Naïve Bayes II

Recall: Naïve Bayes assumption:

features $x_1, x_2, \ldots, x_d$ independent given label y.

$$P(x|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y)$$

Categorical Naïve Bayes

assumption: $x_\alpha \in \{f_1, f_2, f_3, \ldots, f_{K_\alpha}\} = \mathcal{X}_\alpha$

eg. $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \cdots \times \mathcal{X}_d$

$|\mathcal{X}_\alpha| = K_\alpha$.

model: ~~$P(X_\alpha|Y)$~~.

$P(x_\alpha = j|Y) = \theta_{j,Y,\alpha}$

where $\theta_{j,Y,\alpha} \geq 0$ and $\sum_{j \in \mathcal{X}_\alpha} \theta_{j,Y,\alpha} = 1$.

categorical distribution.

estimating θ:

$$\hat{\theta}_{j,Y,\alpha} = \frac{\#\text{ of examples with label } Y \text{ and with feature } j}{\#\text{ of examples with label } Y}$$

$I(\text{true}) = 1$
$I(\text{false}) = 0$.

$$\hat{\theta}_{j,c,\alpha} = \frac{\sum_{i=1}^{n} I(y_i = c) I(x_{i,\alpha} = j) + \ell}{\sum_{i=1}^{n} I(y_i = c) + \ell \cdot K_\alpha}$$

Prediction:

$$h(x) = \arg\max_{c \in y} \hat{P}(y=c|x)$$

$$= \arg\max_{c \in y} \hat{P}(x|y=c) \cdot \hat{P}(y=c)$$

$$= \arg\max_{c \in y} \hat{\pi}_c \cdot \frac{m!}{x_1! x_2! \cdots x_d!} \prod_{\alpha=1}^{d} (\theta_{\alpha c})^{x_\alpha}$$

$$= \arg\max_{c \in y} \hat{\pi}_c \cdot \prod_{\alpha=1}^{d} (\theta_{\alpha c})^{x_\alpha}$$

$$= \arg\max_{c \in y} \underbrace{\log(\hat{\pi}_c)}_{b_c} + \sum_{\alpha=1}^{d} x_\alpha \underbrace{\log(\theta_{\alpha c})}_{w_{\alpha c}}$$

$$= \arg\max_{c \in y} b_c + \sum_{\alpha=1}^{d} w_{\alpha c} \cdot x_\alpha$$

$$= \arg\max_{c \in y} (Wx + b)_c$$

$W \in \mathbb{R}^{C \times d}$

$$\hat{\theta}_{j,c,\alpha} = \frac{|\{i \mid y_i = c \text{ and } x_{i,\alpha} = j\}| + \ell}{|\{i \mid y_i = c\}| + \ell \cdot K_\alpha}$$

$$= \frac{|\{(x,y) \in D \mid y = c \text{ and } x_\alpha = j\}| + \ell}{|\{(x,y) \in D \mid y = c\}| + \ell \cdot K_\alpha}$$

Prediction

$$h(x) = \arg\max_{c \in Y} \hat{P}(y = c \mid x)$$

$$= \arg\max_{c \in Y} \frac{\hat{P}(x \mid y = c) \cdot \hat{P}(y = c)}{\hat{P}(x)} \quad \text{(Bayes Thm)}$$

$$= \arg\max_{c \in Y} \hat{P}(x \mid y = c)\, \hat{P}(y = c)$$

$$= \arg\max_{c \in Y} \hat{P}(y = c) \cdot \prod_{\alpha=1}^{d} \hat{P}(x_\alpha \mid y = c) \quad \text{(Naive Bayes assumption)}$$

$$\boxed{h(x) = \arg\max_{c \in Y} \hat{\pi}_c \prod_{\alpha=1}^{d} \hat{\theta}_{x_\alpha, c, \alpha}}$$

$$\hat{\pi}_c = \frac{|\{(x,y) \in D \mid y = c\}|}{n} = \frac{1}{n} \sum_{i=1}^{n} I(y_i = c)$$

Q: For ~~multi~~categorical naive Bayes, how many parameters are there? ($\hat{\pi}$ and $\hat{\theta}$)
in terms of $n, d, K_\alpha, |Y| = C$ i.e. what is the dimension of the parameter vector?

