Lecture 6 Support Vector Machines

< Continue >

1. Laplace smoothing
   - problems in Naive Bayes:

$$P(y=1 \mid x_j) = \frac{P(x_j \mid y=1) \cdot P(y=1)}{P(x_j \mid y=1) \cdot P(y=1) + P(x_j \mid y=0) \cdot P(y=0)}$$

   If x_j never appears: $P(y=1 \mid x_j) = 0/0+0$

   - Solution: # "0"s +1 ; # "1"s +1

Laplace Smoothing: $\Phi_{j|y=0} = \frac{\sum_{i=1}^{m} I\{x_j^{(i)}=1, y^{(i)}=0\}+1}{\sum_{i=1}^{m} I\{y^{(i)}=0\}+2}$

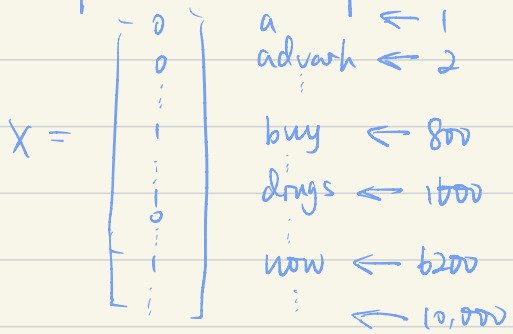
   - When multiple features: $x_i \in \{1, 2, \cdots k\}$

$$P(x \mid y) = \prod_{j=1}^{k} P(x_j \mid y)$$

2. Multinomial Bernoulli Event Model

Multinomial Event Model

Previous example:



$\Rightarrow$

New representation:

$$x = \begin{bmatrix} 1600 \\ 800 \\ 1600 \\ 6200 \end{bmatrix} \in \mathbb{R}^{n_i}$$

$x_j \in \{1, \dots 10000\}$

n_i = length of email i

email: "Drug! Buy drugs now!"

$$P(x,y) = P(x|y)P(y) \overset{\text{assume}}{=\!=} \prod_{j=1}^{n} P(x_j|y) \cdot P(y)$$

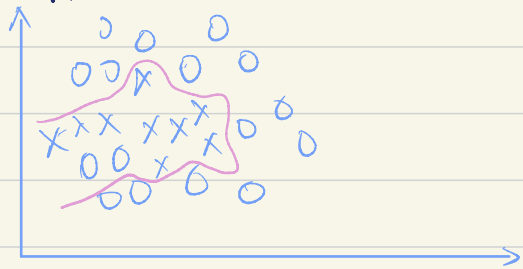
parameters: $\phi_y = P(y=1)$, $\phi_{k|y=0} = P(x_j = k | y=0)$

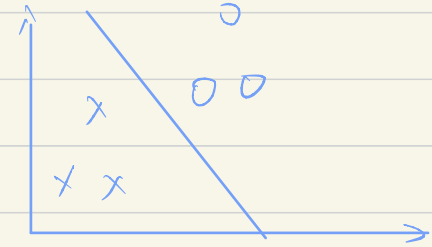
$\phi_{k|y=1} = P(x_j = k | y=1)$ ⇓ chance of word j being k if $y=0$

MLE:

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m} I\{y^{(i)}=0\} \cdot \sum_{j=1}^{n_i} I\{x_j^{(i)}=k\} + 1}{\sum_{i=1}^{m} I\{y^{(i)}=0\} \cdot n_i + 10{,}000}$$

3. Support Vector Machines





- Optimal Margin Classifier (separate case)
  1) Functional margin

     $h_\theta(x) = g(\theta^T x)$

     Predict "1" if $\theta^T x \gg 0$; "0" otherwise

     If $y^{(i)}=1$, hope that $\theta^T x^{(i)} \gg 0$; If $y^{(i)}=0$, hope that $\theta^T x^{(i)} \ll 0$

     <A New Notation for easier discussion of SVM>

     Notation: Labels $y \in \{-1, 1\}$; Have h output values in $\{-1, +1\}$

$g(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{otherwise} \end{cases}$

previously: $h_\theta(x) = g(\theta^T x)$ $\theta \in \mathbb{R}^{n+1}$, $x_0 = 1$

now: $h_{w,b}(x) = g(w^T x + b)$, drop $x_0 = 1$ convention

$\sum_{i=1}^{n} w_i x_i + b$ $\quad w \in \mathbb{R}^n$, $b \in \mathbb{R}$

Functional margin of hyperplane defined by (w,b) w.r.t. $(x^{(i)}, y^{(i)})$

$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$ ⇐ If $y^{(i)} = 1$, want $w^T x^{(i)} + b >> 0$ ⇒ want $\hat{\gamma}^{(i)} >> 0$

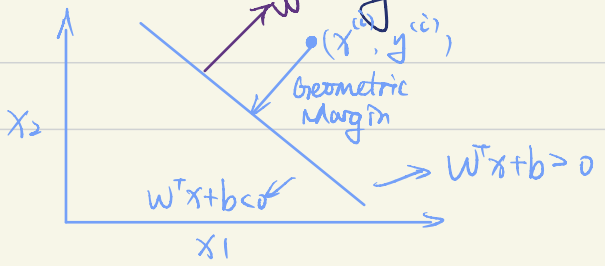
If $y^{(i)} = -1$, want $w^T x^{(i)} + b << 0$

If $\hat{\gamma}^{(i)} > 0$, that means $h(x^{(i)}) = y^{(i)}$

Functional Margin w.r.t training set $\hat{\gamma} = \min \hat{\gamma}^{(i)}$ $(i = 1, 2, \cdots m)$ ← worst trang example

$(w, b) \to \left(\frac{w}{\|w\|}, \frac{b}{\|b\|}\right)$ (normalization, because $g(w^T x + b) = g(2w^T x + 2b)$)

2) Geometric margin



$\gamma^{(i)} = \frac{(w^T x^{(i)} + b) y^{(i)}}{\|w\|}$

$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|w\|}$ ← Functional Margin

↑ Geometric Margin

← when $\|w\| = 1$, functional margin = geometric margin

Geometric Margin w.r.t. training set: $\gamma = \min_i \gamma^{(i)}$

3) Optimal Margin Classifier:

Choose w, b to maximize γ

$$\max_{\gamma, w, b} \gamma \quad \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b)/\|w\| \geq \gamma, \quad i = 1, 2, \cdots m$$

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

$$\Rightarrow \min_{w,b} \|w\|^2 \quad \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$$