Naive Bayes

Recap: MLE : model $P(x, y)$ as P_θ for $\theta \in \Theta$

$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max} \; P_\theta(\text{Data})$$

model θ that maximizes likelihood of data

MAP: $\theta_{MAP} = \underset{\theta \in \Theta}{\arg\max}$ $P(\theta \mid \text{Data}) = \underbrace{P(\text{Data} \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{Prior}}$

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

$X \in \{$Soup, Mac N Cheese, Tacos$\}$

Estimate

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

$\boxed{\text{Estimate } P(Y = y \mid X = x) \text{ ?}}$

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	God father	8
Child	Mac N Cheese	200	Frozen	11
Child	Mac N Cheese	400	Frozen	12
Adult	Tacos	17000	Usual suspects	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

$\hat{P}(Y = \text{Adult} \mid X = (\text{Soup}, 20{,}000, \text{Avengers}, 8))$

Naive Bayes Model

Assumption: $P(X=x \mid Y=y) = \overset{\text{Product}}{\prod_{\alpha=1}} P(X[\alpha]=x[\alpha] \mid Y=y)$

"Given its a child, favorite dish, # words known, hours of sleep are independent"

why is this useful?

$$\boxed{\begin{aligned} P(Y=y \mid X=x) &= \frac{P(X=x \mid Y=y)\, P(Y=y)}{P(X=x)} \\ &= \frac{\prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)}{P(X=x)} \\ &\propto \prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y) \\ &= \frac{\prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)}{\sum_{c\in C} \prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=c)\, P(Y=c)} \end{aligned}}$$

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

$$\boxed{\begin{array}{l} \text{Eg} \quad \text{Estimate} \\ \hat{P}(Y=\text{Adult} \mid X=(\text{Soup}, 20{,}000, \text{Avengers}, 8))\ ? \end{array}}$$

$$\begin{aligned} h(x) &= \underset{y\in C}{\operatorname{argmax}}\ \hat{P}(y=c \mid X=x) \\ &= \underset{y\in C}{\operatorname{argmax}} \prod_{\alpha} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y) \\ &= \underset{y\in C}{\operatorname{argmax}} \sum_{\alpha} \log(P(X[\alpha]=x[\alpha] \mid Y=y)) \\ &\qquad + \log P(Y=y) \end{aligned}$$

When $x_i[\alpha]$ are counts
eg $x[\alpha] = j$ means α^{th} word in the dictionary
occurs j times in the document x_i

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

multinomial distribution: $P(X = x \mid m, Y=y) = \dfrac{m!}{x[1]!\, x[2]! \times \dots \times x[d]!} \prod_{i=1}^{d} (\theta_{\alpha,c})^{x[\alpha]}$

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

m_i: # words in document i

(MAP hallucinate l examples per document)

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

$x[\alpha]$'s are continuous variables: Gaussian distribution conditioned on y

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

Parameter estimation:

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

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

2. For Gaussian case

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

logistic link function.