Outline
Naive Bayes
- Laplace smoothing
- Event models

Kernel methods

Recap:

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} a \\ aardvark \\ \vdots \\ buy \end{matrix} \quad \uparrow d \downarrow \quad n \text{ examples}$$

$x_j = \mathbb{1}\{\text{word } j \text{ appears in email}\}$

Generative Model

$p(x|y)$ $\quad$ $p(y)$

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

$y = \begin{cases} 0 & \text{not spam} \\ 1 & \text{spam} \end{cases}$

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

$P(x_j = 1 | y=0) = \phi_{j|y=0}$ $\quad$ non-spam

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

Maximum Likelihood Estimates:

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

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

$x \in \mathbb{R}^d$

Prediction: $\phi_{j|y=1}$

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

$\phi_{j|y=0}$

COVID

$j = 1273$

$$P(X_{1273} = 1 \mid y = 1) = \frac{0}{\#\{y=1\}} = \phi_{1273|y=1}$$

$$P(X_{1273} = 1 \mid y = 0) = \frac{0}{\#\{y=0\}} = \phi_{1273|y=0}$$

$$P(x \mid y=1) = \prod_{j=1}^{10{,}000} P(x_j \mid y)$$

$= 0$ $\quad \phi_{1273|y=1}$

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

"0" $\quad$ "0" $\quad \phi_{1273|y=0}$

Laplace Smoothing

	Won?
Wakeforest	0
OSU	0
Arizona	0
Caltech	0

Oklohama

$$P(X=1) = \frac{\#\text{"1"s}}{\#\text{"1"s} + \#\text{"0"s}} \quad \begin{matrix} +1 \\ +2 \end{matrix}$$

$$= \frac{0 \overset{+1}{\rightarrow}}{0+4 \overset{+2}{\rightarrow}} = \frac{1}{6}$$

Laplace Smoothing

More generally

$X \in \{1 \dots k\}$

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

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

i: examples j: words

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

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

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

multinomial (vs. bernoulli)

$X =$

	a	1
	aardvark	2
1	account	800
1	bank	1600
1	beneficcary	

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

"_ _ bank account bank"

New representation:

$$X = \begin{bmatrix} 1600 \\ 800 \\ 1600 \\ \vdots \\ \vdots \end{bmatrix} \in \mathbb{R}^{d_i}$$

$x_j \in \{1, \cdots 10{,}000\}$

d_i : length of email i

Multivariate Bernoulli event model

Multinomial event model

$$p(x, y) = p(x|y)\, p(y)$$

$$p(x|y) \overset{\text{assume}}{=} \prod_{j=1}^{d} p(x_j|y)$$

Parameters:

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

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

Chance that word j is k^{th} word in dictionary if $y=0$

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

MLE

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

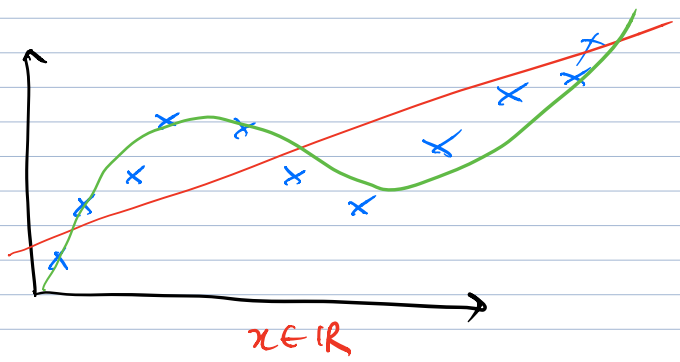
Laplace smoothing numerator : +1
denom : +|dictionary|
+ 10,000

Map rare words to "UNK"

Spam detection
mortgage
mørtgʌge
↳ "UNK"
spoofed headers
fetch URL

Kernel Methods

$\theta^T x$



$h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x_1 + \theta_0$

$\phi : \mathbb{R} \longrightarrow \mathbb{R}^4$

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

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

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$h_\theta(x)$ is linear in θ, $\phi(x)$

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \ldots \quad (x^{(n)}, y^{(n)})\}$$

$$\Downarrow$$

$$\{(\phi(x^{(1)}), y^{(1)}), (\phi(x^{(2)}), y^{(2)}) \ldots \}$$

LMS on new dataset

$$\min_\theta \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)}) \right)^2$$

Gradient Descent:

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

$\theta \in \mathbb{R}^p$ $\quad \phi(x^{(i)}) \in \mathbb{R}^p$ $\quad O(np)$

}

Terminology:

$\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^p$ feature map

attributes $\quad$ features

x : attributes

$\phi(x)$: "features"

How do we handle large values of p?

$(x_1 \dots x_d)$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1^2 \\ \vdots \\ x_i x_j \\ \vdots \\ x_d^2 \\ x_1^3 \\ \vdots \\ x_i x_j x_k \\ \vdots \\ x_d^3 \end{bmatrix} \begin{matrix} \\ \} d \\ \\ \} d^2 \\ \\ \\ \\ \} d^3 \\ \\ \\ \end{matrix} \quad x_i x_j \quad x_j x_i$$

$$\theta^T \cdot \phi(x) = \_\_ \cdot 1 + \_\_ x_1 + \_\_ x_2 \dots + \_\_ x_i x_j$$

$\phi(x)$ is high dimensional!

$p = 1 + d + d^2 + d^3 \qquad O(d^3)$

$d = 10^3, \qquad p \approx 10^9$

Runtime for G.D. depends on $p \qquad O(np)$

Key observation:

If θ is initialized at 0

then at any time

θ can be written as

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \quad \text{for some } \beta_1 \dots \beta_n \in \mathbb{R}$$

$\in \mathbb{R}^p$ $\qquad$ $\in \mathbb{R}^n$

Proof of this observation:

By induction over time

Base Case: Iteration 0

$$\theta = 0 = \sum_{i=1}^{n} 0 \cdot \phi(x^{(i)})$$

($0 \cdot \phi(x^{(i)})$: 0 ↑ β_i)

Assume at some time t

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$$

Next iteration:

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

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

($y^{(i)}$: scalar; $\theta^T \phi(x^{(i)})$: scalar; $\beta_i + \alpha (y^{(i)} - \theta^T \phi(x^{(i)}))$: new β_i)

new algo : update β