_{\text{val},m}$, $Y_{\text{val},m}$.
3. $\theta^* :=$ minimizer of CV error.
4. Report: performance of θ^* on X_{te}, Y_{te}.

Regression: kernel ridge regression

Kernel ridge regression (KRR)

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

Kernel ridge regression (KRR)

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

- Analytical solution:

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)]\,(\mathbf{G} + \lambda n \mathbf{I}_n)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n .$$

Kernel ridge regression (KRR)

- Given: $\{(x_i, y_i)\}_{i=1}^n$, $\mathcal{H} := \mathcal{H}_k$, $y_i \in \mathbb{R}$.
- Task ($\lambda > 0$):

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \|f\|_{\mathcal{H}}^2 \to \min_{f \in \mathcal{H}}.$$

- Analytical solution:

$$f(x) = [k(x_1, x), \ldots, k(x_n, x)] (\mathbf{G} + \lambda n \mathbf{I}_n)^{-1}[y_1; \ldots; y_n],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n.$$

Question

How do we get this solution?

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad (a_i \in \mathbb{R}).$$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad (a_i \in \mathbb{R}).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^\top \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{G}\mathbf{a} + \mathbf{a}^\top \left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right]\mathbf{a}. \end{aligned}$$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad (a_i \in \mathbb{R}).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^\top \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{G}\mathbf{a} + \mathbf{a}^\top \left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right] \mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I}_n)^{-1} \mathbf{y}$

Kernel ridge regression

By the representer theorem

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i), \quad (a_i \in \mathbb{R}).$$

Multiplying the objective by n, using the reproducing property:

$$\begin{aligned} J(f) &= \sum_{j=1}^{n} [y_j - \langle f, k(\cdot, x_j) \rangle_{\mathcal{H}}]^2 + \lambda n \|f\|_{\mathcal{H}}^2 \\ &= \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + (\lambda n)\mathbf{a}^\top \mathbf{G}\mathbf{a} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{G}\mathbf{a} + \mathbf{a}^\top \left[\mathbf{G}^2 + (\lambda n)\mathbf{G}\right] \mathbf{a}. \end{aligned}$$

Solving $\mathbf{0} = \frac{\partial J}{\partial \mathbf{a}}$, one gets $\mathbf{a}^* = (\mathbf{G} + \lambda n \mathbf{I}_n)^{-1} \mathbf{y}$ by

$$\frac{\partial \mathbf{a}^\top \mathbf{B} \mathbf{a}}{\partial \mathbf{a}} = \left(\mathbf{B} + \mathbf{B}^\top\right) \mathbf{a}, \ \frac{\partial \mathbf{c}^\top \mathbf{a}}{\partial \mathbf{a}} = \mathbf{c}.$$

Kernel machines: a simple algorithm = SGD [Kivinen et al., 2004]

Empirical regularized risk:

$$\min_{f \in \mathcal{H}_k} J(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_k}^2 \quad (\lambda > 0).$$

Kernel machines: a simple algorithm = SGD [Kivinen et al., 2004]

Empirical regularized risk:

$$\min_{f \in \mathcal{H}_k} J(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_k}^2 \quad (\lambda > 0).$$

Instantaneous regularized risk:

$$J_{\text{inst}}(f, (x, y)) = \ell(f(x), y) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_k}^2 .$$

Kernel machines: a simple algorithm = SGD [Kivinen et al., 2004]

Empirical regularized risk:

$$\min_{f\in\mathcal{H}_k} J(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_k} \quad (\lambda > 0).$$

Instantaneous regularized risk:

$$J_{\text{inst}}(f, (x, y)) = \ell(f(x), y) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_k}.$$

Update (learning rate: $\eta_t > 0$):

$$f_{t+1} = f_t - \eta_t \underbrace{\left.\frac{\partial J_{\text{inst}}(f, (x_t, y_t))}{\partial f}\right|_{f=f_t}}_{\left.\frac{\partial \ell(z,y)}{\partial z}\right|_{z=f_t(x_t), y=y_t} k(\cdot, x_t) + \lambda f_t}.$$

Note: if ℓ is non-differentiable, subgradient is taken.

