Kernel_methods : Theory and computation

Goal: develop new kernel or other ML methods

Canvas, Github (tentative) lecture schedule

syllabus.pdf

- 2 homeworks: (20%) + in-class
- Project 60% (presentation + report) + proposal (1-page)
- Paper presentation (3 lectures) 20%
- OH by appt

Kernels

Data space: $\mathcal{X}$
(Hilbert space with inner product structure, $\mathbb{R}^n$ ...)

Target computation: classification (decisions)
regression (solving PDEs)
generative modeling, sampling (Bayesian inference)
clustering
dimension reduction
timeseries / sequence modeling
feature extraction / learning

Want: pattern recognition / dependencies within given data that "generalize"

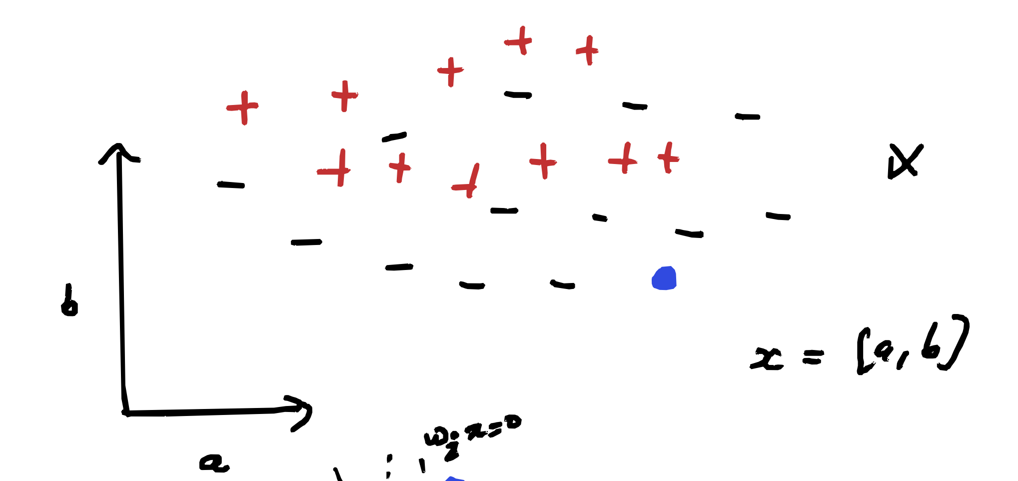
Kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{C}$
"Similarity measure"

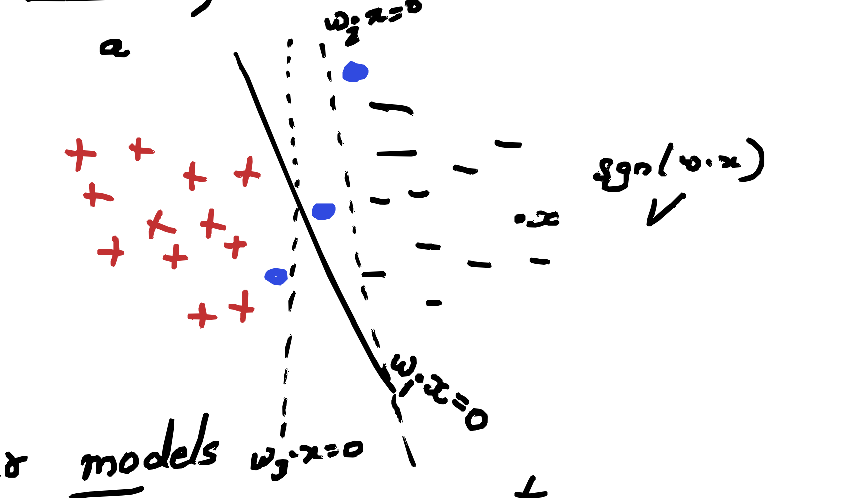
kernel methods: use kernel function for downstream computations

Binary classification

Data: $x_1, x_2 \dots, x_n \in \mathcal{X}$
target label $f(x_1), f(x_2) \dots, f(x_n) \in \{\pm 1\}$] input

Want h s.t.
$h(x) = f(x)$ for any x.

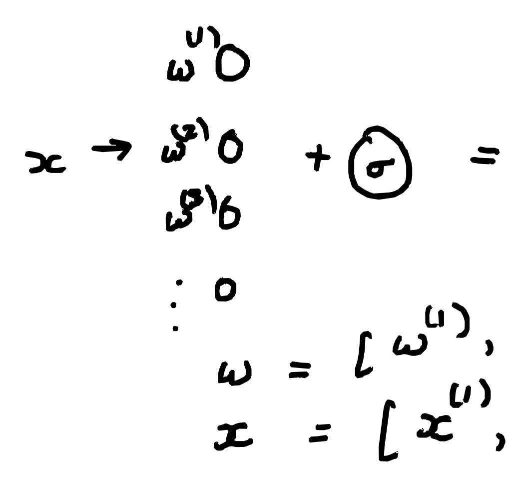




Linear models
Ansatz
Perceptron

$$h(x) = \mathrm{sgn}(\underset{\text{weight}}{w} \cdot x + \underset{\text{bias}}{b})$$

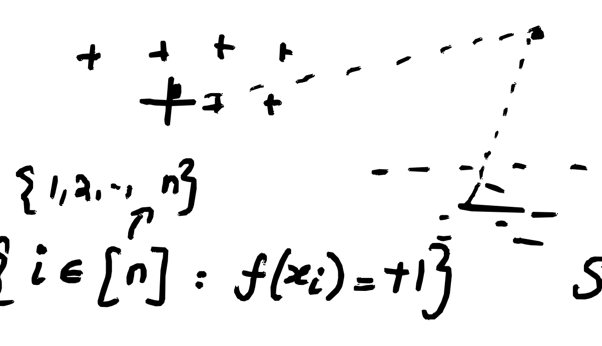
Want: w, b s.t. $h(x) = f(x)$



$$w = [w^{(1)}, \dots, w^{(n)}, b]$$
$$x = [x^{(1)}, \dots, x^{(n)}, 1]$$

$$h(x) = \mathrm{sgn}(w \cdot x)$$
$$\mathrm{sgn}(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$

