

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.

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

(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:

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

Lagrangian problem is:

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

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

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

and

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

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:

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

where as before

$$\begin{aligned}
\text{count}(y) &= \sum_{i=1}^m [\mathbb{I}[y^{(i)} = y]] \\
\text{count}_j(x | y) &= \sum_{i=1}^m [\mathbb{I}[y_i = y \text{ and } x_j^{(i)} = x]]
\end{aligned}$$

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

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

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

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

Applying (i), we can get

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