SGD implementation

1. Initialization: $f_1 = 0$.
2. By the representer theorem:

$$f_t = \sum_{i=1}^{t-1} \alpha_i k(\cdot, x_i), \quad \alpha_i \in \mathbb{R}.$$

SGD implementation

1. Initialization: $f_1 = 0$.
2. By the representer theorem:

$$f_t = \sum_{i=1}^{t-1} \alpha_i k(\cdot, x_i), \quad \alpha_i \in \mathbb{R}.$$

3. Update (from the previous slide):

$$\begin{aligned} f_{t+1} &= f_t - \eta_t \left[\ell'(f_t(x_t), y_t) k(\cdot, x_t) + \lambda f_t\right] \\ &= (1 - \eta_t \lambda) f_t - \eta_t \ell'(f_t(x_t), y_t) k(\cdot, x_t). \end{aligned}$$

SGD implementation

1. Initialization: $f_1 = 0$.
2. By the representer theorem:

$$f_t = \sum_{i=1}^{t-1} \alpha_i k(\cdot, x_i), \quad \alpha_i \in \mathbb{R}.$$

3. Update (from the previous slide):

$$\begin{aligned} f_{t+1} &= f_t - \eta_t \left[\ell'(f_t(x_t), y_t) k(\cdot, x_t) + \lambda f_t\right] \\ &= (1 - \eta_t \lambda) f_t - \eta_t \ell'(f_t(x_t), y_t) k(\cdot, x_t). \end{aligned}$$

4. $\Leftrightarrow$ Update (in terms of coefficients): For $i \in [t]$,

$$\alpha_i := \begin{cases} -\eta_t \ell'(f_t(x_t), y_t) & \text{if } i = t, \\ (1 - \eta_t \lambda) \alpha_i & \text{if } i < t. \end{cases}$$

1. Recall the dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

   This is well-adapted to CD methods (α_i; [Hsieh et al., 2008]).

Optimizers+

1. Recall the dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

   This is well-adapted to CD methods (α_i; [Hsieh et al., 2008]).

2. For KRR:
   - scaling to billions of points [Meanti et al., 2020],
   - idea: Nyström method + pre-conditioned conjugate gradient solver + GPU.

Optimizers+

1. Recall the dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j), \text{ s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\ (\forall i).$$

   This is well-adapted to CD methods (α_i; [Hsieh et al., 2008]).

2. For KRR:
   - scaling to billions of points [Meanti et al., 2020],
   - idea: Nyström method + pre-conditioned conjugate gradient solver + GPU.
3. For large-scale classification (+recent survey), see [Tanji et al., 2023]:
   - Nyström technique + accelerated stochastic subgradient descent.

Maximal correlation: KCCA

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- KCCA measure of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$$\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(X), g(Y)),$$

$$\operatorname{corr}(f(X), g(Y)) = \frac{\operatorname{cov}(f(X), g(Y))}{\sqrt{\operatorname{var}[f(X)] \operatorname{var}[g(Y)]}}.$$

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (X, Y).

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (X, Y).

Recall the reproducing property

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} \quad \forall f \in \mathcal{H}_k, x \in \mathcal{X}.$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N} \varphi(x_i) \right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N} \psi(y_i) \right\rangle_{\mathcal{H}_\ell}}$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i) \right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i) \right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\mathrm{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]^2$$

KCCA: empirical estimate

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]^2 = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2,$$

KCCA: empirical estimate

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)\right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2 = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle^2_{\mathcal{H}_k},$$

$$\widehat{\text{var}}[g(Y)] = \frac{1}{N}\sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n)\rangle^2_{\mathcal{H}_\ell}.$$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$\mathrm{span}\left(\{\tilde{\varphi}(x_n)\}_{n=1}^N\right) = \left\{\sum_{n=1}^N c_n \tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$\mathrm{span}\left(\{\tilde{\varphi}(x_n)\}_{n=1}^N\right) = \left\{\sum_{n=1}^N c_n\tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

Key idea

Enough to consider $f = \sum_{i=1}^N c_i\tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i\tilde{\psi}(y_i)$.

