Outline

- Naive Bayes
  - Laplace Smoothing
  - Event Models
- Comment on applied ML
- Kernel Methods

Recap: Spam filter Naive Bayes

$$X = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} \begin{array}{l} \text{a} \\ \text{add} \\ \text{bee} \\ \\ \\ \\ \end{array} \qquad x_j = \mathbb{1}\{\text{work } j \text{ appears in email}\}$$

Generative Model

$P(x|y) \qquad P(y)$

$$P(x|y) = \prod_{i=1}^{d} P(x_i|y)$$

Parameters:

$P(y=1) = \phi_y$

$P(x_j=1|y=0) = \phi_{j|y=0}$

$P(x_j=1|y=0) = \phi_{j|y=1}$

Max Likelihood estimates

$$\phi_y = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\}}{n}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)}=1, y^{(i)}=0\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\}} \qquad \phi_{j=1|y=1} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)}=1, y^{(i)}=1\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\}}$$

Prediction time

$$P(y=1|x) = \frac{P(x|y=1)\cdot P(y=1)}{\left(P(x|y=1)P(y=1) + P(x|y=0)P(y=0)\right)} \rightarrow P(x)$$

If word x_{5500} never occurs.

$$P(x_{5500}=1 \mid y=1) = \frac{0}{\#\{y=1\}} = 0 = \phi_{5500 \mid y=1}$$

$$P(x_{5500}=1 \mid y=0) = \frac{0}{\#\{y=0\}} = 0 = \phi_{5500 \mid y=1}$$

$$P(x \mid y=1) = \prod_{j=1}^{10000} P(x_j \mid y=1) = 0$$

$$P(x \mid y=0) = \prod P(x_j \mid y=0) = 0$$

$$\Rightarrow \quad P(y=1 \mid x) = \frac{0}{0+0}$$

Laplace Smoothing

\# 1's +1

\# 0's +1

$x \in \{1, \dots, k\}$

Estimate $P(x=j) = \dfrac{\sum_{i=1}^{n} \mathbb{1}\{x^{(i)} = j\} + 1}{n + k}$

space size

$$\sum_{j=1}^{n} P(x=j)$$

- Back to Naive Bayes

$$\phi_{j \mid y=0} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_j^{(i)}=1, y=0\} + 1}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\} + 2}$$

$x_i \in \{1, \dots, k\}$

size x $\langle$ 400 feet^2 (1), 400–800 (2), 800–1200 (3), >1200 (4) $\rangle$ Discrete variable.

$$P(x \mid y) = \prod_{i=1}^{d} \underbrace{P(x_j \mid y)}_{\text{Multinomial (v.s. bernoulli)}}$$

account bank account ...

800 1600 800

$$x = \begin{bmatrix} 800 \\ 1600 \\ 800 \\ \\ \end{bmatrix} \in \mathbb{R}^{d_i}, \quad d_i = \text{length of email } i, \quad x_j \in \{1, \dots, \overset{\text{dict size}}{10{,}000}\}$$

So far: Multivariate Bernoulli event model

New: Multinomial event model

Generative model

$$P(x,y) = P(x \mid y) \cdot P(y)$$

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

$P(x \mid y), P(y)$ generative $\qquad P(y \mid x)$. discriminative

Parameters

$$\phi_y = P(y=1)$$

$$\phi_{k|y=0} = P(x_j = k \mid y=0\}$$

Chance of word j being k if $y=0$

Assume that does not depend on j (independent of j)

$$\phi_{k|y=1} = P(x_j = k \mid y=1\}$$

MLE

$$\phi_{k|y=0} = \frac{\overbrace{\left(\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\} \sum_{j=1}^{d_i} \mathbb{1}\{x_j^{(i)}=k\}\right)}^{\#\text{ of }\{x_j = k \text{ when } y=0\}} + 1}{\underbrace{\left(\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\} \cdot d_i\right)}_{\#\text{ of }\{\text{words, when } y=0\}} + 10{,}000}$$

(Laplace Smoothing)

x_j : index of the j-th word in the email

Original.

$$P(x|y) = \prod_{i=1}^{d} P(x_j \mid y) = P(x_1=1|y)\,P(x_2=0|y)\,P(x_3=1|y=0)\cdots \quad \begin{array}{c} x \\ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{array}$$

New model

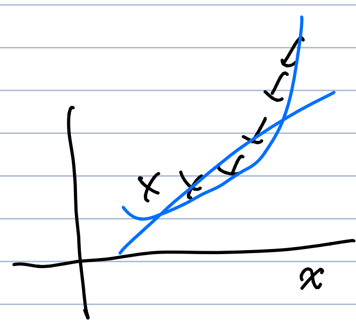
$$P(x|y) = \prod_{j=1}^{d_i} P(x_j \mid y)$$

Kernel Methods $\rightarrow$ SVM

linear fn.

quadratic fn.

cubic fn.



$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \quad \text{(feature mapping)}$$

$$h_\theta(x) = [\theta_0 \ \theta_1 \ \theta_2 \ \theta_3] \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} = \theta^T \phi(x)$$

$(x^{(1)}, y^{(1)}), \dots \quad (x^{(n)}, y^{(n)})$

$\Downarrow$

$(\phi(x^{(1)}), y^{(1)}), \dots, (\phi(x^{(n)}), y^{(n)})$

Recall: linear regression

$\theta : 0$

Loop: $\underset{\mathbb{R}^d}{\theta} := \theta + \alpha \sum \underbrace{(y^{(i)} - \theta^T x^{(i)})}_{\text{scalar}} \underset{\mathbb{R}^d}{x^{(i)}}$

New data set (after feature map)

$\theta : 0$

Loop: $\theta := \overset{\mathbb{R}^p}{\theta} + \sum_{i=1}^{n} \underbrace{(y^{(i)} - \overset{\mathbb{R}^p}{\theta^T} \overset{\mathbb{R}^p}{\phi(x^{(i)})})}_{\text{scaler}} \underset{\mathbb{R}^p}{\phi(x^{(i)})}$ $\qquad O(np)$

$\phi : \mathbb{R}^d \to \mathbb{R}^p$

Terminology

ϕ : feature map

$\phi(x)$: (new) feature

x: attributes

Kernel Method

$d > 1. \qquad x = (x_1, \dots x_d)$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_1 x_1 \\ x_1 x_2 \\ x_1 x_1 x_1 \\ \vdots \\ x_d x_d x_d \end{bmatrix} \begin{array}{l} - 1 \\ \} d \\ \\ \} d^2 \\ \\ \} d^3 \end{array}$$

$\theta^T \phi(x)$ can represent any degree 3 poly in $x_1 \dots x_d$

$p = 1 + d + d^2 + d^3$

Suppose $d = 1000$, $\quad p \cong 10^9$ Bad!

Time per iteration $\xrightarrow[\text{Method}]{\text{Kernel}}$ Improve to $O(n^2)$

$O(np) \quad p \sim d^3$ $\qquad$ Even θ takes $O(p)$

$p \gg n$