Using the means



$$S_+ = \{ i \in \underset{\{1,2,\dots,n\}}{[n]} : f(x_i) = +1 \} \qquad S_- = \{ i \in [n] : f(x_i) = -1 \}$$

$$m_+ = \frac{1}{|S_+|} \sum_{i \in S_+} x_i \qquad |S_+| = n_+$$
$$|S_-| = n_-$$
$$m_- = \frac{1}{|S_-|} \sum_{i \in S_-} x_i$$

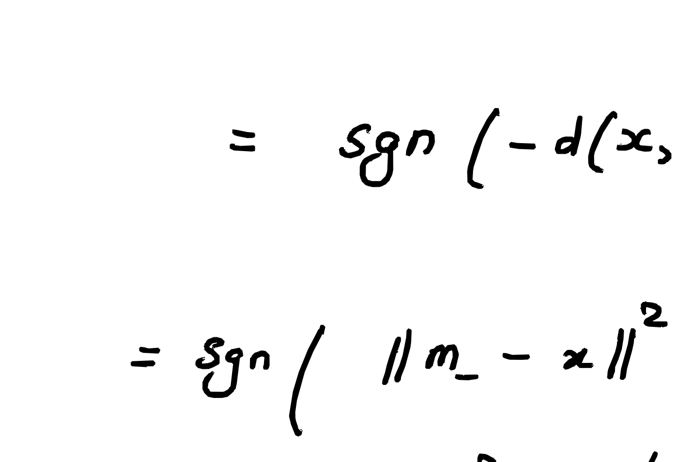
$$h(x) = \begin{cases} +1 & d(x, m_+) < d(x, m_-) \\ -1 & d(x, m_+) \geq d(x, m_-) \end{cases}$$

$$= \mathrm{sgn}(-d(x, m_+) + d(x, m_-))$$

$$= \mathrm{sgn}(\|m_- - x\|^2 - \|m_+ - x\|^2)$$

$$= \mathrm{sgn}(\|m_-\|^2 + \cancel{\|x\|^2} - 2 m_- \cdot x - \|m_+\|^2 - \cancel{\|x\|^2} + 2 m_+ \cdot x)$$

$$= \mathrm{sgn}\left((m_+ - m_1) \cdot x + \frac{1}{2}(\|m_-\|^2 - \|m_+\|^2) \right)$$



X : RV $\quad Y = f(X)$ RV
Bayesian decision rule
Probabilistic assumption on data

$$h(x) = \begin{cases} +1 & \Pr(Y = 1 | X = x) \geq \Pr(Y = -1 | X = x) \\ -1 & \Pr(Y = -1 | X = x) < \Pr(Y = 1 | X = x) \end{cases}$$

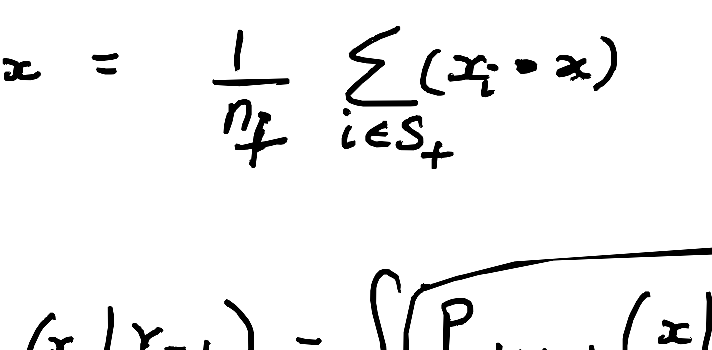
$Y = \pm 1$
$$P_Y(y) = \begin{cases} \frac{1}{2} & y = 1 \\ \frac{1}{2} & y = -1 \end{cases}$$

X
$P_X(x)$ given

Data distribution: joint distribution of X & Y
$$\underset{\text{class conditional}}{P_{X|Y}} \cdot P_Y$$

$$P_{Y|X}(y | X = x) = \frac{P_{X|Y}(X = x | y) P_Y(y)}{P_X(x)}$$

Bayes decision rule
$$h(x) = \mathrm{sgn}(P_{X|Y}(x | Y = 1) - P_{X|Y}(x | Y = -1))$$



$$m_+ = \frac{1}{n_+} \sum_{i \in S_+} x_i \qquad m_- = \frac{1}{n_-} \sum_{i \in S_-} x_i$$

$$m_+ \cdot x = \frac{1}{n_+} \sum_{i \in S_+} (x_i \cdot x)$$

$$P_{X|Y}(x | Y = 1) = \int P_{X|Y,X'}(x | Y = 1, X' = x') P_{X|Y}(x' | Y = 1) \, dx'$$
$$\approx \frac{1}{n_+} \sum_{i \in S^+} \mathcal{K}(x, x_i)$$

estimator for $P_{X|Y}(x | Y = 1)$

$$P_{X|Y}(x_i | Y = 1) = \begin{cases} \frac{1}{n_+} & i \in S^+ \\ 0 & i \in S^- \end{cases}$$

$$P_{X|Y}(x_i | Y = -1) = \begin{cases} \frac{1}{n_-} & i \in S^- \\ 0 & i \in S^+ \end{cases}$$

($\int f(x) \rho(x) \, dx \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i)$
$x_i \sim$ iid according to ρ
MC estimator).