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k}$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n)$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^\top \tilde{\mathbf{G}}_X)_n,$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^N c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^\top \tilde{\mathbf{G}}_X)_n,$$

$$\langle g, \tilde{\psi}(y_n) \rangle_{\mathcal{H}_\ell} = (\mathbf{d}^\top \tilde{\mathbf{G}}_Y)_n,$$

with the centered kernels ($\tilde{k}$, $\tilde{\ell}$) and Gram matrices ($\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$).

Until now

All the objective terms can be expressed by $\mathbf{c}$, $\mathbf{d}$, $\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$.

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

$$\widehat{\mathrm{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2, \widehat{\mathrm{var}}[g(Y)] = \frac{1}{N}\sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}^2,$$

and we have

$$\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} = (\mathbf{c}^\top \tilde{\mathbf{G}}_X)_n, \qquad \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^\top \tilde{\mathbf{G}}_Y)_n.$$

KCCA: empirical estimate

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2, \widehat{\text{var}}[g(Y)] = \frac{1}{N}\sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}^2,$$

and we have

$$\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} = (\mathbf{c}^\top \tilde{\mathbf{G}}_X)_n, \qquad \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^\top \tilde{\mathbf{G}}_Y)_n.$$

Thus,

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d},$$

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\mathbf{c}^\top (\tilde{\mathbf{G}}_X)^2 \mathbf{c}, \quad \widehat{\text{var}}[g(Y)] = \frac{1}{N}\mathbf{d}^\top (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\mathrm{KCCA}}}^{\mathrm{temp}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^\top (\tilde{\mathbf{G}}_X)^2 \mathbf{c}} \sqrt{\mathbf{d}^\top (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^\top (\tilde{\mathbf{G}}_X)^2 \mathbf{c}} \sqrt{\mathbf{d}^\top (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{\text{KCCA}}}(X, Y) := \widehat{\rho_{\text{KCCA}}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa)$$

$$= \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^\top \left(\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N\right)^2 \mathbf{c}} \sqrt{\mathbf{d}^\top \left(\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N\right)^2 \mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^\top (\tilde{\mathbf{G}}_X)^2 \mathbf{c}} \sqrt{\mathbf{d}^\top (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{\text{KCCA}}}(X, Y) := \widehat{\rho_{\text{KCCA}}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa)$$

$$= \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^\top \left(\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N\right)^2 \mathbf{c}} \sqrt{\mathbf{d}^\top \left(\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N\right)^2 \mathbf{d}}}.$$

Question

How do we solve it?

KCCA: solution

Stationary points of $\widehat{\rho_{\mathrm{KCCA}}}(X, Y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\mathrm{KCCA}}}(X, Y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\mathrm{KCCA}}}(X, Y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d} = \frac{(\mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d})(\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \mathbf{c}}{\mathbf{c}^\top (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \mathbf{c}}, \; \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X \mathbf{c} = \frac{(\mathbf{d}^\top \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X \mathbf{c})(\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \mathbf{d}}{\mathbf{d}^\top (\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \mathbf{d}}$$

KCCA: solution

Stationary points of $\widehat{\rho_{\text{KCCA}}}(X,Y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X,Y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X,Y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} = \frac{(\mathbf{c}^\top\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d})(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N)^2\mathbf{c}}{\mathbf{c}^\top(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N)^2\mathbf{c}}, \quad \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X\mathbf{c} = \frac{(\mathbf{d}^\top\tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X\mathbf{c})(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N)^2\mathbf{d}}{\mathbf{d}^\top(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N)^2\mathbf{d}}$$

Normalization:

- $(\mathbf{c},\mathbf{d})$: solution $\Rightarrow$ $(a\mathbf{c}, b\mathbf{d})$: solution $a,b \in \mathbb{R}\backslash\{0\}$.
- denominators $:= 1$.