"I am spam. Spam I am. Do you like perceptrons?"

I → 2
am → 2
spam → 2
do → 1
you → 1
like → 1
perceptron → 1.

$x = (\ldots, 2, \ldots, 1, \ldots, 1, \ldots)$ — sparse!

d = # of english words

$nnz = 7$

Multinomial Naive Bayes

features $x_\alpha \in \mathbb{N}$, sum up to $\sum_{\alpha=1}^{d} x_\alpha = m$. (length)

Model: x are multinomially distributed

$$P(x|y) = \frac{m!}{x_1! x_2! \cdots x_d!} \prod_{\alpha=1}^{d} (\theta_{\alpha y})^{x_\alpha}$$

e.g. a 6-sided die: $\sum_{x \in \mathbb{N}^6} P(x) = \frac{m!}{x_1! x_2! x_3! x_4! x_5! x_6!} \cdot \prod_{\alpha=1}^{6} \left(\frac{1}{6}\right)^{x_\alpha}$

Estimate parameters:

$$\hat{\theta}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c) \cdot x_{i,\alpha} + \ell}{\sum_{i=1}^{n} I(y_i = c) \cdot m_i + \ell \cdot d}$$

length of i^{th} training example $m_i = \sum_{\alpha=1}^{d} x_{i,\alpha}$

# of params: $C + dC$

Gaussian Naive Bayes!

Features: $x_\alpha \in \mathbb{R}$, $X \in \mathbb{R}^d$

Model: $X_\alpha \sim N(\mu_{\alpha c}, \sigma^2_{\alpha c})$

$$P(x_\alpha | y=c) = \frac{1}{\sqrt{2\pi\sigma^2_{\alpha c}}} \exp\left(-\frac{(x_\alpha - \mu_{\alpha c})^2}{2\sigma^2_{\alpha c}}\right)$$

Estimating Parameters:

$$\hat{\mu}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c) x_{i\alpha}}{\sum_{i=1}^{n} I(y_i = c)}$$

$$\hat{\sigma}^2_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c)(x_{i\alpha} - \hat{\mu}_{\alpha c})^2}{\sum_{i=1}^{n} I(y_i = c) - 1}$$

or!

$$P\ x_\alpha \sim N(\mu_{\alpha c}, \sigma^2_\alpha)$$

here: $\hat{\sigma}^2_\alpha = \frac{1}{n} \sum_{i=1}^{n} (x_{i\alpha} - \hat{\mu}_{\alpha y_i})^2$

Parameters are $\hat{\pi}_1, \hat{\pi}_2 \cdots \hat{\pi}_C, \cdots \hat{\mu}_{\alpha c} \cdots, \hat{\sigma}_\alpha^2$

$$\hat{P}(x_\alpha | y) = \frac{1}{\sqrt{2\pi}\cdot\hat{\sigma}_\alpha} \cdot \exp\left(-\frac{(x_\alpha - \hat{\mu}_{\alpha y})^2}{2\hat{\sigma}_\alpha^2}\right).$$

What is $\hat{P}(y|x)$? (As a function of π, μ, σ, and x).

