→ Convex optimization

→ SVM, KKT conditions to SVM

(derivation via margin maximization)

Variants of SVM

$$\min_{\substack{w, b \in \mathbb{R} \\ \in \mathbb{R}^d}} \frac{\|w\|^2}{2}$$

$$\text{s.t.} \quad y_i (w \cdot x_i + b) \geq 1$$

$d = \dim(x)$

Margin $\dfrac{1}{\|w\|}$

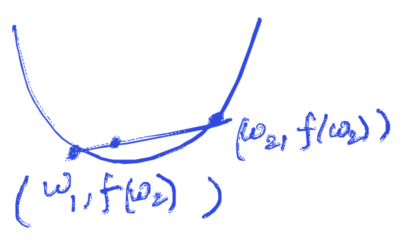
Convex optimization

Convex function f a convex set Ω

for any $\omega_1, \omega_2 \in \Omega$,

$$f(\alpha \omega_1 + (1-\alpha)\omega_2) \leq \alpha f(\omega_1) + (1-\alpha) f(\omega_2)$$

$<$ strictly convex



Thm: Suppose a convex function $f : \Omega \to \mathbb{R}$ has a minimum. Then, the set of minima forms a convex subset of Ω.

If f is strictly convex, there is a unique minimum.

Proof:

Let $M_f \subseteq \Omega$ be the set of minima of f. Let m_f be the minimum value of f on Ω.

Let $\omega, \omega' \in M_f$, for any $\alpha \in [0,1]$, $\alpha\omega + (1-\alpha)\omega' \in M_f$.

$$f(\alpha\omega + (1-\alpha)\omega') \leq \alpha f(\omega) + (1-\alpha) f(\omega')$$
$$= \alpha m_f + (1-\alpha) m_f$$
$$= m_f$$

$$\Rightarrow \alpha\omega + (1-\alpha)\omega' \in M_f$$

when f is strictly convex, $\omega = \omega'$.

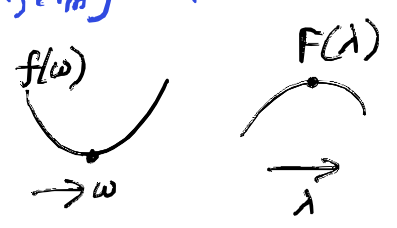
M_f is a singleton set.

KKT theorem for constrained optimization

$$\left.\begin{aligned} \min_{\omega \in \mathbb{R}^d} \; & f(\omega) = P_{opt} \\ \text{s.t.} \; & g_i(\omega) \leq 0 \quad i = [m] \end{aligned}\right\} \text{ primal}$$

Lagrangian: $\mathcal{L}(\omega, \underset{\text{dual}}{\lambda}) = f(\omega) + \sum_{i=1}^{m} \lambda_i g_i(\omega)$

$\lambda \in \mathbb{R}^m \qquad \lambda = [\lambda_1, \ldots, \lambda_m]^T \in \mathbb{R}^m$



Dual problem:

$$\max_{\lambda \in \mathbb{R}^m} \; F(\lambda) := \inf_{\omega} \mathcal{L}(\omega, \lambda) = d_{opt}$$

$$\text{st. } \lambda_i \geq 0 \qquad i \in [m].$$

Weak duality: $d_{opt} \leq P_{opt}$

with equality for convex problem.

Assumption ①: f is convex, g is convex & diff.

KKT: There exists a $\lambda \in \mathbb{R}^m$ s.t. at a minimum $\omega \in \mathbb{R}^d$ of f, the following conditions are satisfied:

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

(ii) $\nabla_{\lambda} \mathcal{L}(\omega, \lambda) = [g_i(\omega) \leq 0]_{i=1,\ldots,m}$

(iii) Complementarity constraints: either $\lambda_i = 0$ or $g_i(\omega) = 0$. for every $i = 1, \ldots, m$.

② constraints are qualified: $\exists\, \omega \in \text{int}(\Omega)$ s.t. $g_i(\omega) < 0 \;\; \forall\, i \in [m]$.

or $g_i(\omega) \leq 0$ and g_i is affine

(Slater's conditions) $\qquad g_i(\omega) = x \cdot \omega + b$

Ref: Chapter 6 of "Learning with Kernels" Smola & Scholkopf.

