Lecture 9. Support vector machines (SVM)

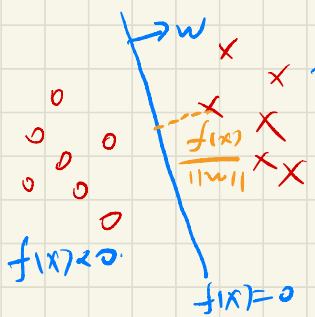
Linear classifier:

$$f(x) = w^T x + b \qquad y^* = \text{sign}(f(x)) = \begin{cases} +1, & f(x) \geq 0 \\ -1 & f(x) < 0 \end{cases}$$

distance from point X to boundary: $\frac{f(x)}{\|w\|}$

"Margin" — distance from the boundary to the closest point in the training set.

$$\gamma = \min_i \frac{|f(x_i)|}{\|w\|} = \min_i \frac{\|w^T x_i + b\|}{\|w\|}$$

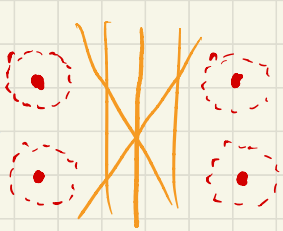


Idea: maximize the region, i.e. the seperation btwn the boundary and the points.

1) Perception — margin determines the complexity of learning.

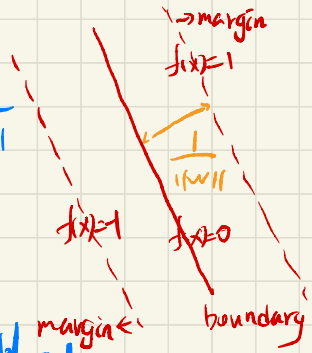
2) training points are random: leave a margin to be safe against the noise.

3) w is an uncertain estimate
- max margin → all more variance of w (boundary)



Need normalization: fix the numerator.

$$\min_i |w^T x_i + b| = 1 \quad \Rightarrow \quad \gamma = \frac{1}{\|w\|}$$



Maximize margin:

$$w^* = \arg\max_{w,b} \gamma, \text{ s.t. } \min_i |w^T x_i + b| = 1$$

$$= \arg\max_{w,b} \frac{1}{\|w\|}, \text{ s.t. } \min_i |w^T x_i + b| = 1$$

$$= \arg\min_{w,b} \|w\|^2, \text{ s.t. } \min_i |w^T x_i + b| = 1$$

$$= \arg\min_{w,b} \frac{1}{2}\|w\|^2, \text{ s.t. } |w^T x_i + b| \geq 1$$

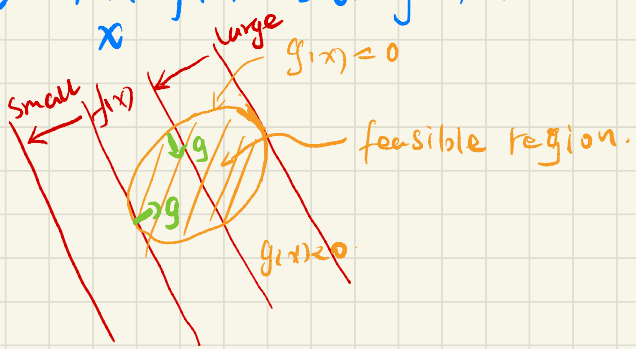
as optimum w will shrink s.t. $|w^T x_i + b| = 1$ for at least one i

$$w^* = \arg\min_{w,b} \frac{1}{2}\|w\|^2 \text{ s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall i.$$

Svm problem: assuming linearly separable data.

Optimization inequality constriants.

Goal: $\min_x f(x)$ s.t. $g(x) \geq 0$



Note: $\nabla g(x)$ points inside feasible region.

Consider 2 posibilities of x^*

1) x^* is on boundary, $g(x^*) = 0$ (active) equality

minimum when $\nabla f(x) = \lambda \nabla g(x)$ $\lambda > 0$

- both ∇f and ∇g point into the feasible region.

  if $\lambda < 0$. $f(x)$ is maximize. not minimize.

- f cannot decrease without leaving $g(x) \geq 0$

2) x^* is in feasible region, $g(x^*) > 0$ (inactive)

minimum when $\nabla f(x) = 0$ or $\lambda = 0$

combine 2 cases: find stationary point of lagrangian.

$$L(x, \lambda) = f(x) - \lambda g(x)$$

$$\nabla f(x) - \lambda \nabla g(x) = 0$$

$$\text{S.t.} \begin{cases} g(x) \geq 0 \\ \lambda \geq 0 \\ \lambda g(x) = 0 \end{cases}$$

KKT Conditions.

($g(x) > 0$ and $\lambda = 0$ OR

$g(x) = 0$ and $\lambda > 0$.)

Duality:

Suppose we have optimal λ^*, then minimize

$L(x, \lambda^*)$

$$L^* = \min_x L(x, \lambda^*) = \min_x f(x) - \lambda^* g(x)$$

Since $\lambda^* g(x^*) = 0$ at minimum.

