Learning Kernels - Tutorial

Part I: Introduction to Kernel Methods.

Corinna Cortes
Google Research
corinna@google.com

Mehryar Mohri
Courant Institute &
Google Research
mohri@cims.nyu.edu

Afshin Rostami
UC Berkeley
arostami@eecs.berkeley.edu

Outline

- Part I: Introduction to kernel methods.
- Part II: Learning kernel algorithms.
- Part III: Theoretical guarantees.
- Part IV: Software tools.

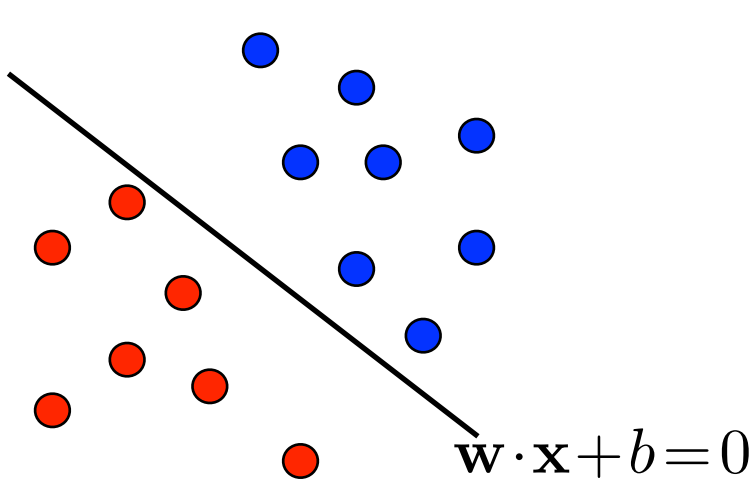
Binary Classification Problem

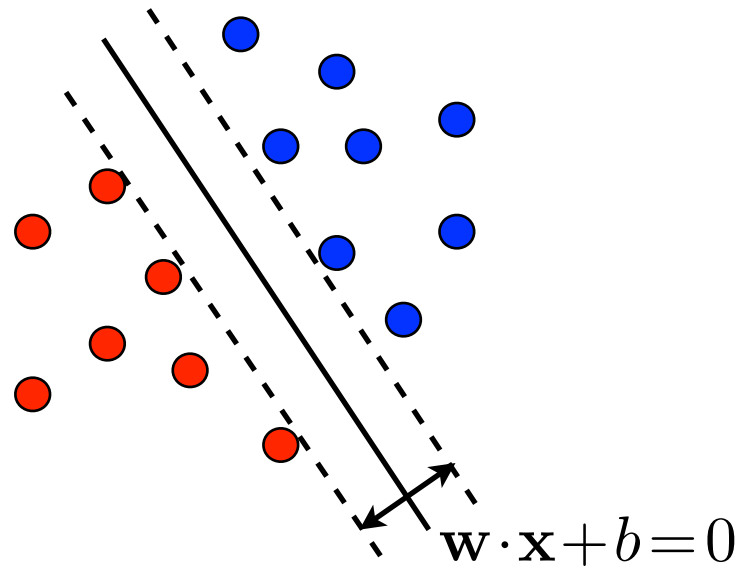
- Training data: sample drawn i.i.d. from set $X \subseteq \mathbb{R}^N$ according to some distribution D,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in X \times \{-1, +1\}.$$

- Problem: find hypothesis $h : X \mapsto \{-1, +1\}$ in H (classifier) with small generalization error $R_D(h)$.
- Linear classification:
  - Hypotheses based on hyperplanes.
  - Linear separation in high-dimensional space.