SVM

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

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

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

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

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

At min w, b, $\exists\ \lambda \geq 0$

$$\left.\begin{aligned} \nabla_w \mathcal{L}(w, b, \lambda) &= 0 \\ \nabla_b \mathcal{L}(w, b, \lambda) &= 0 \end{aligned}\right]$$

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

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

Comp. cond:

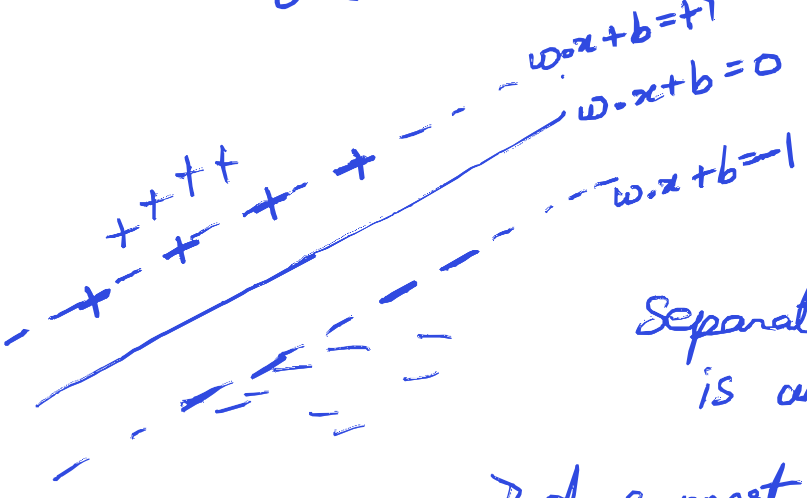
$\lambda_i = 0$ or $y_i(w \cdot x_i + b) = 1$, $i \in [m]$.

Last class: Force + torque balance

x_i for nonzero $\lambda_i \neq 0$ are called support vectors.

For support vectors x_i,

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



Separating hyperplane is unique.

$> d$ support vectors, w is not unique.

$$\omega = \sum_{i=1}^{m} \lambda_i y_i x_i \qquad \sum_{i=1}^{m} \lambda_i y_i = 0.$$

$$\lambda_i \neq 0, \quad y_i(\omega \cdot x_i + b) = 1$$

Dual problem

$$\inf_{\omega, b} \mathcal{L}(\omega, b, \lambda) = \inf_{\omega, b} \frac{\|\omega\|^2}{2} - \sum_{i=1}^{m} \lambda_i \left((\omega \cdot x_i + b) y_i - 1 \right)$$

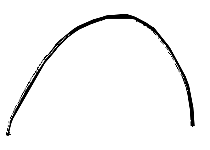
$$\max_{\lambda} F(\lambda) = \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j x_i \cdot x_j - \sum_{i=1}^{m} \lambda_i \left(y_i \left(\sum_{j=1}^{m} \lambda_j y_j x_j \cdot x_i + b \right) - 1 \right)$$

$$\lambda_i \geq 0 \qquad i \in [m]$$

$$\max_{\lambda} \; -\frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j x_i \cdot x_j + \sum_{i=1}^{m} \lambda_i$$

s.t. $\lambda_i \geq 0 \quad i \in [m]$.

Ex: Convex problem



Remark:
To evaluate objective function, we only need dot products on data space

• $h_{SVM}(x) = sgn(\omega \cdot x + b)$
$= sgn\left(\sum_{i=1}^{m} \lambda_i y_i x_i \cdot x + b \right)$

output of SVC also needs only dot products

⇒ So can replace $x_i \cdot x \to \mathcal{K}(x_i, x)$ to get nonlinear classifiers.

$$h_{svm}(x) = sgn\left(\sum_{i=1}^{m} \lambda_i y_i \mathcal{K}(x_i, x) + b \right)$$

When $\mathcal{K}$ PD kernel → linear classifier with features $\Phi(x)$ s.t.

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

(ℓ^2 inner product if $\Phi(x)$ are inf. dimensional).

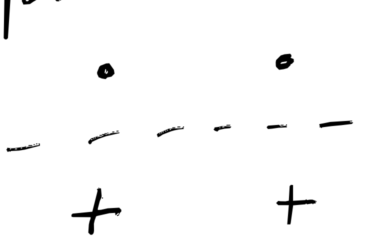
$x \longrightarrow \Phi(x)$

Data — feature (could be finite but high-dimensional OR inf-dimensional)

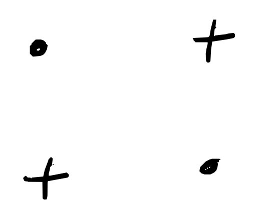
SVM → Kernel SVM

linear classifier on $\mathcal{X}$ — nonlinear classifier on $\mathcal{X}$; linear classifier on feature space.

e.g.
XOR function



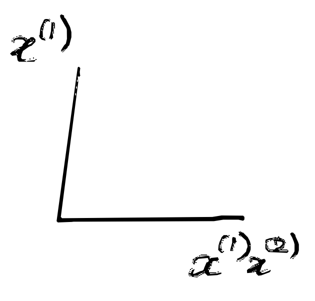
linearly separable



non-linearly separable

$$\mathbb{R}^2 \to \mathbb{R}^6$$

$$\begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \to \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ {x^{(1)}}^2 \\ {x^{(2)}}^2 \\ x^{(1)} x^{(2)} \\ c \end{bmatrix}$$



