CS3780/5780

Naive Bayes

- Recap: MLE: model $P(x, y)$ with P_θ where $\theta \in \Theta$

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; P_\theta(D)$$

model θ that maximizes likelihood of da

MAP:

$$\theta_{MAP} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; P(\theta \mid D) = P(D \mid \theta) \; P(\theta)$$

$P(D \mid \theta)$: likelihood; $P(\theta)$: Prior

Eg

Y	X = Favorite dish
Adult	Soup
child	Mac N Cheese
child	Mac N Cheese
Adult	Tacos
Adult	Soup
Child	Tacos
Adult	Soup
Adult	Mac N Cheese
child	Mac N Cheese

$X = \{$ Soup, Mac N cheese, Tacos$\}$

Estimate

$P(Y = \text{"child"} \mid X = \text{" Mac or cheese"})$?

Estimate $P(Y = y \mid X = x)$?

What is the issue with this?

Eg

Y	$X_{[1]}$ = Favorite dish	$X_{[2]}$ = # words known	$X_{[3]}$ = Favorite Movie	$X_{[4]}$ = Hours of sleep
Adult	Soup	20000	Godfather	8
child	Mac N Cheese	200	Frozen	11
child	Mac N Cheese	400	Frozen	12
Adult	Tacos	17000	La la land	6
Adult	Soup	15000	Godfather	5
Child	Tacos	1000	Eternals	10
Adult	Soup	21000	Avengers	10
Adult	Mac N Cheese	11000	Avengers	8
child	Mac N Cheese	700	Avengers	11

$\hat{P}(Y = \text{"Adult"} \mid X = (\text{"soup"}, 20000, \text{"Avengers"}, 8))$?

Naive Bayes Model

Assumption: $P(X = \vec{x} \mid Y = y) = \prod_{\alpha=1}^{d} P(X[\alpha] = x[\alpha] \mid Y = y)$

"Given it's a child, favorite dish, # words known, hours of sleep. are all independent"

why is this useful?

$$P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y)\, P(Y=y)}{P(X=x)}$$

$$= \frac{\prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)}{P(X=x)}$$

$$= \frac{\prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)}{\sum_{c \in \mathcal{Y}} \prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=c)\, P(Y=c)}$$

$P(Y=y)$ and $\forall \alpha$, $P(X[\alpha]=x[\alpha] \mid Y=y)$ are easy to estimate

Eg. estimate

$\hat{P}(Y = \text{"Adult"} \mid X = (\text{"soup"}, 20000, \text{"Avengers"}, 8))$?

$$h(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \hat{P}(Y=y \mid X=x)$$

$$= \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \prod_{\alpha} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)$$

$$= \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{\alpha} \log\left(P(X[\alpha]=x[\alpha] \mid Y=y)\right) + \log P(Y=y)$$

When $X_j[\alpha]$ are counts

Eg $X[\alpha] =$ ___ means α^{th} word in the dictionary occurs j times in the document

x is an m word document: $X[\alpha] \in \{0, 1, \ldots, m\}$ $\sum_{\alpha=1}^{d} X[\alpha] = m$

Multinomial Distribution $p(X = x \mid m, Y = y) = \frac{m!}{X[1]!\, X[2]! \times \cdots \times X[d]!} \prod_{i=1}^{d} (\theta_{\alpha,y})^{X[\alpha]}$

MLE estimate: $\hat{\theta}_{\alpha,y} = \frac{\sum_{i=1}^{n} 1\{y_i = y\}\, x_i[\alpha]}{\sum_{i=1}^{n} 1(y_i = y)\, m_i}$

m_i = # words in document i

$$h(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \hat{p}(Y = y) \prod_{\alpha=1}^{d} \hat{\theta}_{\alpha,y}^{x[\alpha]}$$

$X[\alpha]$'s are continuous variables.: Gaussian distribution conditioned on Y

$$p(x_{[\alpha]} = x \mid Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x - \mu_y)^2}{2\sigma_y^2}\right)$$

Parameter estimation:

$$\hat{\mu}_y[\alpha] = \frac{\sum_{i=1}^{n} 1(y_i = y)\, x_i[\alpha]}{\sum_{i=1}^{n} 1(y_i = y)} \qquad \hat{\sigma}_y^2 = \frac{\sum_{i=1}^{n} 1(y_i = y)\, (x_i[\alpha] - \hat{\mu}_y[\alpha])^2}{\sum_{i=1}^{n} 1(y_i = y)}$$

1. For both multinomial case and Gaussian case (with variance between class per feature fixed) classification boundary is linear.

2. For Gaussian case

$$P(Y = y \mid x) = \frac{1}{1 + \exp\left[-y(w^T x + b)\right]}$$

logistic link function.