Linear Separation

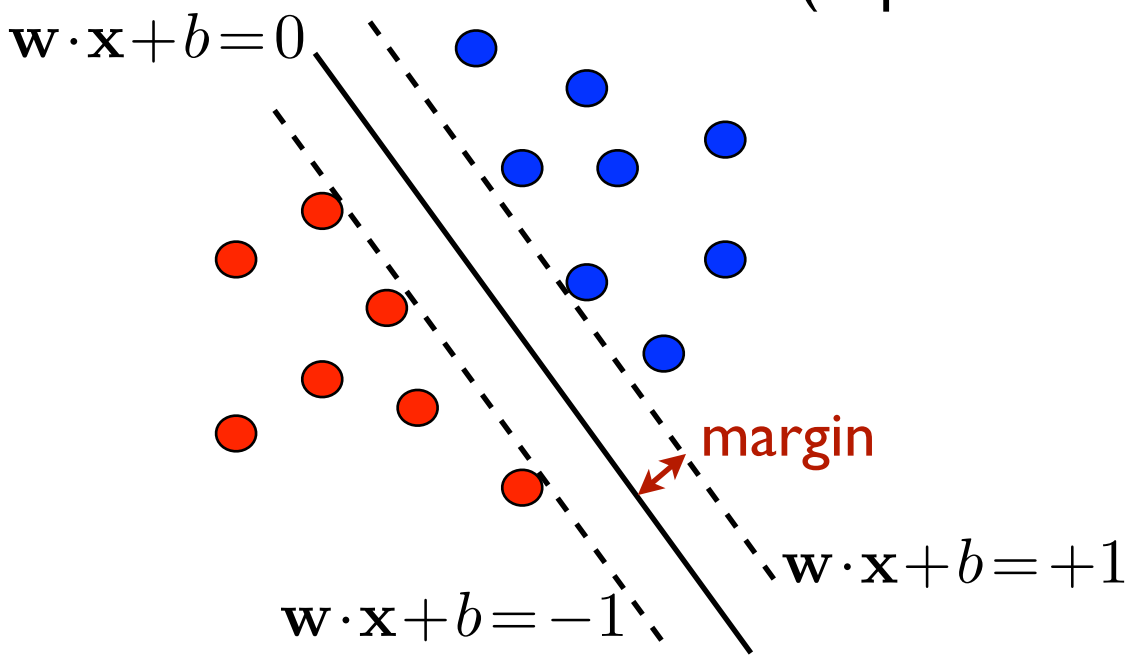




- Classifiers: $H=\{\mathbf{x}\mapsto\operatorname{sgn}(\mathbf{w}\cdot\mathbf{x}+b):\mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}\}$.

Optimal Hyperplane: Max. Margin

(Vapnik and Chervonenkis, 1964)



- Canonical hyperplane: $\mathbf{w}$ and b chosen such that for closest points $|\mathbf{w}\cdot\mathbf{x}+b|=1$.
- Margin: $\rho=\min_{\mathbf{x}\in S}\frac{|\mathbf{w}\cdot\mathbf{x}+b|}{\|\mathbf{w}\|}=\frac{1}{\|\mathbf{w}\|}$.

Optimization Problem

- Constrained optimization:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, i \in [1, m].$$

- Properties:
  - Convex optimization (strictly convex).
  - Unique solution for linearly separable sample.

Support Vector Machines

(Cortes and Vapnik, 1995)

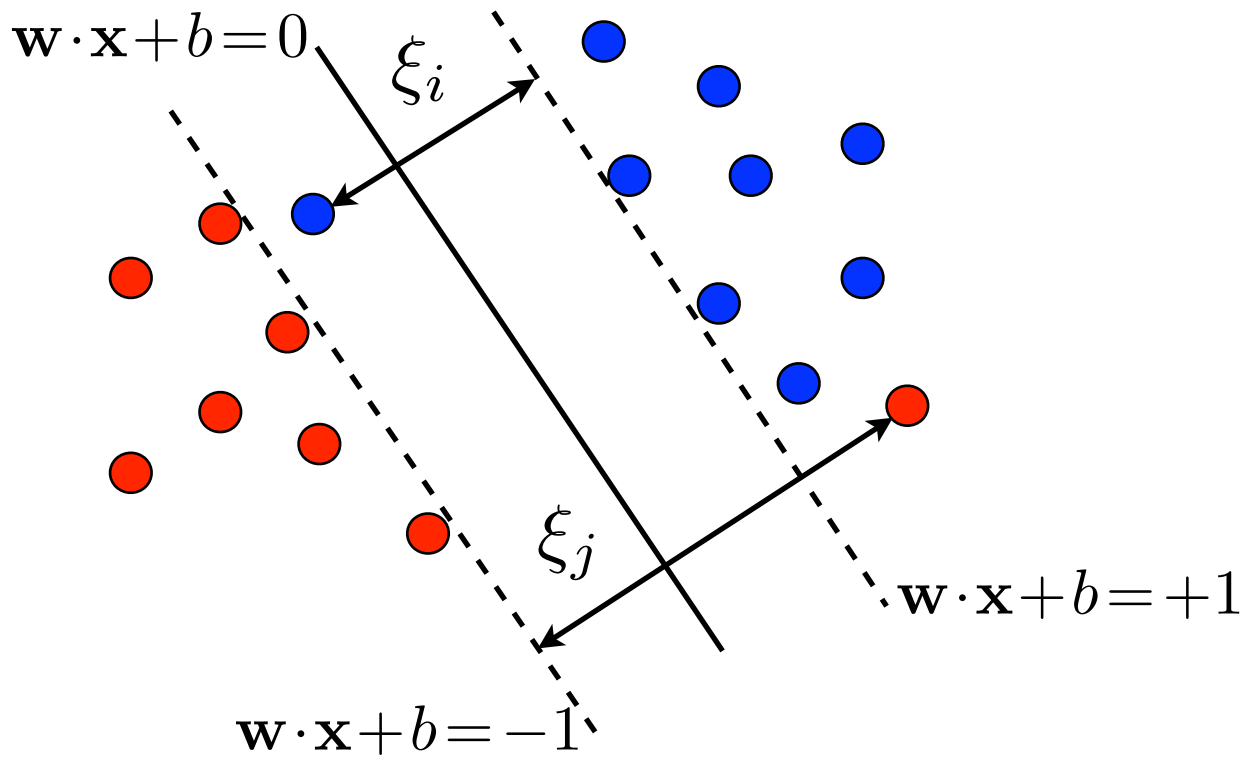
- Problem: data often not linearly separable in practice. For any hyperplane, there exists $\mathbf{x}_i$ such that