KCCA: final task

Find the maximal eigenvalue, $\lambda := \mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}$, of the generalized eigenvalue problem:

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$$\mathbf{A}\mathbf{z} = \lambda \mathbf{B}\mathbf{z}.$$

Note: Python implementation in the ITE toolbox ($M \geq 2$, with acceleration).

Summary

- Kernel, RKHS: generalized inner product, - linear methods.
- Computational tractability: representer theorem.
- Classification: SVMC.
- Regression: kernel ridge regression.
- Maximal correlation: KCCA.

Summary

- Kernel, RKHS: generalized inner product, - linear methods.
- Computational tractability: representer theorem.
- Classification: SVMC.
- Regression: kernel ridge regression.
- Maximal correlation: KCCA.

Contents (KCCA: questions)

1. Is KCCA an independence measure? $(\Leftarrow$ universality$)$
2. Meaning/handling of the regularization (κ).
3. $M \geq 2$ components.
4. Computation of $\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$.

Q1 (independence measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005].

Q1 (independence measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005].
- Enough: universal kernel on a compact metric domain.

Q1 (independence measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005].
- Enough: universal kernel on a compact metric domain.
- Example ($\gamma > 0$):
  - Gaussian: $k(\mathbf{x}, \mathbf{x}') = e^{-\gamma \|\mathbf{x}-\mathbf{x}'\|_2^2}$.
  - Laplacian kernel: $k(\mathbf{x}, \mathbf{x}') = e^{-\gamma \|\mathbf{x}-\mathbf{x}'\|_2}$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow$ $C(\mathcal{X}) = C_b(\mathcal{X})$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow k$: bounded.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow k$: bounded.
- k: continuous, bounded $\Rightarrow \mathcal{H}_k \subset C(\mathcal{X})$ [Steinwart and Christmann, 2008].

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow k$: bounded.
- k: continuous, bounded $\Rightarrow \mathcal{H}_k \subset C(\mathcal{X})$ [Steinwart and Christmann, 2008].
- Extensions of c-universality to non-compact spaces:
  - c_0-universality, cc-universality, … [Carmeli et al., 2010, Sriperumbudur et al., 2010b, Simon-Gabriel and Schölkopf, 2018].

Q1: properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.

Q1: properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.

Q1: properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_k}$$

  is a metric.

Q1: properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_k}$$

  is a metric.
- The normalized kernel (recall: corr)

$$\tilde{k}(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

  is universal.

Q1: universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^\infty \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r), \; r \in (0, \infty].$$

Q1: universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^{\infty} \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r), \; r \in (0, \infty].$$

- If $a_n > 0 \; \forall n$, then

$$k(\mathbf{x}, \mathbf{y}) = f(\langle \mathbf{x}, \mathbf{y} \rangle)$$

is universal on $\mathcal{X} := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \sqrt{r} \right\}$.

Q1: universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha \langle \mathbf{x}, \mathbf{y} \rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.

Q1: universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha\langle \mathbf{x}, \mathbf{y}\rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.
- $k(\mathbf{x}, \mathbf{y}) = e^{-\alpha\|\mathbf{x}-\mathbf{y}\|_2^2}$: exp. kernel & normalization.

Q1: universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = (1 - \langle \mathbf{x}, \mathbf{y} \rangle)^{-\alpha}$ binomial kernel
  - on $\mathcal{X}$ compact $\subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < 1\}$.
  - $f(t) = (1 - t)^{-\alpha} = \sum_{n=0}^{\infty} \underbrace{\binom{-\alpha}{n} (-1)^n}_{>0} t^n \quad (|t| < 1)$,

where $\binom{b}{n} = \sum_{i=1}^{n} \frac{b-i+1}{i}$.

