Lecture 6 : 2/6/17

Naive Bayes Model

$$D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$$

Bernoulli Naive Model:

$$\Theta = \begin{bmatrix} \Theta_{10} & \Theta_{11} \\ \vdots & \vdots \\ \Theta_{M0} & \Theta_{M1} \end{bmatrix}$$

$$\begin{aligned} p(\vec{x}, y \mid \phi, \Theta) &= p(x_1, \ldots, x_M, y \mid \phi, \Theta) \\ &= p(y \mid \phi) \prod_{m=1}^{M} p(x_m \mid y, \Theta) \\ &= \left[(\phi)^{y}(1-\phi)^{(1-y)}\right] \prod_{m=1}^{M} (\Theta_{my})^{x_m} (1-\Theta_{my})^{(1-x_m)} \end{aligned}$$

Naives Bayes Assumption:

Recall: $p(\vec{x}, y) = p(\vec{x} \mid y)\, p(y)$

$$p(\vec{x} \mid y) = \prod_{m=1}^{M} p(x_m \mid y)$$

each x_q is conditionally independent of x_r given y $\forall q, r$

Def: two r.vs X, Y are cond. indep. given r.v. Z
written $X \perp Y \mid Z$
iff $P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$

Q: Why is this "naive"?

A: in real data.

Q: Why is it useful?

A: Count parameters:

Case #1: w/o NB assumption:

$p(x_1, \ldots, x_M \mid y) =$

x_1	x_2	$\ldots$	x_M	y	$p(\vec{x} \mid y)$
0	0	...	0	0	·
0	0	...	0	1	·
0	0	...	1	0	·
0	0	...	1	1	·

2^{M+1} rows

$2^{M+1} - 2$ params

Case #2 w/ NBA

$p(x_1 = 1 \mid y = 0) = \Theta_{10}$

$p(x_1 = 0 \mid y = 0) = 1 - \Theta_{10}$

$4M$ rows

$2M$ params

$p(x_1 \mid y) =$

x_1	y	$p(x_1 \mid y)$
0	0	·
0	1	·
1	0	·
1	1	

$p(x_2 \mid y) =$ $\ldots$ $p(x_M \mid y) =$

MLE for Naive Bayes

① Data Likelihood

$$\ell(\phi,\theta) = \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)} \mid \phi, \theta)$$

$\theta_{m0}, (1-\theta_{m0})$

$\theta_{m1}, (1-\theta_{m1})$

$$= \sum_{i=1}^{N} \left[\log p(y^{(i)} \mid \phi) + \sum_{m=1}^{M} \underline{\log p(x_m^{(i)} \mid y^{(i)}, \theta)} \right]$$

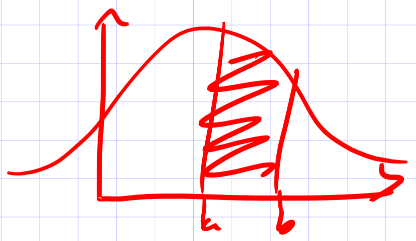
# times $y^{(i)}$ is 1 in D

$$= N_{y=1} \log \phi + N_{y=0} \log(1-\phi)$$

$$+ \sum_{m=1}^{M} \left[N_{x_m=1,y=1} \log \theta_{m1} + N_{x_m=0,y=1} \log(1-\theta_{m1}) \right]$$

# times that $x_m^{(i)}=1$ and $y^{(i)}=0$

$$+ \sum_{m=1}^{M} \left[N_{x_m=1,y=0} \log \theta_{m0} + N_{x_m=0,y=0} \log(1-\theta_{m0}) \right]$$



MLE for ϕ and θ:

$$(\hat{\phi}, \hat{\theta}) = \underset{\phi,\theta}{\operatorname{argmax}}\ \ell(\phi,\theta)$$

② Take partials wrt ϕ:

$$\frac{d\,\ell(\phi,\theta)}{d\phi} = \frac{N_{y=1}}{\phi} + \frac{N_{y=0}}{\phi-1}$$

we already know the MLE!

set equal to zero, solve for ϕ...

$$\phi^{MLE} = \frac{N_{y=1}}{N_{y=1}+N_{y=0}} = \frac{N_{y=1}}{N}$$

③ Take partials wrt θ_{my}:

Case where $y=0$, θ_{m0}

$$\frac{d\,\ell(\phi,\theta)}{d\theta_{m0}} = \frac{N_{x_m=1,y=0}}{\theta_{m0}} + \frac{N_{x_m=0,y=0}}{\theta_{m0}-1}$$

$$\Rightarrow \theta_{m0}^{MLE} = \frac{N_{x_m=1,y=0}}{N_{x_m=1,y=0}+N_{x_m=0,y=0}} = \frac{N_{x_m=1,y=0}}{N_{y=0}}$$

MAP Estimation for NB

The Problem w/MLE:

