SVM

Amr

April 1

Outline

- Margin and VC-dimension
- The separable case
- Non- separable case: Hing Loss
- Kernels

SVM: Intuition

- Remember LR
  - Predict Y=1 if $\frac{1}{1+\exp(w.x+b)} > 0.5$
  - or if w.x > 0
  - The more prob. >>> 0.5, the more confident we are about our prediction
  - Or the more w.x+b >>0 (margin), the more we are confident about our prediction
  - Same in SVM

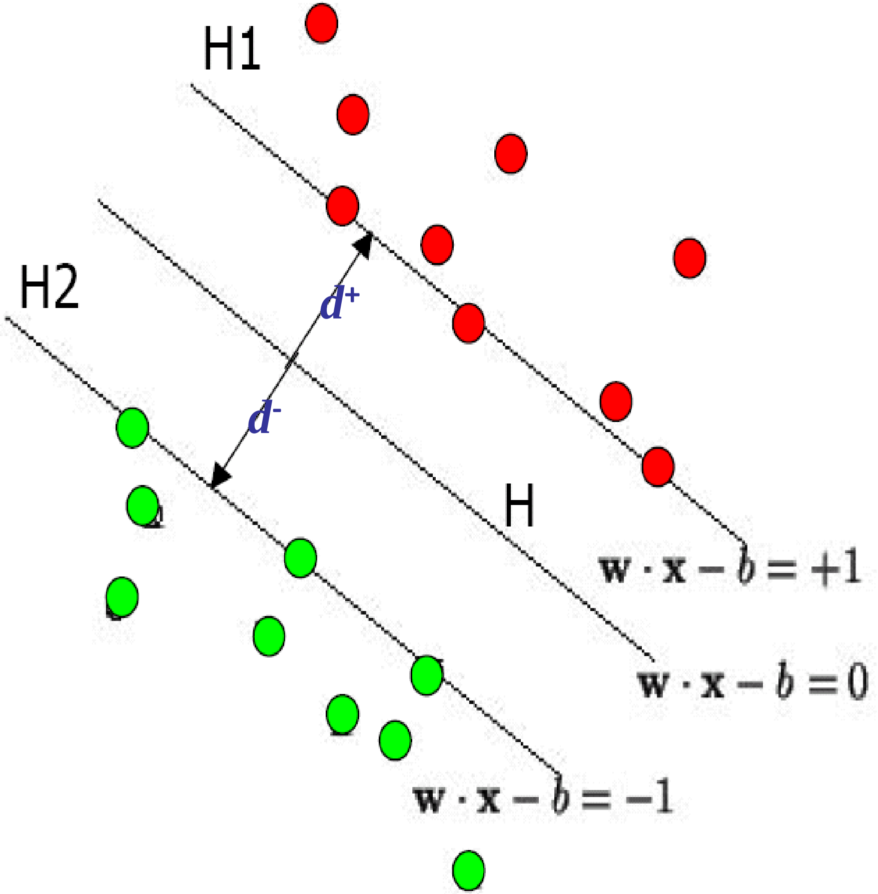
SVM

- SVM problem

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

- Or equivalently

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


H1
H2
d^+
d^-
H
$\mathbf{w} \cdot \mathbf{x} - b = +1$
$\mathbf{w} \cdot \mathbf{x} - b = 0$
$\mathbf{w} \cdot \mathbf{x} - b = -1$

SVM using VC-dimension

VC Theory

(Vapnik, 1982)
Given $x_1, \dots, x_n \in \mathbb{R}^d$ iid and $||x_i||_2 \leq D$, if $\mathcal{H}_\gamma$ is the hypothesis space of linear classifiers in $\mathbb{R}^d$ with margin γ,

$$\boxed{VC(\mathcal{H}_\gamma) \leq \min\left\{d, \left\lceil \frac{4D^2}{\gamma^2} \right\rceil\right\}.}$$

$$\boxed{error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}}$$

SVM using VC-dimension

- Thus large-margin → small VC-dim → better generalization bound
- Recall that d+1 is the upper bound for a linear classifier in d-space

$$\boxed{VC(\mathcal{H}_\gamma) \leq \min\left\{d, \left\lceil \frac{4D^2}{\gamma^2} \right\rceil\right\}.}$$

$$\boxed{error_{true}(h) < error_{train}(h) + \sqrt{\frac{VC(H)(\ln \frac{2m}{VC(H)} + 1) + \ln \frac{4}{\delta}}{m}}}$$