Contents

Q2 (κ)

In fact, we estimated

$$\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(X), g(Y); \kappa),$$

$$\operatorname{corr}(f(X), g(Y); \kappa) = \frac{\operatorname{cov}(f(X), g(Y))}{\sqrt{\operatorname{var}[f(X)] + \kappa \|f\|_{\mathcal{H}_k}^2} \sqrt{\operatorname{var}[g(Y)] + \kappa \|g\|_{\mathcal{H}_\ell}^2}}.$$

Q2 (κ)

In fact, we estimated

$$\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f\in\mathcal{H}_k, g\in\mathcal{H}_\ell} \operatorname{corr}(f(X), g(Y); \kappa),$$

$$\operatorname{corr}(f(X), g(Y); \kappa) = \frac{\operatorname{cov}(f(X), g(Y))}{\sqrt{\operatorname{var}[f(X)] + \kappa \|f\|^2_{\mathcal{H}_k}}\sqrt{\operatorname{var}[g(Y)] + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

For consistent KCCA estimate:

- $\kappa_N \to 0$ [Leurgans et al., 1993](spline-RKHS), [Fukumizu et al., 2007] (general RKHS).
- analysis: covariance operators.

Contents

Q3 ($M \geq 2$): symmetry, other form

For

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^\top \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

Q3 ($M \geq 2$): symmetry, other form

For

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^\top \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

Adding the r.h.s. to both sides:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_\mathbf{Y} + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1 + \lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

with eigenvalues $\{1 + \lambda_1, 1 - \lambda_1, \ldots, 1 + \lambda_N, 1 - \lambda_N\}$.

Q3 $(M \geq 2)$

2-variables $[(X, Y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{X}} + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_{\mathbf{Y}} + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1 + \lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

Q3 $(M \geq 2)$

2-variables $[(X, Y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{X}} + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_{\mathbf{Y}} + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

For M-variables (pairwise dependence):

$$\begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_1 \tilde{\mathbf{G}}_2 & \dots & \tilde{\mathbf{G}}_1 \tilde{\mathbf{G}}_M \\ \tilde{\mathbf{G}}_2 \tilde{\mathbf{G}}_1 & (\tilde{\mathbf{G}}_2 + \kappa \mathbf{I}_N)^2 & \dots & \tilde{\mathbf{G}}_2 \tilde{\mathbf{G}}_M \\ \vdots & \vdots & & \vdots \\ \tilde{\mathbf{G}}_M \tilde{\mathbf{G}}_1 & \tilde{\mathbf{G}}_M \tilde{\mathbf{G}}_2 & \dots & (\tilde{\mathbf{G}}_M + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix} =$$

$$\gamma \begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa \mathbf{I}_N)^2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_2 + \kappa \mathbf{I}_N)^2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & \dots & (\tilde{\mathbf{G}}_M + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix}.$$

Q4: Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$(\tilde{\mathbf{G}}_X)_{ij} = \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k}$$

Q4: Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k}
\end{aligned}
$$

Q4: Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N \times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_X)_{nm}
\end{aligned}
$$

Q4: Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N \times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_X)_{nm} \\
&= \left(\mathbf{G}_X - \mathbf{G}_X \frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_X - \frac{\mathbf{E}_N}{N}\mathbf{G}_X\frac{\mathbf{E}_N}{N} \right)_{ij},
\end{aligned}
$$

Q4: Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \left\langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_X)_{nm} \\
&= \left(\mathbf{G}_X - \mathbf{G}_X \frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_X - \frac{\mathbf{E}_N}{N}\mathbf{G}_X \frac{\mathbf{E}_N}{N} \right)_{ij}, \\
&= (\mathbf{H}\mathbf{G}_X\mathbf{H})_{ij},
\end{aligned}
$$

$\mathbf{H}$: symmetric ($\mathbf{H} = \mathbf{H}^\top$), idempotent ($\mathbf{H}^2 = \mathbf{H}$). Contents

Bach, F. and Jordan, M. (2002).
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48.

Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E. (2020).
Local-global nested graph kernels using nested complexity traces.
Pattern Recognition Letters, 134:87–95.

Balanca, P. and Herbin, E. (2012).
A set-indexed Ornstein-Uhlenbeck process.
Electronic Communications in Probability, 17:1–14.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020).

Graph kernels: State-of-the-art and future challenges.
Foundations and Trends in Machine Learning, 13(5-6):531–712.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.
In *International Conference on Data Mining (ICDM)*, pages 74–81.

