

Outline

- Naive Bayes
- Laplace smoothing
- Event models

Kernel methods

Recap:

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \quad \begin{array}{l} \text{a} \\ \text{aardvark} \\ \vdots \\ \text{buy} \end{array}$$

d

n examples

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

Generative Model

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

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

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

<https://powcoder.com>

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

$P(x_j=1|y=0) = \phi_{j|y=0}$ not spam

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

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

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

COVID

$j = 1273$

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

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

$$P(x | y=1) = \sum_{j=1}^{10,000} P(x_j | y) = \phi_{1273 | y=1}$$

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

<https://powcoder.com>

Laplace Smoothing

Add WeChat powcoder

Wakeforest

0

OSU

0

Arizona

0

Caltech

0

Oklahoma

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

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

Laplace Smoothing

More generally

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

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

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

i: examples

j: words

Assignment Project Exam Help

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

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

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

multinomial (vs. bernoulli)

$$X = \begin{bmatrix} & a \\ 1 & aardvark \\ 1 & account \\ 1 & bank \\ 1 & beneficiary \end{bmatrix} \quad \begin{array}{c} \frac{1}{2} \\ 800 \\ 1600 \end{array}$$

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

"~ bank account bank"

New representation:

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

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

d_i : length of email i

Multivariate Bernoulli event model

Multinomial event model

Assignment Project Exam Help

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

Add WeChat powcoder

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

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

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

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

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

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

map rare words to "UNK"

Spam detection

mortgage

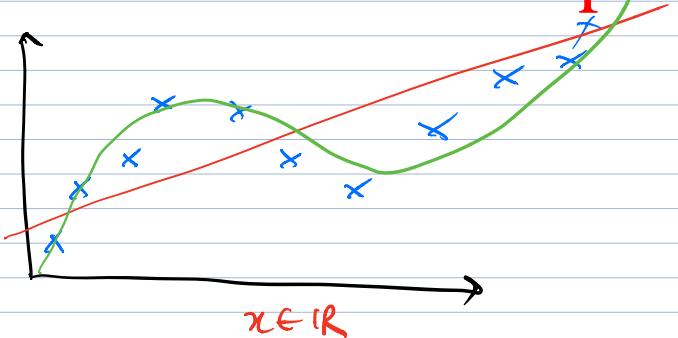
↙
mørtgøage
"UNK"

spoofed headers

Assignment Project Exam Help
 fetch URL

Kernel <https://powcoder.com>

$\theta^T x$ Add WeChat powcoder



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

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

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

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

$$\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \\ \Theta_3 \end{bmatrix}$$

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

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

Assignment Project Exam Help

LMS <https://powcoder.com> On new dataset

$$\min_{\theta} \frac{1}{2} \sum_i^n (y^{(i)} - \theta^\top \phi(x^{(i)}))^2$$

Gradient Descent :

$$\text{Loop } \{ \quad \theta := \theta + \alpha \sum_{i=1}^n (y^{(i)} - \theta^T \phi(x^{(i)})) \phi(x^{(i)}) \quad \in \mathbb{R}^p \quad \in \mathbb{R}^p \quad O(n p) \}$$

Terminology:

α : Attributer

$\phi(x)$: "features"

How do we handle large values of p?

$(x_1 \dots x_d)$

$$\phi(x) = \begin{bmatrix} x_1^0 \\ x_1^1 \\ x_d^0 \\ x_1^2 \\ \vdots \\ x_i x_j \\ \vdots \\ x_d^2 \\ x_1^3 \\ \vdots \\ x_i x_j x_k \\ \vdots \\ x_d^3 \end{bmatrix}$$

Annotations: x_1^0 is labeled d^0 , x_1^1 is labeled d^1 , x_d^0 is labeled d^0 , x_1^2 is labeled d^2 , $x_i x_j$ is labeled d^2 , x_d^2 is labeled d^2 , x_1^3 is labeled d^3 , and x_d^3 is labeled d^3 .

$$\theta^T \cdot \phi(x) = _\cdot 1 + _x_1 + _x_2 + \dots + _x_i x_j$$

Assignment Project Exam Help

<https://powcoder.com>
 $\phi(x)$ is high dimensional!

p = Add WeChat powcoder

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

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

Key observation:

If θ is initialized at 0

then at any time

θ can be written as

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

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

Proof of this observation:

By induction over time

Base Case: Iteration 0

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

Assume at some time t

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

Next iteration:

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

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

Add WeChat ^{scalar} _{new β_i} powcoder

new algo : update β