Outline

Naive Bayes

- Laplace Smoothing
- Event models

Comments on applied ML

Kernel Methods

Recap:

$$X = \begin{bmatrix} 0 \\ 1 \\ \\ \\ \\ \\ \end{bmatrix} \begin{matrix} a \\ \text{aardvark} \\ \\ \text{buy} \\ \vdots \\ \vdots \end{matrix} \qquad X_j = \mathbb{1}\{\text{word } j \text{ appears in email}\}$$

Generative Model

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

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

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

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

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

Max Likelihood estimates

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

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

$\phi_{j|y=1}$

At prediction time

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

NeurIPS

$j = 5500$

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

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

$$P(x|y=1) = \prod_{j=1}^{10,000} P(x_j|y=1) \longrightarrow \phi_{5500|y=1}$$

$$P(x|y=0) = \prod_{j=1}^{10,000} P(x_j|y=0) \longrightarrow \phi_{5500|y=1}$$

Laplace Smoothing

		Won?
2009 9/12	Wakeforest	0
10/10	OSU	0
10/17	Arizona	0
11/21	Caltech	0
12/31	Oklahoma	0

$$P(X=1) = \frac{\#1's}{\#1's + \#0's} \qquad +1 \quad +2$$

$$= \frac{0}{0+4} \qquad \frac{1}{6}$$

$$= 0 \qquad = \frac{1}{6}$$

Laplace Smoothing. $\#1's + 1$, $\#0's + 1$

More generally $X \in \{1 \ldots k\}$

Estimate

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

$$\sum_j P(X=j) = 1$$

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

$P(X_j = 1 \mid y = 0)$
$+ P(X_j = 0 \mid y = 0) = 1$

$X_i \in \{1 \ldots k\}$

size	< 400 feet^2	400–800	800–1200	>1200
X	1	2	3	4

$$P(x \mid y) = \prod_{i=1}^{d} P(x_j \mid y)$$

multinomial (vs. bernoulli)

Is $P(X_j = 1) + P(X_j = 0) = 1$

$P(X_j = 1 \mid y = 0) + P(X_j = 0 \mid y = 0) = 1$

$$X = \begin{bmatrix} 0 \\ 1 \\ 1 \\ \\ \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \\ \text{account} \\ \text{bank} \\ \text{beneficiary} \end{matrix}$$

$X_i \in \{0, 1\}$

account bank account ....

800 1600 800

$$X = \begin{bmatrix} 800 \\ 1600 \\ 800 \\ \vdots \\ \vdots \end{bmatrix} \in \mathbb{R}^{d_i} \qquad d_i = \text{length of email } i$$

$$X_j \in \{1, \dots 10{,}000\}$$

So far: Multivariate Bernoulli event model

New: Multinomial event model

Generative model

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

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

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 this does not depend on j

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

MLE

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} \sum_{j=1}^{d_i} 1\{x_j^{(i)} = k\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} \cdot d_i}$$

$+1$ / $+10{,}000$

Laplace Smoothing

x_j = index of j^{th} word in email

↓

position in the dictionary

map Rare words to UNK

mortgage mφrtgage → UNK

$\phi_{k|y=1} = p(j^{th} \text{ word} = k \mid y = 1)$

↓ independent of j

$$p(x|y) = \prod_{i=1}^{d} p(x_j|y)$$

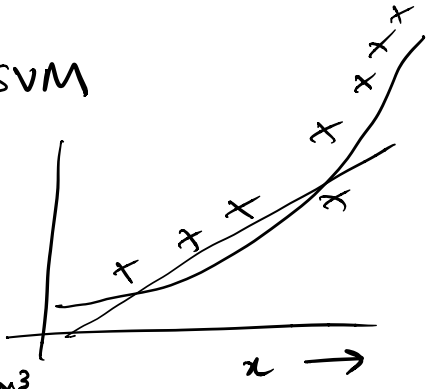
$$= p(x_1=1|y) \cdot p(x_2=0|y) \cdot p(x_3=1|y) \cdots$$

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

$$p(x|y) = \prod_{i=1}^{d_i} p(x_j|y)$$

Kernel Methods → SVM

linear fn
quadratic fn
cubic fn of data



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

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \qquad \phi : \mathbb{R} \to \mathbb{R}^4$$

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

$$= \theta^T \phi(x)$$

$(x^{(1)}, y^{(1)}) \; \cdots \; (x^{(n)}, y^{(n)})$

$\Downarrow$

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

Recall: linear regression

$\theta = 0$

Loop $\theta := \theta + \alpha \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)}) x^{(i)}$

$\theta \in \mathbb{R}^d$; $(y^{(i)} - \theta^T x^{(i)})$ scalar; $x^{(i)} \in \mathbb{R}^d$

New data set

$\theta = 0$

Loop $\theta := \theta + \alpha \sum_{i=1}^{n} (y^{(i)} - \theta^T \phi(x^{(i)})) \phi(x^{(i)})$

$\theta \in \mathbb{R}^p$; $\theta^T \phi(x^{(i)})$ scalar; $\phi(x^{(i)}) \in \mathbb{R}^p$

$\phi : \mathbb{R}^d \to \mathbb{R}^p$ ($d=1$, $p=4$ in example)

Terminology

ϕ: feature map

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

x: attributes

Kernel methods:

$d > 1 \qquad x = (x_1 \cdots x_d)$

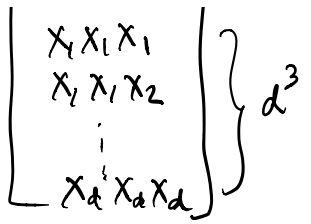
$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_d x_d \end{bmatrix} \begin{matrix} \} 1 \\ \} d \\ \} d^2 \end{matrix}$$

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

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

Suppose $d = 1000$

$p \cong 10^9$



time per iteration $O(np)$ $\quad p \sim d^3$

Improve this to $O(n^2)$ per iteration

Even storing Θ takes time $O(p)$

$p \gg n$