Outline

- Margin and VC-dimension
- The separable case
- Kernels
- Non-seperable case: Hing Loss

Solution Sketch

$$\begin{array}{ll} \min_{w,b} & \frac{1}{2} w^t w \\ \text{s.t} & y_i(w^T x_i + b) \geq 1, \quad \forall i \end{array}$$

- Form the Langrangian
- Optimize with respect to primal variable
- Subs. Into Lagrangian to get dual problem
- Exploit the KKT condition

Lagrangian

$$\operatorname*{argmin}_{w,b} \frac{1}{2}||w||^2$$

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

$$L(w, b, \alpha) = \frac{1}{2} w \cdot w - \sum_i \alpha_i [y_i(w \cdot x_i + b) - 1]$$

- One dual variable per constraints

$$L(w, b, \alpha) = \frac{1}{2} w \cdot w - \sum_i \alpha_i [y_i (w \cdot x_i + b) - 1]$$

$$\frac{\partial}{\partial w} L(w, b, \alpha) = w - \sum_i \alpha_i y_i x_i = 0 \quad \rightarrow \quad w = \sum_i \alpha_i y_i x_i .$$

$$\frac{\partial}{\partial b} L(w, b, \alpha) = \sum_i \alpha_i y_i = 0.$$

$$\boxed{\begin{aligned} &\operatorname*{argmax}_{\alpha} L(w, b, \alpha) = \operatorname*{argmax}_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ &s.t. \quad \alpha_i \geq 0, \\ &\qquad \sum_i \alpha_i y_i = 0. \end{aligned}}$$

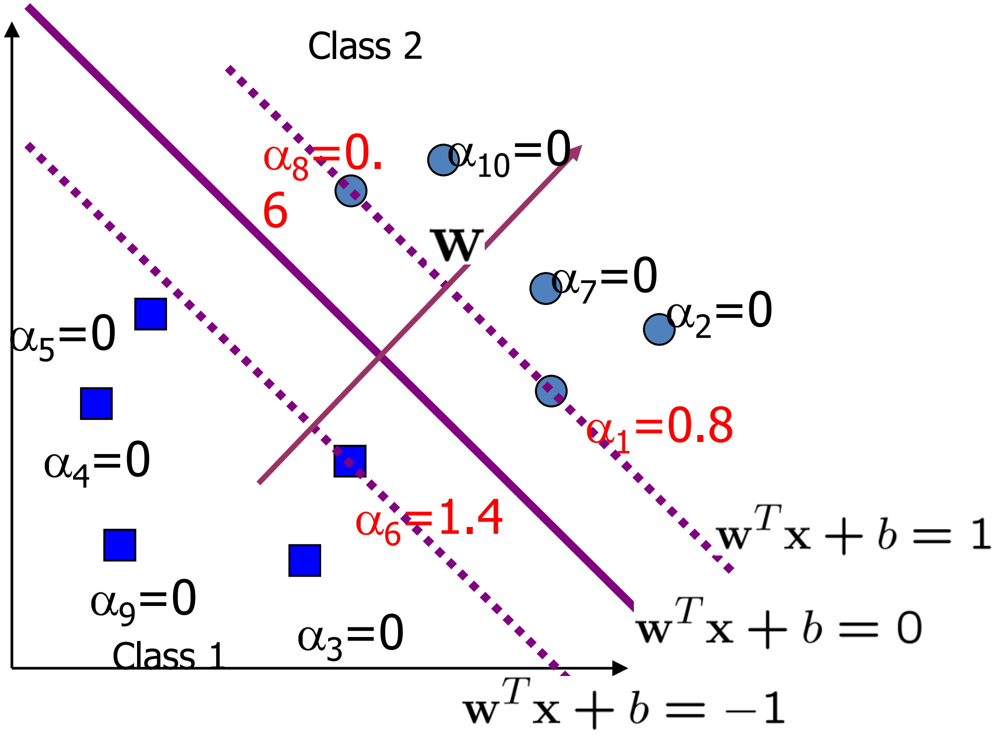
KKT conditions

- For convex objective and affine constraints we have

$$\alpha_i \left[1 - y_i \left(w.x_i + b\right)\right] = 0, \quad i = 1, \dots, m$$

