Recap

kernel perceptron $\qquad x \in \mathbb{R}^d$

$$h(x) = \mathrm{sgn}\Big(\sum_{i \in \mathcal{I}} \eta\, y_i\, \mathcal{K}(x_i, x)\Big)$$

incorrectly classified indices

Polynomial kernel

$$\mathcal{K}(x, x') = (c + x \cdot x')^m$$

Last time: feature space associated is finite-dimensional

$$\Phi(x) = \begin{bmatrix} x^{(1)}(x) \\ x^{(2)}(x) \\ \sqrt{2}\, x^{(1)}(x)\, x^{(2)}(x) \\ x^{(1)^2}(x) \\ x^{(2)}(x) \\ \sqrt{c} \end{bmatrix}$$

$$\mathcal{K}(x, x') = \Phi(x) \cdot \Phi(x')$$

Euclidean dot product

Dim of feature space = # of monomials of upto degree m.

=

$$x^{(1)^{m_1}} x^{(2)^{m_2}} \cdots\cdots x^{(d)^{m_d}}$$

degree $\quad m_1 + m_2 + \cdots + m_d \leq m$

$\circ \circ \cdots \circ \cdots \circ \, ||||$

$$\binom{m+d}{d}$$

Linear model on data space (in finite dimensions) $\xrightarrow{\text{Lift}}$ Linear model on feature space

$\downarrow$ feature space : inf. or finite dimensional

Nonlinear models evaluated using kernels.

Kernel: a scalar function on $\mathcal{X} \times \mathcal{X}$.

Positive definite kernels

A kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is called PD kernel if for any integer $p \in \mathbb{N}$ and $c \in \mathbb{R}^p$, $\sum_{i=1}^{p} c_i c_j \mathcal{K}(x_i, x_j) \geq 0$.

$\mathcal{K}(x_i, x_j) = \mathcal{K}(x_j, x_i)$ (symmetric)

Gram matrix : Given p data points $x_1, \ldots, x_p \in \mathcal{X}$ a gram matrix, G, associated with a kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a $p \times p$ matrix with elements $G_{ij} = \mathcal{K}(x_i, x_j)$

$1 \leq i \leq p$

$1 \leq j \leq p$.

PD kernel $\iff$ Gram matrix symmetric positive semi-definite.

$$\sum_{i=1}^{p}\sum_{j=1}^{p} c_i c_j \kappa(x_i, x_j) = c^T G c \geq 0$$

$\iff$ G is PSD

$\iff$ all eigenvalues of G are real, non-negative

claim: if κ satisfies $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$

for some $\Phi \in \mathcal{H}$ (feature space),

then κ is a PD kernel

(Part of Mercer's theorem)

Want: G is SPSD.

$$\sum_{i,j=1}^{p} c_i c_j \kappa(x_i, x_j) = \sum_{i,j=1}^{p} c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle$$

$$= \left\langle \sum_{i=1}^{p} c_i \Phi(x_i), \sum_{j=1}^{p} c_j \Phi(x_j) \right\rangle$$

$$= \left\| \sum_{i=1}^{p} c_i \Phi(x_i) \right\|^2 \geq 0$$

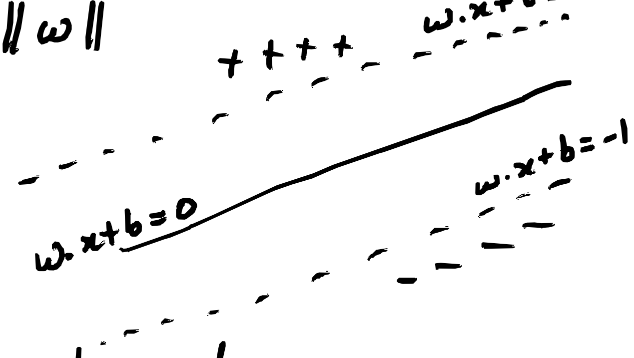
Support Vector Machine

Geometric Margin : min distance b/w training points & separating hyperplane

Last time : Max margin $\rightarrow$ separating hyperplane.

$$h(x) = \mathrm{sgn}(w \cdot x + b)$$

$$\max_{w,b} \min_i \frac{(w \cdot x_i + b) y_i}{\|w\|}$$



Canonical form of hyperplane :

$$\min_i y_i (w \cdot x_i + b) = 1$$

$$\max_{w,b} \min_i \left| \frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|} \right|$$

$$\max_{v \in S^{d-1}, b} \min_i | v \cdot x_i + b |$$

Derivation of max-margin classifier

(A) $\forall i \in [n], \quad y_i (w \cdot x_i + b) \geq 1$

$$\max_{w,b} \min_i \frac{y_i (w \cdot x_i + b)}{\|w\|}$$

(B)