$$y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] \not\geq 1.$$

- Idea: relax constraints using slack variables $\xi_i \geq 0$

$$y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] \geq 1 - \xi_i.$$

Soft-Margin Hyperplanes



- Support vectors: points along the margin or outliers.
- Soft margin: $\rho = 1/\|\mathbf{w}\|$.

Optimization Problem

(Cortes and Vapnik, 1995)

- Constrained optimization:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \quad \wedge \quad \xi_i \geq 0, i \in [1, m].$$

- Properties:
  - $C \geq 0$ trade-off parameter.
  - Convex optimization (strictly convex).
  - Unique solution.

Dual Optimization Problem

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to: } \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$ for any SV $\mathbf{x}_i$.

Kernel Methods

- Idea:
  - Define $K : X \times X \to \mathbb{R}$, called kernel, such that:

$$\Phi(x) \cdot \Phi(y) = K(x, y).$$

  - K often interpreted as a similarity measure.
- Benefits:
  - Efficiency: K is often more efficient to compute than Φ and the dot product.
  - Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (Mercer's condition).

Example - Polynomial Kernels

- **Definition:**

$$\forall x, y \in \mathbb{R}^N, \; K(x,y) = (x \cdot y + c)^d, \quad c > 0.$$

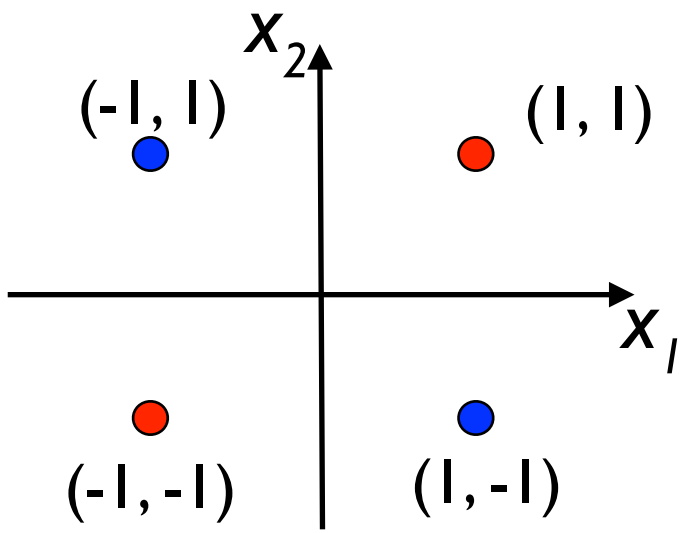
- **Example:** for $N = 2$ and $d = 2$,

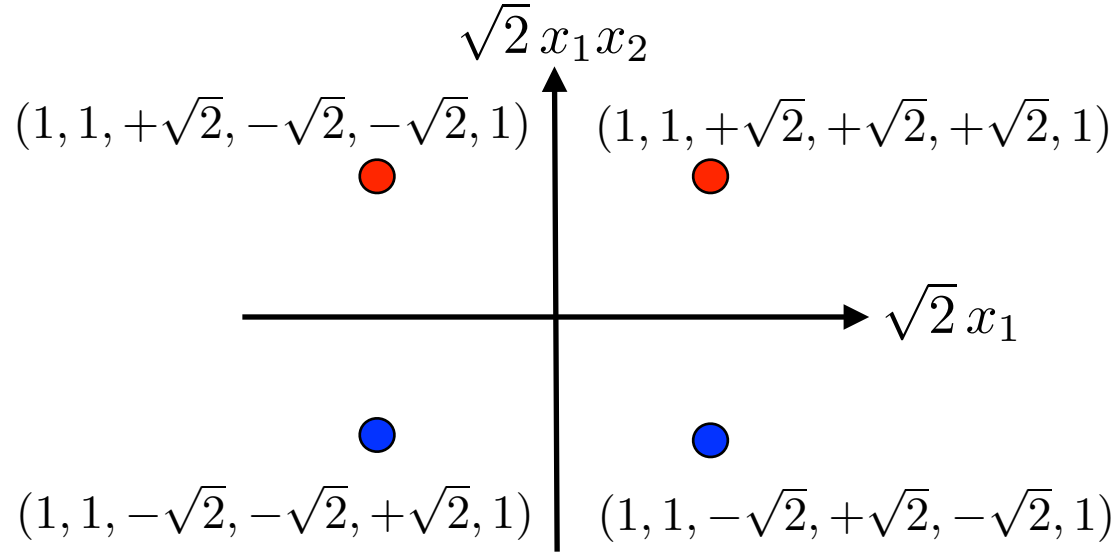
$$K(x,y) = (x_1 y_1 + x_2 y_2 + c)^2$$

$$= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}\, y_1 y_2 \\ \sqrt{2c}\, y_1 \\ \sqrt{2c}\, y_2 \\ c \end{bmatrix}.$$

XOR Problem

- Use second-degree polynomial kernel with $c = 1$:





Linearly non-separable

Linearly separable by $x_1x_2 = 0$.

Other Standard PDS Kernels

- Gaussian kernels:

$$K(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right), \ \sigma \neq 0.$$

- Sigmoid Kernels:

$$K(x, y) = \tanh(a(x \cdot y) + b), \ a, b \geq 0.$$

Consequence: SVMs with PDS Kernels

(Boser, Guyon, and Vapnik, 1992)

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{subject to: } 0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(x_j, x_i)$ for any x_i with $0 < \alpha_i < C$.