Carmeli, C., Vito, E. D., Toigo, A., and Umanitá, V. (2010).
Vector valued reproducing kernel Hilbert spaces and universality.
Analysis and Applications, 8:19–61.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).
Fast global alignment kernels.

In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Cuturi, M. and Vert, J.-P. (2005).
The context-tree kernel for strings.
Neural Networks, 18(8):1111–1123.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).
A kernel for time series based on global alignments.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

Fellmann, N., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., and Sinoquet, D. (2023).
Kernel-based sensivity analysis for (excursion) sets.
Technical report.
(https://arxiv.org/abs/2305.09268).

Fukumizu, K., Bach, F. R., and Gretton, A. (2007).
Statistical consistency of kernel canonical correlation analysis.
Journal of Machine Learning Research, 8(14):361–383.

Gärtner, T., Flach, P., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gärtner, T., Flach, P., and Wrobel, S. (2003).
On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6(70):2075–2129.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.

In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S., and Sundararajan, S. (2008).
A dual coordinate descent methhod for large-scale linear SVM.
In *International Conference on Machine Learning (ICML)*, pages 408–415.

Jaakkola, T. S. and Haussler, D. (1999).
Exploiting generative models in discriminative classifiers.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.
In *International Conference on Machine Learning (ICML)*, volume 37, pages 2982–2990.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003).

Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*, pages 321–328.

Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.
Journal of Machine Learning Research, 20:1–45.

Kivinen, J., Smola, A., and Williamson, R. (2004).
Online learning with kernels.
IEEE Transactions on Signal Processing, 52(8):2165–2176.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kondor, R. I. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete input.
In *International Conference on Machine Learning (ICML)*, pages 315–322.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.
Journal of Bioinformatics and Computational Biology, 13(4):527–550.

Leslie, C., Eskin, E., and Noble, W. S. (2002).
The spectrum kernel: A string kernel for SVM protein classification.
Biocomputing, pages 564–575.

Leslie, C. and Kuang, R. (2004).
Fast string kernels using inexact matching for protein sequences.
Journal of Machine Learning Research, 5:1435–1455.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993).
Canonical correlation analysis when the data are curves.

Journal of the Royal Statistical Society, Series B (Methodological), 55(3):725–740.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. (2020).
Kernel methods through the roof: handling billions of points efficiently.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14410–14422.

Nikolentzos, G. and Vazirgiannis, M. (2023).
Graph alignment kernels using Weisfeiler and Leman hierarchies.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2019–2034.

Rényi, A. (1959).

On measures of dependence.
Acta Mathematatica Academiae Scientiarum Hungaricae,
10:441–451.

Rüping, S. (2001).
SVM kernels for time series analysis.
Technical report, University of Dortmund.
(http://www.stefan-rueping.de/publications/rueping-2001-a.pdf).

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In *Conference on Learning Theory (COLT)*, pages 416–426.

Schulz, T. H., Welke, P., and Wrobel, S. (2022).
Graph filtration kernels.

In *AAAI Conference on Artifical Intelligence (AAAI)*, pages 8196–8203.

Seeger, M. (2002).
Covariance kernels from Bayesian generative models.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018).
Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions.
Journal of Machine Learning Research, 44:1–29.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.

In *Algorithmic Learning Theory (ALT)*, pages 13–31.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010a).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2010b).
On the relation between universality, characteristic kernels and RKHS embedding of measures.
In *International Conference on AI and Statistics (AISTATS)*, pages 781–788.

Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 6(3):67–93.

Steinwart, I. and Christmann, A. (2008).

Support Vector Machines.
Springer.

Tanji, S., Vecchia, A. D., Glineur, F., and Villa, S. (2023).
Snacks: a fast large-scale kernel SVM solver.
In *European Control Conference (ECC)*, pages 1–6.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

Watkins, C. (1999).
Dynamic alignment kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Yu, Y., Cheng, H., Schuurmans, D., and Szepesvári, C. (2013).
Characterizing the representer theorem.
In *International Conference on Machine Learning (ICML; PMLR)*, volume 28, pages 570–578.