$$\hat{P}(y|x) = \frac{\hat{P}(x|y)\cdot\hat{P}(y)}{\hat{P}(x)} \quad \text{(Bayes rule)} \qquad \text{Naive Bayes (Assumption + Gaussian)}$$

$$= \frac{\hat{P}(y)}{\hat{P}(x)} \cdot \prod_{\alpha=1}^{d} \hat{P}(x_\alpha|y) = \frac{\hat{\pi}_y}{\hat{P}(x)} \cdot \prod_{\alpha=1}^{d} \frac{1}{\sqrt{2\pi}\cdot\hat{\sigma}_\alpha} \cdot \exp\left(-\frac{(x_\alpha - \hat{\mu}_{\alpha y})^2}{2\hat{\sigma}_\alpha^2}\right)$$

let $\frac{1}{Z} = \frac{1}{\hat{P}(x)} \cdot \prod_{\alpha=1}^{d} \frac{1}{\sqrt{2\pi}\cdot\hat{\sigma}_\alpha} \Rightarrow = \frac{\hat{\pi}_y}{Z} \prod_{\alpha=1}^{d} \exp\left(-\frac{(x_\alpha - \hat{\mu}_{\alpha y})^2}{2\hat{\sigma}_\alpha^2}\right)$

$$= \frac{\hat{\pi}_y}{Z} \cdot \exp\left(\sum_{\alpha=1}^{d} -\frac{1}{2\hat{\sigma}_\alpha^2}(x_\alpha - \hat{\mu}_{\alpha y})^2\right)$$

$(x_\alpha - \hat{\mu}_{\alpha y})^2 = x_\alpha^2 - 2\hat{\mu}_{\alpha y}x_\alpha + \hat{\mu}_{\alpha y}^2$

$$= \frac{\hat{\pi}_y}{Z} \cdot \exp\left(\sum_{\alpha=1}^{d} -\frac{x_\alpha^2}{2\hat{\sigma}_\alpha^2}\right) \exp\left(\sum_{\alpha=1}^{d} \frac{2\hat{\mu}_{\alpha y}x_\alpha - \hat{\mu}_{\alpha y}^2}{2\hat{\sigma}_\alpha^2}\right)$$

$\frac{\hat{\pi}_y}{Z'}$ i.e. $\frac{1}{Z'} = \frac{1}{Z}\exp\left(\sum_{\alpha=1}^{d} -\frac{x_\alpha^2}{2\hat{\sigma}_\alpha^2}\right)$

$$= \frac{\hat{\pi}_y}{Z'} \exp\left(\sum_{\alpha=1}^{d} \frac{\hat{\mu}_{\alpha y}x_\alpha}{\hat{\sigma}_\alpha^2}\right) \cdot \exp\left(\sum_{\alpha=1}^{d} -\frac{\hat{\mu}_{\alpha y}^2}{2\hat{\sigma}_\alpha^2}\right)$$

let $\beta_y = \hat{\pi}_y \cdot \exp\left(\sum_{\alpha=1}^{d} -\frac{\hat{\mu}_{\alpha y}^2}{2\hat{\sigma}_\alpha^2}\right)$.

$$= \frac{\beta_y}{Z'} \cdot \exp\left(\sum_{\alpha=1}^{d} \frac{\hat{\mu}_{\alpha y}}{\hat{\sigma}_\alpha^2} \cdot x_\alpha\right).$$

$$\hat{P}(y|x) = \frac{\beta_y}{Z'} \exp\left(\sum_{\alpha=1}^{d} \frac{\hat{\mu}_{\alpha y}}{\hat{\sigma}_\alpha^2} \cdot x_\alpha\right).$$

let $W \in \mathbb{R}^{C \times d}$ s.t. $W_{y\alpha} = \frac{\hat{\mu}_{\alpha y}}{\hat{\sigma}_\alpha^2}$

$$= \frac{\beta_y}{Z'} \cdot \exp\left((Wx)_y\right).$$

C = # of classes $= |\mathcal{Y}|$

$$= \frac{1}{Z'} \exp\left((Wx+b)_y\right).$$

let $b \in \mathbb{R}^C$ s.t. $b_y = \log \beta_y$

Know: $\sum_{y=1}^{C} \hat{P}(y|x) = 1 = \sum_{y=1}^{C} \frac{1}{Z'} \exp\left((Wx+b)_y\right)$.