SVMs with PDS Kernels

- Constrained optimization:

$$\max_{\boldsymbol{\alpha}} 2\,\boldsymbol{\alpha}^\top \mathbf{1} - \boldsymbol{\alpha}^\top \mathbf{Y}^\top \mathbf{K} \mathbf{Y} \boldsymbol{\alpha}$$
$$\text{subject to: } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \wedge \boldsymbol{\alpha}^\top \mathbf{y} = 0.$$

- Solution:

$$h = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i K(x_i, \cdot) + b\Big),$$

with $b = y_i - (\boldsymbol{\alpha} \circ \mathbf{y})^\top \mathbf{K} \mathbf{e}_i$ for any x_i with $0 < \alpha_i < C$.

Regression Problem

- **Training data**: sample drawn i.i.d. from set X according to some distribution D,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in X \times Y,$$

with $Y \subseteq \mathbb{R}$ is a measurable subset.

- **Loss function**: $L : Y \times Y \to \mathbb{R}_+$ a measure of closeness, typically $L(y, y') = (y' - y)^2$ or $L(y, y') = |y' - y|^p$ for some $p \geq 1$.

- **Problem**: find hypothesis $h : X \to \mathbb{R}$ in H with small generalization error with respect to target f

$$R_D(h) = \mathop{\mathrm{E}}_{x \sim D} \left[L\big(h(x), f(x)\big) \right].$$

Kernel Ridge Regression

(Saunders et al., 1998)

- Optimization problem:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y} - \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$

or $$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} -\boldsymbol{\alpha}^\top (\mathbf{K} + \lambda \mathbf{I}) \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{y}.$$

- Solution:

$$h(x) = \sum_{i=1}^{m} \alpha_i K(x_i, x),$$

with $$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}.$$

Questions

- How should the user choose the kernel?
  - problem similar to that of selecting features for other learning algorithms.
  - poor choice $\longrightarrow$ learning made very difficult.
  - good choice $\longrightarrow$ even poor learners could succeed.
- The requirement from the user is thus critical.
  - can this requirement be lessened?
  - is a more automatic selection of features possible?