Naive Bayes

Feature vector x (ex. spam classification)

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \\ \\ \end{bmatrix} \begin{array}{l} \text{a} \\ \text{aardvark} \\ \vdots \\ \\ \\ \text{zymurgy} \end{array} \quad \Bigg\updownarrow \text{ top-}d \text{ words}$$

$x \in \{0,1\}^d$

$x_i : \mathbb{1}\{\text{word } i \text{ in email}\}$

Want to model $p(x|y)\,p(y)$

2^d possible values of x

Assume x_i's are conditionally independent given y

$$\Rightarrow p(x_1, \dots x_d \mid y) = p(x_1|y)\,p(x_2|x_1,y)\cdots p(x_d|x_{d-1},\dots x_1, y)$$

$$\overset{\text{assume}}{=} p(x_1|y)\,p(x_2|y)\cdots p(x_d|y)$$

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

Naive Bayes assumption

Parameters

$\phi_{j|y=1} = p(x_j=1|y=1)$ if it is a spam

$\phi_{j|y=0} = p(x_j=1|y=0)$ if it is not spam

$\phi_y = p(y)$ $\quad$ Pr(spam)

Joint Likelihood

$$\mathcal{L}(\phi_y, \phi_{j|y}) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi_y, \phi_{j|y})$$

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

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

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

Prediction

$$p(y=1|x) = \frac{\overset{\phi_{j|y=1}}{p(x|y=1)}\,p(y=1)}{\underset{\phi_{j|y=1}}{p(x|y=1)}\,p(y=1) + \underset{\phi_{j|y=0}}{p(x|y=0)}\,p(y=0)}$$

With new word (i.e. COVID) word k

$$p(x_k=1|y=1) = \frac{0}{\#\{y=1\}} = \phi_{k|y=1}$$

$$p(x_k=1|y=0) = \frac{0}{\#\{y=0\}} = \phi_{k|y=0}$$

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

$$p(y=1|x) = \frac{\overset{0:\ \text{never seen}}{p(x|y=1)}\,p(y=1)}{\underset{0}{p(x|y=1)}\,p(y=1) + \underset{0}{p(x|y=0)}\,p(y=0)}$$

Laplace Smoothing

Add 1 to count of 1s

Add 1 to count of 0s

Multivariate Bernoulli Event Model (Multinomial Event Model)

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

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

no longer binary

$x_j \in \{1 \dots |V|\}$

ex.

$$V = \begin{bmatrix} \text{a} & 1 \\ \text{aardvark} & 2 \\ \vdots & \\ \text{account} & 800 \\ \vdots & \\ \text{bank} & 1600 \\ \vdots & \end{bmatrix}$$

$$x = \begin{bmatrix} 1600 \\ \vdots \\ 800 \\ \vdots \\ 1600 \\ \vdots \\ 6200 \end{bmatrix} \in \mathbb{N}^d$$

Parameters

$\phi_y = p(y=1)$

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

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

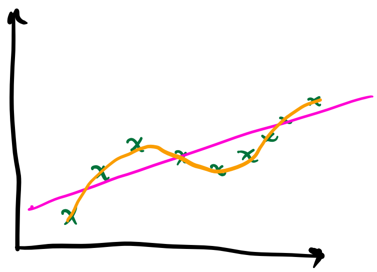
MLE

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

Laplace Smoothing: +1 to numerator
+|V| to denominator

Map rare words to UNK

Kernel Methods



Linear methods: $\theta^T x$

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

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

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

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

$h_\theta(x)$: linear in θ, $\phi(x)$

Dataset

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

Cubic polynomial on old dataset,
linear on new dataset

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:

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

Terminology

$$\phi : \underset{\substack{\uparrow \\ \text{attributes} \\ x}}{\mathbb{R}^d} \rightarrow \underset{\substack{\uparrow \\ \text{features} \\ \phi(x)}}{\mathbb{R}^p}$$ feature map

What to do if p is very large?

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

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

Problem: $\phi(x)$ is high dimensional

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

$d \sim 10^3 \Rightarrow p \sim 10^9$

Runtime for 1 iteration of GD is $O(np)$

Key observation:

If θ initialized as 0, then at any time,

$$\underset{\in \mathbb{R}^p}{\theta} = \sum_{i=1}^{n} \beta_i \, \phi(x^{(i)}) \text{ for } \underset{\in \mathbb{R}^n}{\beta_1 \ldots \beta_n \in \mathbb{R}}$$