$$\Rightarrow Z' = \sum_{y=1}^{C} \exp\left((Wx+b)_y\right)$$

$$\hat{P}(y|x) = \frac{\exp\left((Wx+b)_y\right)}{\sum_{j=1}^{C} \exp\left((Wx+b)_j\right)} \qquad \left(\underset{u \in \mathbb{R}^d}{\text{Softmax}}(u)_i = \frac{\exp(u_i)}{\sum_{j=1}^{d} \exp(u_j)}\right)$$

$$\hat{P}(y|x) = \text{Softmax}(Wx+b)_y.$$

$$h(x) = \underset{c \in \mathcal{Y}}{\arg\max}\ \text{Softmax}\left((Wx+b)_c\right) = \underset{c \in \mathcal{Y}}{\arg\max}\ (Wx+b)_c$$

If we have just 2 classes:

$$\hat{P}(y=1|x) = \frac{\exp\left((Wx+b)_1\right)}{\exp\left((Wx+b)_1\right) + \exp\left((Wx+b)_2\right)} = \frac{1}{1+\exp\left((Wx+b)_2 - (Wx+b)_1\right)}$$

now let $w \leftarrow w_2 - w_1$ $\quad w \in \mathbb{R}^d$ $\quad$ let $b \leftarrow b_2 - b_1$ $\quad b \in \mathbb{R}$

$$= \frac{1}{1+\exp(w^T x + b)} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1+e^z}.$$

predict $y=1$, when $w^T x + b \gtrless 0$.

Discriminative modeling: model just $P(y|x)$.

Logistic Regression. $y \in \{-1, 1\}$. Params $w \in \mathbb{R}^d$ $b \in \mathbb{R}$.

Assume:

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

if $x \leftarrow \begin{bmatrix} x \\ 1 \end{bmatrix}$ $w \leftarrow \begin{bmatrix} w \\ b \end{bmatrix}$ $w^T x \leftarrow w^T x + b$

So: ~~$P(y|x) = \frac{1}{1+\exp(-y(w^T x + b))}$~~

$$P(y|x) = \frac{1}{1+\exp(-yw^T x)}.$$

$h(x) = \text{sign}(w^T x)$. $\Leftarrow$ a linear model.

~~MAP~~ MLE estimate for w:

~~max~~ $\underset{w \in \mathbb{R}^d}{\arg\max}$ $P(D\$; w)$

$$\hat{\omega} = \arg\max_{\omega \in \mathbb{R}^d} P(D; \omega)$$

$$= \arg\max_{\omega \in \mathbb{R}^d} P((x_1, y_1), (x_2, y_2) \cdots (x_n, y_n); \omega)$$

$$= \arg\max_{\omega \in \mathbb{R}^d} \prod_{i=1}^{n} P((x_i, y_i); \omega)$$

$$= \arg\max_{\omega \in \mathbb{R}^d} \prod_{i=1}^{n} P(y_i | x_i; \omega) \underbrace{P(x_i; \omega)}_{\text{not modeling}}$$

no distribution

$$\hat{\omega}_{MLE} = \arg\max_{\omega \in \mathbb{R}^d} P(y | x; \omega) \quad \text{(discriminative MLE)}$$

$$= \arg\max_{\omega \in \mathbb{R}^d} \prod_{i=1}^{n} P(y_i | x_i; \omega)$$

$$= \arg\max_{\omega \in \mathbb{R}^d} \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i x_i^T \omega)}$$

$\log \frac{1}{x} = -\log x$

$$= \arg\max_{\omega \in \mathbb{R}^d} \sum_{i=1}^{n} -\log\left(1 + \exp(-y_i x_i^T \omega)\right)$$

$$\hat{\omega}_{MLE} = \arg\min_{\omega \in \mathbb{R}^d} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^T \omega)\right)$$

logistic regression loss.