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:

$$\begin{gathered} \max \sum_{y \in \mathcal{Y}} c_y \log p_y \\ \text{s.t. } \sum_{y \in \mathcal{Y}} p_y = 1 \\ p(y) \geq 0, \forall y \in \mathcal{Y} \end{gathered}$$

Lagrangian problem is:

$$\begin{gathered} F = \sum_{y \in \mathcal{Y}} c_y \log p_y + \lambda \left(\sum_{y \in \mathcal{Y}} p_y - 1 \right) \\ \frac{\partial F}{\partial \lambda} = \sum_{y \in \mathcal{Y}} p_y - 1 \\ \frac{\partial F}{\partial p_y} = \frac{c_y}{p_y} + \lambda \\ \lambda = -\sum_{y \in \mathcal{Y}} c_y, p_y = \frac{c_y}{\sum_{y \in \mathcal{Y}} c_y} \\ p_y = \frac{c_y}{N} \end{gathered}$$

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

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

and

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

Answer:

We now prove the result in theorem 1. Our first step is to re-write the log-likelihood function in a way that makes direct use of "counts" taken from the training data:

$$
\begin{aligned}
L(\underline{\theta}) &= \sum_{i=1}^{m} \log q\left(y_i\right) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log q_j\left(x_{i,j} \mid y_i\right) \\
&= \sum_{y \in \mathcal{Y}} \text{count}(y) \log q(y) \\
&\quad + \sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in \{-1,+1\}} \text{count }_j(x \mid y) \log q_j(x \mid y)
\end{aligned}
$$

where as before

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

$$
\text{count}_j(x \mid y) = \sum_{i=1}^{m} \left[\left[y_i = y \text{ and } x_j^{(i)} = x\right]\right]
$$

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

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

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

$$
\sum_{y \in \mathcal{Y}} \text{count}(y) \log q(y)
$$

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

$$
q(y) = \frac{\text{count}(y)}{\sum_{y=1}^{k} \text{count}(y)} = \frac{\text{count}(y)}{n}
$$

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

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

Applying (i), we can get

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