$$\mathcal{K}(x_i, x_j) = (x_i \cdot x_j + c)^2$$

polynomial kernel

Other kernels

hyperparameter

$$\kappa(x,y) = e^{-\frac{\|x-y\|^2}{2\eta^2}} \qquad \eta$$

Gaussian $\|x-y\|^2$ for given data points

$O(\|x-y\|^2) = O(\eta^2)$

big O

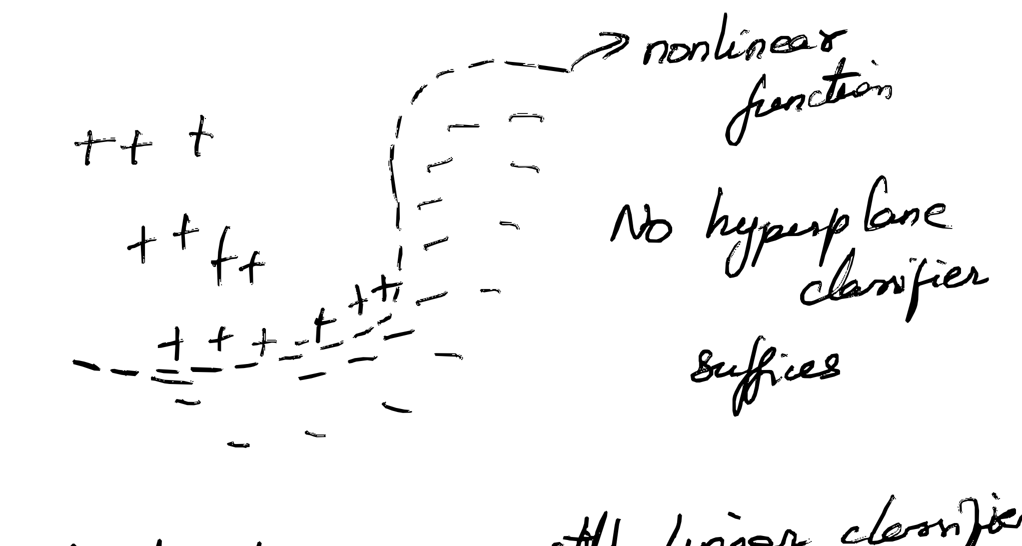
(inf-dimensional feature space)

- Sigmoid kernel (neural network)

$$\kappa(x,y) = \tanh(\sigma(x \cdot y) + b)$$

σ, b

Beyond linear classifiers



Kernel classifiers are still linear classifiers in feature space

Cover's theorem:

m points in d dimensions in "general position"

how many "labelings"/sets of m points are linearly separable?

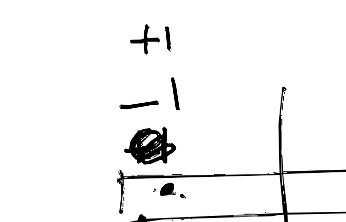
Subsets of m points/labeling in dimensions that are linearly separable = $C(m,d)$

- if $m \le d+1$, $C(m,d) = 2^m$
- if $m > d+1$, $C(m,d) = 2\sum_{i=0}^{d}\binom{m-1}{i}$

Sub exponential in m.

Example 2D

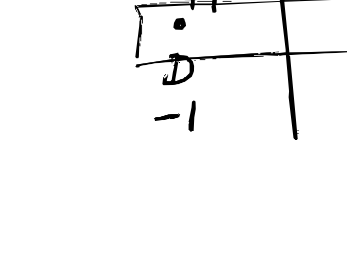
does lin. class. exist



+1 +1 +1
-1 -1 -1 } Yes

+1 +1 -1 Yes

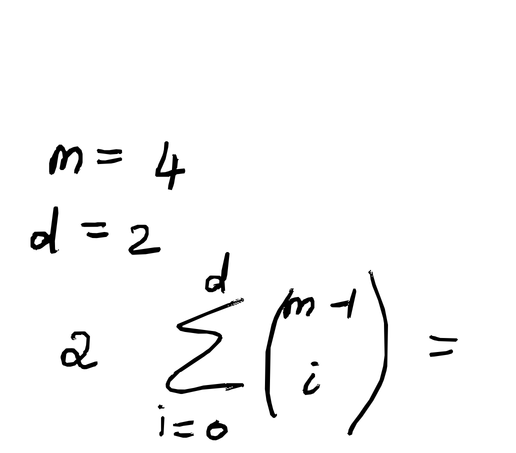
$\binom{3}{2}$



+1 -1 -1 Yes

$\binom{3}{2}$

$8 = 2^3$



$2^4 = 16$ subsets or labelings

$C(4,2)$ (# pts, dim)

$= C(3,2) + C(4,1)$
$= 8 + 6$

$= 14$

$m = 4$
$d = 2$

$$2\sum_{i=0}^{d}\binom{m-1}{i} = 2\left(\binom{3}{0}+\binom{3}{1}+\binom{3}{2}\right)$$

$$= 2(1+3+3) = 14$$

Ex:

$$C(m,d) = C(m-1,d) + C(m,d-1)$$

Fix m. As you increase d, there is a higher prob. of points in general position being linearly separable.