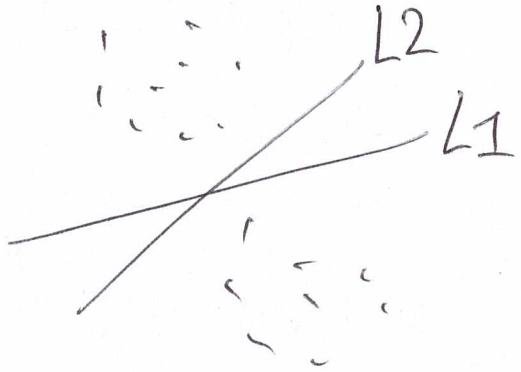
Lecture 8 : Support Vector Machines

* Linear classifier



$D = \{(x_i, y_i)\}$

$x_i \in \mathbb{R}^n$, $y_i \in \{\pm 1\}$

$F(x) = \langle w, x \rangle - b$

w : weight vector , b : bias

① $\begin{cases} \langle w, x_i \rangle - b > 0 \text{ if } y_i = +1 \\ \langle w, x_i \rangle - b < 0 \text{ if } y_i = -1 \end{cases}$

$\hookrightarrow y_i(\langle w, x_i \rangle - b) > 0 \quad \forall (x_i, y_i) \in D$

If possible : D is linearly separable

② Margin = distance hyperplane to closest data points

$y_i(\langle w, x_i \rangle - b) \geq 1 \quad \forall (x_i, y_i) \in D$

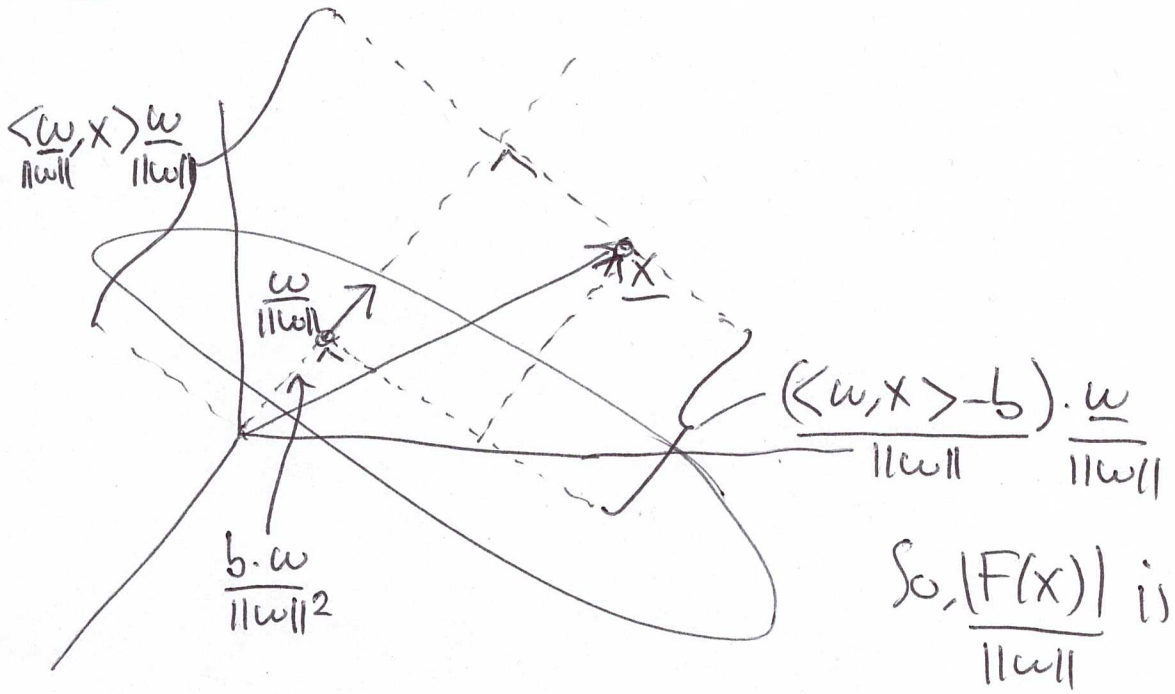
After rescaling ; Some x_i will meet equality ⇒ support vectors

$= |F(x_i)|$

Distance x_i to hyperplane : $\frac{|F(x_i)|}{\|w\|}$

Margin $= \frac{1}{\|w\|}$

Interme220:



So, $\frac{|F(x)|}{\|w\|}$ is a proper distance measure.

Training problem:

Primal $\begin{cases} \text{minimize} & \Phi(w) = \frac{1}{2}\|w\|^2 \quad \text{(for convenience)} \\ \text{s.t.} & y_i(\langle w, x_i\rangle - b) \geq 1 \quad \forall (x_i, y_i) \in I \end{cases}$

* Φ is convex in w
* constraints linear in w

Lagrange multipliers:

$$J(w,b,\alpha) = \frac{1}{2}\langle w,w\rangle - \sum_{i=1}^{m} \alpha_i \{ y_i(\langle w, x_i\rangle - b) - 1 \}$$

Minimize w.r.t. w and b $\Leftrightarrow$ $\frac{\partial J}{\partial w} = 0$ & $\frac{\partial J}{\partial b} = 0$ (J convex)

$\Updownarrow$ $\Updownarrow$

$w = \sum \alpha_i y_i x_i$ (a) $\quad \sum \alpha_i y_i = 0$ (b)

Maximize w.r.t. $\alpha (\geq 0)$

Dual Problem

* Same optimal value as primal problem
* α_i's provide optimal solution

$$J(w,b,\alpha) = \underbrace{\frac{1}{2}\langle w,w\rangle - \sum \alpha_i y_i \langle w, x_i\rangle}_{} \underbrace{- b\sum \alpha_i y_i}_{=0 \text{ by } (b)} + \sum \alpha_i$$

$$\langle w,w\rangle \overset{\text{by } (a)}{=} \sum \alpha_i y_i \langle w, x_i\rangle = \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

(by (a))

$$J(w,b,\alpha) = Q(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

Dual
$$\begin{cases} \text{maximize } Q(\alpha) \\ \text{s.t. } \sum \alpha_i y_i = 0,\ \alpha \geq 0 \end{cases}$$

* Just training data !!
* $Q(\alpha)$ only depends on $\langle x_i, x_j\rangle$!!

