0.75 × 0.25   21

1. **Naive Bayes with Binary Features.** Consider a group of 50 Cornell Students. 20 of them are Master's students, while the rest 30 of them are PhD students. There are 5 Master's students who bike, and there are 5 Master's students who ski. On the other hand, 20 PhD students bike, and 15 PhD students ski.

15   10

We can formulate this as a machine learning problem by modeling the students with features $x = (x_1, x_2) \in \{0,1\}^2$, where x_1 is a binary indicator of whether the students bike and x_2 is a binary indicator of whether they ski, and the target y equals 1 if they are PhD students and 0 if they are Master's students.

(a) Please elaborate in this context what is the Naive Bayes assumption.

(b) With the Naive Bayes assumption, find the probability of a student in this group who neither bikes or skis being a Master's student

(c) Suppose we know that every PhD who skis also bikes. Does it make sense to still assume that probability of biking and skiing are conditionally independent for a PhD student? If not, how would your answer to part (b) change with this knowledge (you can still assume probability of biking and skiing are conditionally independent for a Master's student)?

(a)

Naive Bayes assumption is each feature is independent to each other, e.g. people being master/phD, can or cannot ski, can or cannot bike are independent

(b) $$P(y=0 \mid x_1=0, x_2=0) = \frac{P(x_1=0, x_2=0 \mid y=0) \cdot P(y=0)}{P(x_1=0, x_2=0)}$$

$$= \frac{P(y=0) \cdot P(x_1=0 \mid y=0) \cdot P(x_2=0 \mid y=0)}{P(y=1) \cdot P(x_1=0 \mid y=1) \cdot P(x_2=0 \mid y=1) + P(y=0) \cdot P(x_1=0 \mid y=0) \cdot P(x_2=0 \mid y=0)}$$

$$= \frac{0.4 \times 0.75 \times 0.75}{0.6 \times 0.33 \times 0.5 + 0.4 \times 0.75 \times 0.75}$$

$$= 0.6923$$

(c) No, because students who ski can bike, the feature has certain connections, it's not independent.

denominator part $P(y=1) \cdot P(x_1=0 \mid y=1) \cdot P(x_2=0 \mid y=1)$ is different, it should be $P(y=1 \mid x_1=0, y=0)$, and it will increase and answer of (b) will decrease since its denominator increases

(a) Show that the maximum likelihood estimate for the parameters ϕ is

$$\phi^* = \frac{n_k}{n},$$

where n_k is the number of data points with class k.

$$\arg\max_{\theta} \frac{1}{n} \sum_i \log P_\theta(x^{(i)}, y^{(i)})$$

$$= \arg\max_{\theta} \sum_i \sum_j \log P_\theta(x_j^{(i)} \mid y^{(i)}) + \sum_i \log P_\theta(y^{(i)}) \quad ①$$

$\because$ Naive bayes

$$\therefore ① = \arg\max_{\theta} \sum_i \sum_j \log P_\theta(x_j^{(i)} \mid y^{(i)}; \psi_{jlk}) + \sum_i \log P_\theta(y^{(i)}, \varphi)$$

$$\therefore \arg\max_{\theta} \sum_i \log P_\theta(y = y^{(i)}; \varphi)$$

$$= \arg\max_{\theta} \sum_i \log \varphi_{y^{(i)}} - n \cdot \log \sum_k \varphi_k$$

$$= \sum_k \sum_{i : y^{(i)} = k} \log \varphi_k - n \cdot \log \sum_k \varphi_k$$

Take the derivative on the expression

$$\therefore \frac{\varphi_k}{\sum_l \varphi_l} = \frac{n_k}{n}$$

$$\because \sum_l \varphi_k = 1$$

$$\therefore \varphi^* = \frac{n_k}{n}$$

(b) Show that the maximum likelihood estimate for the parameters $\psi_{jk\ell}$ is

$$\psi^*_{jk\ell} = \frac{n_{jk\ell}}{n_k},$$

where $n_{jk\ell}$ is the number of data points with class k for which the j-th feature equals ℓ.

$$\arg\max_{\varphi} \sum_i \sum_j \log P(\theta)\,(x_j^{(i)} \mid y_j^{(i)} \,;\, \varphi_{jlk})$$

$$= \sum_i \sum_j \log \varphi_{jx^{(i)}y^{(i)}} - \sum_i \sum_j \sum_{i:y^{(i)}=k} \log \sum_l \varphi_{jkl}$$

$$= \sum_k \sum_l \sum_j \sum_{i:x^{(i)}=\ell \text{ and } y^{(i)}=k} \log \varphi_{jx^{(i)}y^{(i)}} - \sum_k \sum_j \sum_{i:y^{(i)}=k} \log \sum_l \varphi_{jkl}$$

$$= \sum_k \sum_l \sum_j \sum_{i:x^{(i)}=\ell \text{ and } y^{(i)}=k} \log \varphi_{jkl} - \sum_k \sum_j \sum_{i:y^{(i)}=k} \log \sum_l \varphi_{jkl}$$

$$= n_{jkl} \cdot \log \varphi_{jkl} - n_k \cdot \log \sum_l \varphi_{jkl}$$

Take derivative and set it to 0

$$\frac{n_{jkl}}{\varphi_{jkl}} = \frac{n_k}{\sum_l \varphi_{jkl}}$$

$$\frac{\varphi_{jkl}}{\sum_l \varphi_{jkl}} = \frac{n_{jkl}}{n_k}$$

$$\because \sum_l \varphi_{jkl} = 1$$

$$\therefore \varphi_{jkl} = \frac{n_{jkl}}{n_k}$$