$$\max_{w,b} \frac{1}{\|w\|}$$

s.t. $y_i (w \cdot x_i + b) \geq 1$
$i \in [n]$

(C)

$$\min_{w,b} \frac{\|w\|^2}{2}$$

s.t. $y_i (w \cdot x_i + b) \geq 1$

(SVM original formulation)

quadratic program convex problem with affine constraints

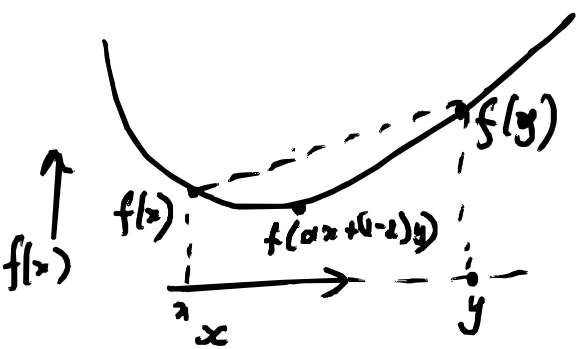
Convex optimization

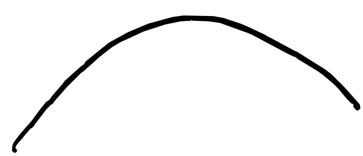
Convex function: line drawn between points on the image lies above the function

Epi. of function is convex set

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y)$$

$\alpha \in (0,1)$





max concave or min convex

→ convex optimization

KKT condition: necessary & sufficient conditions for solution of convex program

Constrained optimization problem:

Primal problem
$$\begin{cases} p^* = \min_{\omega} f(\omega) \\ \text{s.t.} \quad g_i(\omega) \leq 0 \\ \quad i = 1, \dots, p \\ \omega \in \mathbb{R}^d. \end{cases}$$

Convex problem:
f, g_i convex & differentiable

Lagrangian
$$\mathcal{L}(\omega, \lambda) = f(\omega) + \sum_{i=1}^{p} \lambda_i g_i(\omega)$$
λ: dual variables $\quad \lambda \in \mathbb{R}^p$

Dual problem:
$$d^* = \max_{\lambda} \min_{\omega} \mathcal{L}(\omega, \lambda)$$
$$\lambda_i \geq 0$$

$$d^* \leq p^*$$
$$d^* = p^* \quad \text{(strong duality)}$$

All local minima are global minima.

<u>KKT</u>:
- f is convex, diff
- g_i are convex, diff
- Slater's condition: $\exists \ \omega_0 \in \text{dom}(\omega)$ s.t. $g_i(\omega_0) < 0$ for every i

  or

  $g_i(\omega_0) \leq 0$ for any i & g_i is affine.

Then, $\exists \ \lambda$, $\lambda_i \geq 0$, s.t. at a minimum ω, the following hold: $\quad \lambda = [\lambda_1, \dots, \lambda_p]^T \in \mathbb{R}^p$

i) $\nabla_\omega \mathcal{L}(\omega, \lambda) = 0$

(ii) $\nabla_\lambda \mathcal{L}(\omega, \lambda) \leq 0$
$$g_i(\omega) \leq 0 \quad \forall i$$

(iii) Complementarity conditions
$$\lambda_i = 0 \quad \text{or} \quad g_i(\omega) = 0. \quad \forall i$$
$$\lambda_i g_i(\omega) = 0 \quad \forall i.$$

SVM

$$\min_{w, b} \quad \frac{\|w\|^2}{2}$$

$$\text{s.t.} \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i \in [n]$$

$$\mathcal{L}([w, b], \lambda) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \lambda_i (y_i (w \cdot x_i + b) - 1)$$

$$f(w) = \frac{\|w\|^2}{2}$$

$$g_i(w) = - y_i (w_i \cdot x + b) + 1$$

KKT conditions

$$\nabla_w \mathcal{L}([w, b], \lambda) = w - \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$= 0$$

$$\Rightarrow \quad w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$\nabla_b \mathcal{L}(w, \lambda) = 0$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

Physical meaning of term "support vector"

x_i where $\lambda_i \neq 0$.

Think of each data point x_i as exerting a force $\lambda_i y_i w$ on separating hyperplane.

Then, force balance on hyperplane gives:

$$\sum_i F_i = \sum_{i=1}^{m} \lambda_i y_i w = 0 \quad (\nabla_b \mathcal{L} = 0)$$

Torque balance gives:

$$\sum_{i=1}^{m} r_i \underset{\text{cross product}}{\times} F_i = \sum_{i=1}^{m} x_i \times \lambda_i y_i w$$

$$= \sum_{i=1}^{m} \lambda_i y_i x_i \times w$$

$$\left(\because \sum_{i=1}^{m} \lambda_i y_i x_i = w \text{ from } \nabla_w \mathcal{L} = 0\right) \quad = w \times w = 0$$