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

Kuhn-Tucker $\rightsquigarrow$ x_i is a support vector iff $\alpha_i^* \neq 0$

$$\downarrow$$

$$b^* = 1 - \langle w^*, x_i\rangle$$

Not linearly separable

Allow mislabeled data → use slack variables ξ_i

measure degree of misclassification

Primal

Minimize $Q(w, b, \xi) = \frac{1}{2}\|w\|^2 + C \cdot \sum \xi_i$

C: Trade-off:
- margin size
- amount of error in training

s.t. $y_i(\langle w, x_i \rangle - b) \geq 1 - \xi_i \quad \forall (x_i, y_i) \in D$

$\xi_i \geq 0$

Dual

Maximize $Q(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$ (mistake in book)

s.t. $\sum_i \alpha_i y_i = 0, \quad C \geq \alpha > 0$

(In the proof: use Lagrange variables μ_i to enforce non-negativity of slack variables ξ_i

$\xi_i = 0$ if $\alpha_i < C$

$\xi_i \geq 0$ if $\alpha_i = C$)

Slides [6-11] + [15-21] tutorial

Kernel trick

Non-Linear SVMs: 1) $x_i \rightarrow \varphi(x_i)$ in higher dim. space

2) $\{(\varphi(x_i), y_i)\}$ can be linearly separated

$$Q(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \underbrace{\langle \varphi(x_i), \varphi(x_j) \rangle}_{K(x_i, x_j)}$$

$$\exists_\varphi \; K(u,v) = \langle \varphi(u), \varphi(v) \rangle$$

iff

$$\forall_\psi \left[\int \psi(x)^2 dx \leq 0 \Rightarrow \int K(u,v) \psi(u) \psi(v) \, du \, dv \leq 0 \right]$$

We do not need to know φ

We do not need to compute a large inner product !!

$$F(x) = \sum_i \alpha_i^* y_i K(x_i, x) - b$$

Slides [26-28] tutorial

SVM Regression : estimate a function that

maps input objects x_i ($\in \mathbb{R}^n$)

$\rightarrow$ real number y_i ($\in \mathbb{R}$, not necessarily in $\{\pm 1\}$)

based on training data

$$F(x) = \langle w, x \rangle - b$$

How to best fit the data?

$|y_i - F(x_i)| \leq \varepsilon$, margin $\frac{1}{\|w\|}$

$$\begin{cases} \text{Minimize } L(w) = \frac{1}{2}\|w\|^2 \\ \text{s.t. } |y_i - F(x_i)| \leq \varepsilon \quad \forall (x_i, y_i) \in D \end{cases}$$

Allow noise in training data

$$\text{Minimize } L(w, \xi, \hat{\xi}) = \frac{1}{2}\|w\|^2 + C \cdot \sum_i (\xi_i^2 + \hat{\xi}_i^2)$$

$$\text{s.t. } y_i - (\langle w, x_i \rangle - b) \leq \varepsilon + \xi_i$$

$$-y_i + (\langle w, x_i \rangle - b) \leq \varepsilon + \hat{\xi}_i$$

$$\xi_i, \hat{\xi}_i \geq 0$$

etc.

SVM Ranking : learning a ranking/preference function

$\{(x_i, y_i)\}$: x_i is preferred to x_j $(x_i \succ x_j)$

if

$y_i < y_j$

Output score (not class) : $F(x_i) > F(x_j)$ for any $x_i \succ x_j$

Linear ranking function $\equiv$ $\langle w, x_i \rangle > \langle w, x_j \rangle$

Approximate solution using slack variables:

minimize $\frac{1}{2}\langle w, w \rangle + C \sum \xi_{ij}$

s.t. $\forall_{y_i < y_j}$ $\langle w, x_i \rangle \geq \langle w, x_j \rangle + 1 - \xi_{ij}$

$\xi_{ij} \geq 0$

$\Leftarrow$ (dual)

$$F(z) = \sum_{ij} \alpha_{ij}^* K(x_i - x_j, z)$$

1-Norm ranking SVM

$$F(x_u) > F(x_v) \Rightarrow \sum_{(i,j)\in P} \alpha_{ij} \langle (x_i - x_j), x_u \rangle > \sum_{(i,j)\in P} \alpha_{ij} \langle (x_i - x_j), x_v \rangle$$

$(u,v) \in P$

$$P = \{(i,j) : y_i < y_j\}$$

$$L(\alpha, \xi) = \sum_{(i,j)\in P} \alpha_{ij} + C \sum_{(i,j)\in P} \xi_{ij}$$

$$\forall_{y_u < y_v} \quad \sum_{(i,j)\in P} \alpha_{ij} k((x_i - x_j), (x_u - x_v)) \geq 1 - \xi_{uv}$$

$$\alpha_{ij} \geq 0, \ \xi_{u,v} \geq 0$$
