

## 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:

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

Lagrangian problem is:

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

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

• For  $\lambda$ ,

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

• 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:

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

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:

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:

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

where as before

$$\begin{aligned} \text{count}(y) &= \sum_{i=1}^m 1(y^{(i)} = y) \\ \text{count}_j(x | y) &= \sum_{i=1}^m 1(y^{(i)} = y \wedge x_j^{(i)} = x) \end{aligned}$$

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

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

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

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

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

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

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

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

Applying ( i ), we can get

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