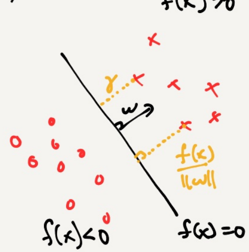
Lecture 9 Support Vector Machines (SVM)

Linear classifier

$$f(x) = w^T x + b$$

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



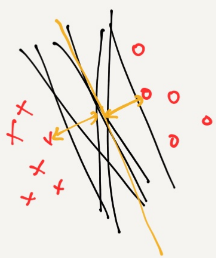
distance from point x to boundary:

$$\frac{f(x)}{\|w\|} \qquad \text{(PS 9-1)}$$

"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 margin, i.e. the separation b/wn the boundary & the points.

1) Perceptron - 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 → allow more variance of w (boundary)

Need normalization: fix the numerator

$$\min_i |w^T x_i + b| = 1$$

$$\Rightarrow \gamma = \frac{1}{\|w\|}$$

Maximize margin

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

$$= \underset{w,b}{\text{argmax}} \left(\frac{1}{\|w\|}\right) \quad , \text{s.t. } \min_i |w^T x_i + b| = 1$$

change max to min.

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

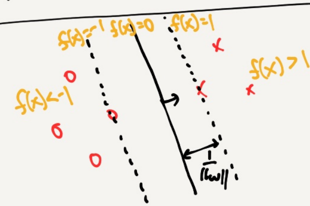
at optimum, w will shrink s.t. $|w^T x_i + b| = 1$ for at least one i.

$$= \underset{w,b}{\text{argmin}}\ \|w\|^2 \quad \text{s.t. } |w^T x_i + b| \geq 1 \quad \forall i$$

$$\boxed{w^* = \underset{w,b}{\text{argmin}}\ \tfrac{1}{2}\|w\|^2 \quad \text{s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall i}$$

SVM problem

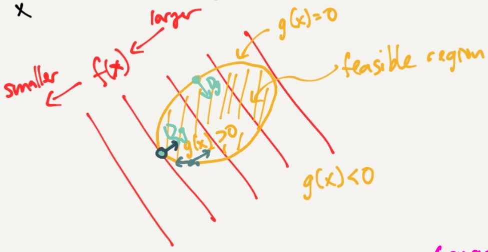
(assuming linearly separable data)



(what is a support vector?)

Optimization w/ inequality constraints

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



Note: $\nabla g(x)$ point inside feasible region. (gradient points to increasing direction)

Consider 2 possibilities of x^*

1) x^* is on boundary, $g(x^*) = 0$ (active) equality.
   minimum when $\nabla f(x) = \lambda \nabla g(x)$, $\lambda > 0$. (otherwise, f can still be decreased)
   - both ∇f & ∇g point into the feasible region.
   - f cannot decrease w/o leaving $g(x) \geq 0$

2) x^* is in feasible region, $g(x^*) > 0$ (inactive)
   minimum when $\nabla f(x) = 0$, or $\lambda = 0$ in other word.

Combine 2 cases: find stationary point of Lagrangian.

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

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

s.t. $\begin{cases} g(x) \geq 0 \\ \lambda \geq 0 \\ \lambda g(x) = 0 \end{cases}$ KKT conditions (Karush-Kuhn-Tucker)

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

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

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

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

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

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

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

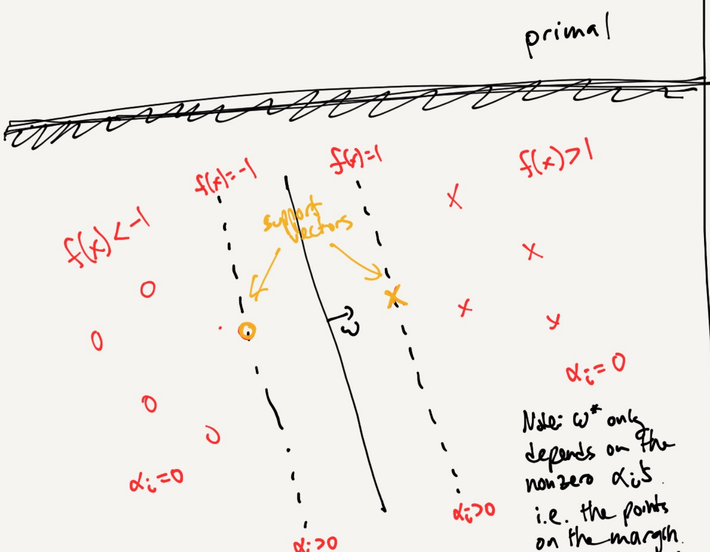
The dual problem: $g^* = \max_{\lambda \geq 0} g(\lambda)$