- Only a few α_i can be non-zero.

$$w = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i$$

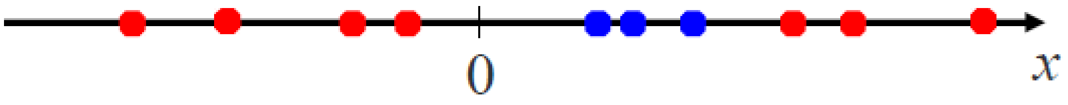


- How about *b*?

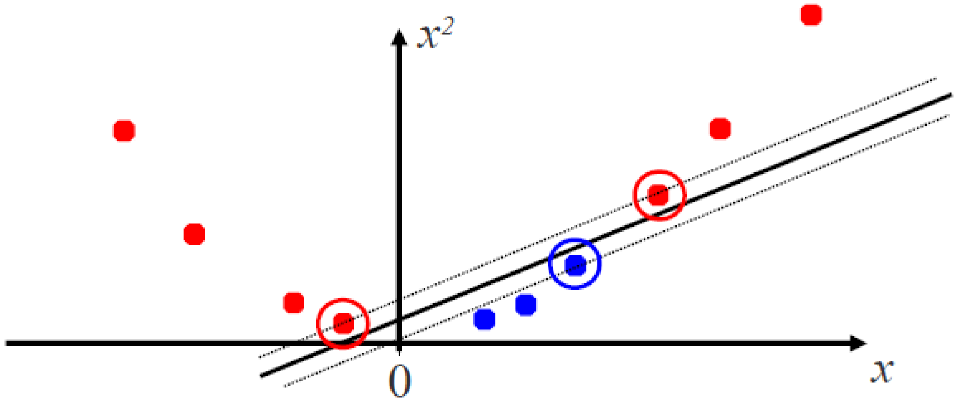
$$y^*(z) = \operatorname{sign}\left(\sum_{i \in SV} \alpha_i y_i \left(x_i^T z\right) + b\right)$$

Kernel Trick

- Is this data linearly-separable?

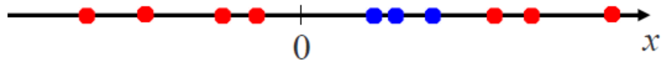


- How about a quadratic mapping $\phi(x_i)$?

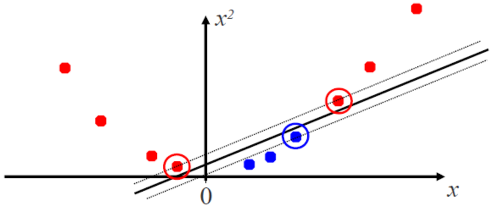


Kernel Trick

Input space



feature space



- Simply replace x_i with $\phi(x_i)$!

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \left(\phi(\mathbf{x}_i)^t \phi(\mathbf{x}_j)\right)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

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

$$y^*(z) = \operatorname{sign}\left(\sum_{i \in SV} \alpha_i y_i \left(\phi(\mathbf{x}_i)^t \phi(z)\right) + b \right)$$

- So what is the deal?