Suppose we never observe "Brexit" in Onion article

$\forall i$ where $y^{(i)} = \text{Onion}$, $x^{(i)}_{\text{Brexit}} = 0$

Q: What is the MLE $\theta_{\text{Brexit, Onion}} \overset{?}{=} 0$

$$p(y = \text{Onion} \mid \vec{x}^{(\text{new})}) = 0$$

↑ contains Brexit

Beta Priors:

$\phi \sim \text{Beta}(\alpha, \beta)$ ← not in HW

$\theta_{m,y} \sim \text{Beta}(\alpha, \beta) \quad \forall y \in \{0,1\} \quad \forall m \in \{1, \ldots, M\}$

for i in $1 \ldots N$

$y^{(i)} \sim \text{Bern}(\phi)$

$x_1^{(i)} \sim \text{Bern}(\quad)$

$\vdots$

$x_M^{(i)} \sim \text{Bern}(\quad)$

$$l_{MAP}(\phi, \theta) = \log\left[p(\phi, \theta \mid \alpha, \beta)\, p(D \mid \phi, \theta)\right]$$

$$= \log\left(\left[p(\phi \mid \alpha, \beta) \prod_{m=1}^{M} p(\theta_{m0} \mid \alpha, \beta)\, p(\theta_{m1} \mid \alpha, \beta)\right] \left[\prod_{i=1}^{N} p(x^{(i)}, y^{(i)} \mid \phi, \theta)\right]\right)$$

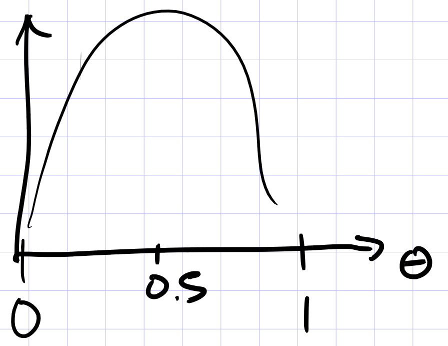
MAP Estimates:

$$(\hat{\phi}, \hat{\theta}) = \underset{\phi, \theta}{\text{argmax}}\; l_{MAP}(\phi, \theta)$$

Take partials, set equal to zero, and solve.

$$\theta_{m,0} = \frac{N_{x_m=1, y=0} + (\alpha - 1)}{N_{y=0} + (\alpha - 1) + (\beta - 1)}$$

$$1 - \theta_{m,0} = \frac{N_{x_m=0, y=0} + (\beta - 1)}{N_{y=0} + (\alpha - 1) + (\beta - 1)}$$



Hyperparameters α, β :

- Special case: where $\alpha = 2, \beta = 2$ called "Add-1 Smoothing"
- Chosen by hand, with world knowlede ~~with~~ about the model.
- Can learn them (Beyond Scope)

Gaussian Naive Bayes

Gaussian NB Model:

$y \sim \text{Bern}(\phi)$ $= p(y)$

$x_1 \sim \text{Gaussian}(\mu_{1y}, \sigma^2_{1y}) = p(x_1 | y)$

$\vdots$

$x_M \sim \text{Gaussian}(\mu_{My}, \sigma^2_{My}) = p(x_M | y)$

Data: $y \in \{0,1\}$
$\vec{x} \in \mathbb{R}^M$

Recall:

Bernoulli NB Model:

$y \sim \text{Bern}(\phi)$

$x_1 \sim \text{Bern}(\theta_{1y})$

$\vdots$

$x_M \sim \text{Bern}(\theta_{My})$

$y \in \{0,1\}$
$\vec{x} \in \{0,1\}^M$

$$p(\vec{x}, y) = p(y) \prod_{m=1}^{M} \underbrace{p(x_m | y)}_{\text{event model}}$$

Gaussian NB Parameters

$\phi \in \mathbb{R}$

$$\mu = \begin{array}{cc} \left[\begin{array}{cc} \mu_{10} & \mu_{11} \\ \vdots & \vdots \\ \mu_{M0} & \mu_{M1} \end{array}\right] & \begin{array}{l} \leftarrow \text{for } x_1 \\ \\ \leftarrow \text{for } x_M \end{array} \\ \begin{array}{cc} \uparrow & \uparrow \\ \text{if } y=0 & \text{if } y=1 \end{array} & \end{array}$$

$$\sigma^2 = \begin{bmatrix} \sigma^2_{10} & \sigma^2_{11} \\ \vdots & \vdots \\ \sigma^2_{M0} & \sigma^2_{M1} \end{bmatrix}$$

$\mu_{my} \in \mathbb{R} \quad \forall y, m$

$\sigma^2_{my} > 0 \quad \forall y, m$

Details :

① Data likelihood :

② Learning (MLE or MAP)

③ Prediction :

$$\hat{y} = h(\vec{x}) = \operatorname*{argmax}_{y \in \{0,1\}} p(y \mid \vec{x}, \phi, \mu, \sigma^2)$$

Everything just works like "Bernoulli" NB!