Weak duality thm: $g^* \leq f^*$ (if $g^* \neq f^*$, then there is a "duality gap".)

Strong duality thm:

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 $g^* = f^*$ (solving the dual problem is equivalent to solving the primal)



Note: w^* only depends on the nonzero α_i's. i.e. the points on the margin (called support vectors)

SVM dual problem

let $\alpha_i \geq 0$ be the Lagr multiplier for i^{th} constraint. $y_i(w^T x_i + b) \geq 1$

"$g(x)$" $\boxed{y_i(w^T x_i + b) - 1} \geq 0$

Lagrangian

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

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

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w^* = \sum_i \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

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

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

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

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

Recall KKT

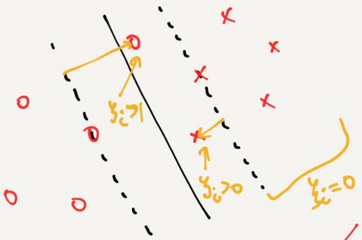
1) $g(x) = 0$ active: $y_i(w^T x_i + b) - 1 = 0$, $\alpha_i > 0$ → x_i is on the margin, $y_i(w^T x_i + b) = 1$

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

Soft-SVM

What about the non separable case?

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



Slack variable - allows some points to violate margin when > 0.

New constraint w/ slack

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

ξ xi

New objective:

penalize large ξ_i, prevent too much slack.

$$\min_{w,\xi} \|w\|^2 + C\sum_i \xi_i$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i$$

$$\xi_i \geq 0, \quad \forall i$$

Dual problem:

$$\max_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t. } \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

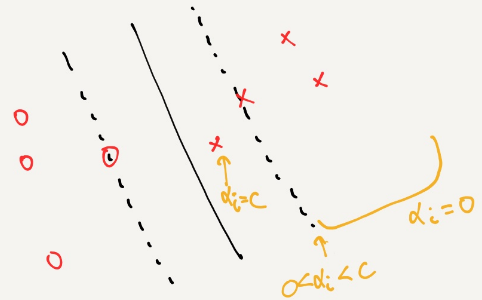
New upper-constraint on α_i.

Geometrically

$\alpha_i = 0 \Rightarrow x_i$ beyond margin

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

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



Reconstruct

$$w^* = \sum_i \alpha_i y_i x_i$$

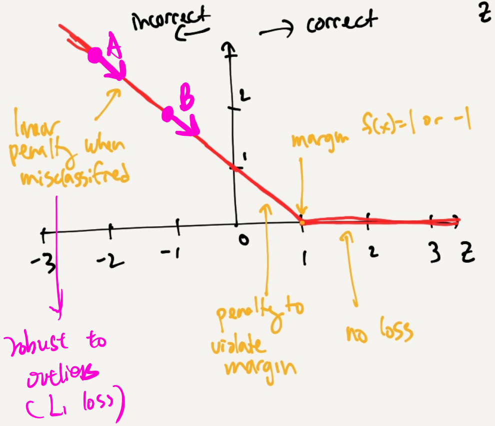
Note: $\alpha_i = C$ when x_i is an outlier, thus prevents a single outlier from dominating the w^*.

(The importance of outliers are controlled.)

SVM loss function

$L(z) = \max(0, 1-z)$ ← "hinge loss"

$z = yf(x)$



For A, B.

(move towards boundary have equivalent effect)

For a extremely outlier, unit decreasing gives same effect