Kernel Trick

- Computation depends on feature space
  - Bad if its dimension is much larger than input space

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

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

Where $K(x_i,x_j) = \phi(x_i)^t \, \phi(x_j)$

$$y^*(z) = \operatorname{sign}\left(\sum_{i \in SV} \alpha_i y_i K(\mathbf{x}_i, z) + b \right)$$

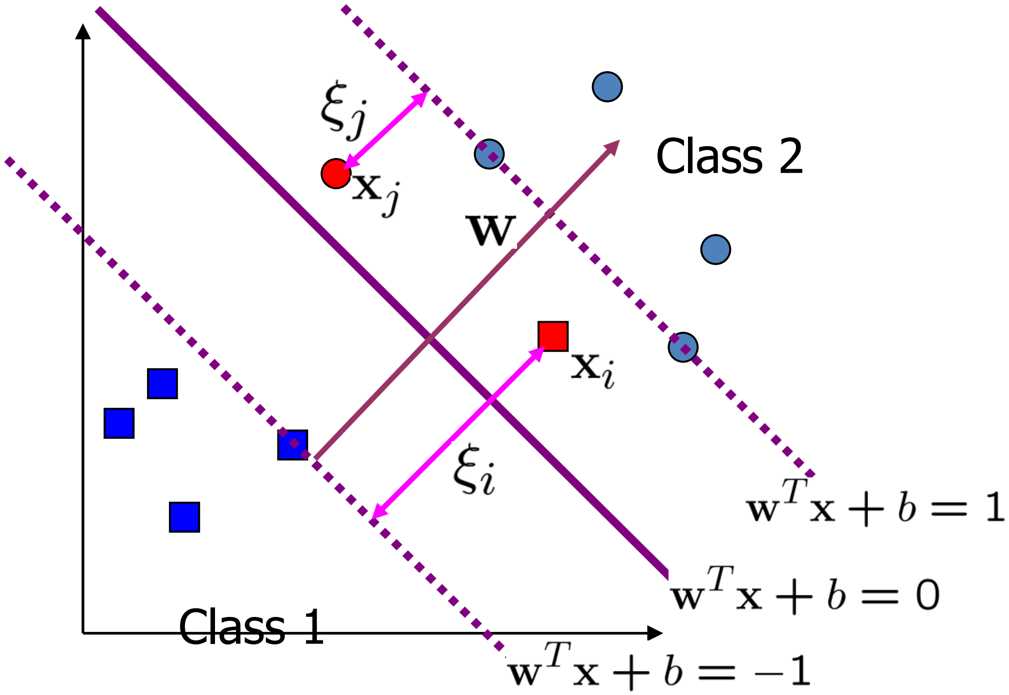
Example Kernel

- x_i is a bag of words
- Define $\phi(x_i)$ as a count of every n-gram up to n=k in x_i.
  - This is huge space 26^k
  - What are we measuring by $\phi(x_i)^t \phi(x_j)$?
- Can we compute the same quantity on input space?
  - Efficient linear dynamic program!
- Kernel is a measure of similarity
- Must be positive semi-definite

Outline

- Margin and VC-dimension
- The separable case
- Kernels
- Non-separable case: Hing Loss

Non-separable case


ξ_j
$\mathbf{x}_j$
$\mathbf{w}$
Class 2
$\mathbf{x}_i$
ξ_i
$\mathbf{w}^T\mathbf{x}+b=1$
$\mathbf{w}^T\mathbf{x}+b=0$
Class 1
$\mathbf{w}^T\mathbf{x}+b=-1$

$$\min_{w,b} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t} \quad \begin{aligned} & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \end{aligned}$$

Remember Ridge regression

- Min [squared loss + λ $w^t w$]
- How about SVM?

$$\text{argmin}_{\{w,b\}} \underbrace{w^t w}_{\text{regularization}} + \underbrace{\lambda \sum_1^m \max(1 - y_i(w^t x_i + b), 0)}_{\text{Loss: hinge loss}}$$

$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t} \quad \begin{aligned} & y_i(w^T x_i + b) \ge 1 - \xi_i, \quad \forall i \\ & \xi_i \ge 0, \quad \forall i \end{aligned}$$

$$\xi_i \ge \max(0, 1 - y_i(w^T x_i + b))$$

⇓ Why?

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

$$\text{argmin}_{\{w,b\}} \underbrace{w^t w}_{\text{regularization}} + \lambda \underbrace{\sum_{1}^{m} \max(1 - y_i(w^t x_i + b), 0)}_{\text{Loss: hinge loss}}$$