$\Rightarrow L^* = f(x^*)$ ← the minimum we are trying to find

Define $q(\lambda) = \min_x L(x, \lambda) = \min_x [f(x) - \lambda g(x)]$

for every λ, find min of $L(x, \lambda)$ w.r.t. x

Note: $\lambda \geq 0$, $g(x) \geq 0$ $\Rightarrow$ $\lambda g(x) \geq 0$,

$$q(\lambda) \leq \min_{\substack{x \\ g(x) \geq 0}} f(x) = f(x^*)$$

($q(\lambda)$ is a lower-bound to $f(x^*)$)

Hence, maximizing $q(\lambda)$ could yield $f(x^*)$ (under some conditions).

The dual problem: $q^* = \max_{\lambda \geq 0} q(\lambda)$

weak dual theory: $q^* \leq f^*$ (if $q^* \neq f^*$, then there is a "duality gap"

Strong duality theory:

if 1) $f(x)$ is convex

2) the feasible region is convex $\{x \mid g(x) \geq 0\}$

3) not degenerate $\{x \mid g(x) > 0\} \neq \emptyset$

then $q^* = f^*$ (solving the dual problem is equal to solving the primal problem)

SVM dual problem:

let $\alpha_i \geq 0$ be the Lagrange multiplier for ith constraint

Lagrange

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1]$$

Find dual function $L(\alpha)$: set derivative to 0

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i [y_i x_i] = 0 \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$$

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

plug in w^* to $L(w, b, \alpha)$

$$L(\alpha) = \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{n} \alpha_j y_j x_j\right)$$

$$- \sum_{i=1}^{n} \alpha_i \left[y_i \left(\left(\sum_{j=1}^{n} \alpha_j y_j x_j \right)^T x_i + b \right) - 1 \right]$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j x_j^T x_i - \underbrace{\sum_{i=1}^{n} \alpha_i y_i b}_{0} + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

SVM dual problem:
$$\begin{cases} \max_{\alpha} L(\alpha) \\ \text{s.t. } \alpha_i \geq 0, \ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Given α^*, then $w^* = \sum_i \alpha_i y_i x_i$

Recall KKT

1) $g(x) = 0$ active $\Rightarrow$ $y_i(w^T x_i + b) - 1 = 0$, $\alpha_i > 0$ $\rightarrow$ x_i is on the margin $y_i(w^T x_i + b) = 1$

2) $g(x) > 0$ inactive $\Rightarrow$ $y_i(w^T x_i + b) - 1 > 0$, $\alpha_i = 0$ $\rightarrow$ x_i is beyond the margin $y_i(w^T x_i + b) > 1$

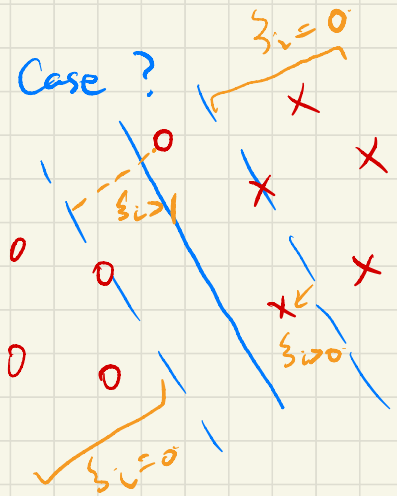
Note: w^* only depends on the non-zero α_i.

i.e. the points on the margin (called support vectors)

Soft SVM:

What about the non-separable case?

Soft-margin: most points satisfy the margin constriant, but some can violate the margin.



New constraint with slack: $y_i(w^T x_i + b) \geq 1 - \xi_i$

ξ_i: slack variable - allows some points to violate margin when $\xi_i > 0$

$\xi_i \geq 0$

New objective: $\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$

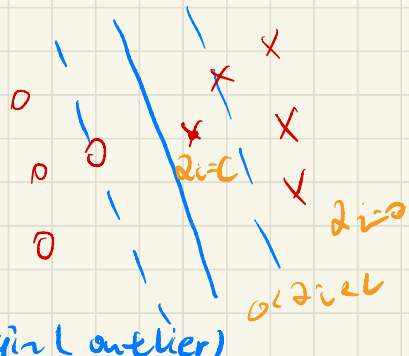
Penalize large ξ_i, prevent too much slack.

Dual objective: $\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$

$$s.t \quad \sum_{i=1}^{n} x_i y_i = 0$$

$0 \le \alpha_i \le C$ ← new upper-constraint on α_i

Geometrically



$\alpha_i = 0 \Rightarrow x_i$ beyond margin

$0 < \alpha_i < C \Rightarrow x_i$ on the margin.

$\alpha_i = C \Rightarrow x_i$ violates the margin (outlier)

Reconstruct: $w^* = \sum_{i=1}^{n} \alpha_i y_i x_i$

Note: $\alpha_i = C$ when x_i is an outlier, thus prevents a single outlier from dominatly the w^*

SVM Loss function: $L(z) = \max(0, 1 - z)$ ← "hinge loss"