MLE for Multinomial Naive Bayes

Consider the following definition of MLE problem for multinomials. The input to the problem is a finite set $\mathcal{Y}$, and a weight $c_y \geq 0$ for each $y \in \mathcal{Y}$. The output from the problem is the distribution p^* that solves the following maximization problem.

$$p^* = \arg\max_{p \in \mathcal{P}_y} \sum_{y \in \mathcal{Y}} c_y \log p_y$$

(i) Prove that, the vector p^* has components

$$p_y^* = \frac{c_y}{N}$$

for $\forall y \in \mathcal{Y}$, where $N = \sum_{y \in \mathcal{Y}} c_y$. (Hint: Use the theory of Lagrange multiplier)

Answer:

$$\begin{aligned} &\max \sum_{y \in \mathcal{Y}} c_y \log p_y \\ &\text{s.t. } \sum_{y \in \mathcal{Y}} p_y = 1 \\ &\quad p_y \geq 0, \qquad \forall y \in \mathcal{Y} \end{aligned}$$

Lagrangian problem is:

$$\begin{cases} F = -\sum_{y \in \mathcal{Y}} c_y \log p_y + \lambda(\sum_{y \in \mathcal{Y}} p_y - 1) - \sum_{y \in \mathcal{Y}} \mu_y p_y \\ \mu_y \geq 0 \qquad \forall y \in \mathcal{Y} \end{cases}$$

$$\Rightarrow \begin{cases} \frac{\partial F}{\partial p_y} = -\frac{c_y}{p_y} + \lambda - \mu_y = 0 & \forall y \in \mathcal{Y} \\ \mu_y p_y = 0 & \forall y \in \mathcal{Y} \end{cases}$$

- For λ,

$$\begin{aligned} &\because \mu_y p_y = 0 \text{ and } c_y = \lambda p_y - \mu_y p_y \\ &\therefore c_y = \lambda p_y \Leftrightarrow p_y = \tfrac{c_y}{\lambda} \\ &\because \textstyle\sum_{y \in \mathcal{Y}} p_y = 1 \\ &\therefore \textstyle\sum_{y \in \mathcal{Y}} p_y = \frac{1}{\lambda} \sum_{y \in \mathcal{Y}} c_y = 1 \\ &\therefore \lambda = \textstyle\sum_{y \in \mathcal{Y}} c_y \end{aligned}$$

- Therefore, $p_y = \frac{c_y}{\lambda} = \frac{c_y}{\sum_{y \in \mathcal{Y}} c_y}$

(ii) Using the above consequence, prove that, the maximum-likelihood estimates for Naive Bayes model are as follows:

$$p(y)=\frac{\sum_{i=1}^{m} 1\left(y^{(i)}=y\right)}{m}$$

and

$$p_j(x \mid y)=\frac{\sum_{i=1}^{m} 1\left(y^{(i)}=y \wedge x_j^{(i)}=x\right)}{\sum_{i=1}^{m} 1\left(y^{(i)}=y\right)}$$

Answer:

The first step is to re-write the log-likelihood function in a way that makes direct use of “counts” taken from the training data:

$$\begin{aligned} l(\Omega) &= \sum_{i=1}^{m} \log p\left(y^{(i)}\right)+\sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j\left(x_j^{(i)} \mid y^{(i)}\right) \\ &= \sum_{y \in \mathcal{Y}} \operatorname{count}(y) \log p(y) \\ &\quad +\sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in\{-1,+1\}} \operatorname{count}_j(x \mid y) \log p_j(x \mid y) \end{aligned}$$

where as before

$$\operatorname{count}(y)=\sum_{i=1}^{m} 1\left(y^{(i)}=y\right)$$

$$\operatorname{count}_j(x \mid y)=\sum_{i=1}^{m} 1\left(y^{(i)}=y \wedge x_j^{(i)}=x\right)$$

Consider first maximization of this function with respect to the $q(y)$ parameters. It is easy to see that the term

$$\sum_{j=1}^{d} \sum_{y \in \mathcal{Y}} \sum_{x \in\{-1,+1\}} \operatorname{count}_j(x \mid y) \log p_j(x \mid y)$$

does not depend on the $p(y)$ parameters at all. Hence to pick the optimal $p(y)$ parameters, we need to simply maximize

$$\sum_{y \in \mathcal{Y}} \operatorname{count}(y) \log p(y)$$

Subject to the constraints $p(y) \geq 0$ and $\sum_{y=1}^{k} p(y)=1$, by the consequence of (i) , the values for $q(y)$ which maximize this expression under these constraints is simply

$$p(y)=\frac{\operatorname{count}(y)}{\sum_{y=1}^{k} \operatorname{count}(y)}=\frac{\operatorname{count}(y)}{n}$$

By a similar argument, we can maximize each term of the form

$$\sum_{x \in\{-1,+1\}} \operatorname{count}_j(x \mid y) \log p_j(x \mid y)$$

Applying (i), we can get

$$p_j(x \mid y) = \frac{\mathsf{count}_j(x \mid y)}{\sum_{x\in\{-1,1\}} \mathsf{count}_j(x \mid y)} = \frac{\mathsf{count}_j(x \mid y)}{\mathsf{count}(y)}$$