MLE Derivation for $P(C_j)$ in Naive Bayes Classification

Yizhou Sun (yzsun@ccs.neu.edu)

September 24, 2014

Following Page 12 to 13 in the slides, given a dataset $D = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i$ is a p-Dimensional feature vector, y_i is the class label taking values from $\{C_1, C_2, \ldots, C_m\}$, the goal is to find the best parameters $P(C_j)$ and $P(X_k = v|C_j)$ that can maximize the likelihood of the observed dataset:

$$\begin{aligned} L &= \prod_i P(\mathbf{x}_i, y_i) = \prod_i P(\mathbf{x}_i|y_i)P(y_i) = \prod_i (\prod_k P(x_{ik}|y_i))P(y_i) \\ &= \prod_i (\prod_k P(x_{ik}|y_i)) \prod_j P(C_j)^{\mathbb{1}(y_i = C_j)} \end{aligned} \tag{1}$$

where $\mathbb{1}$ is the indicator function, which equals to 1 if the predicate holds, otherwise, 0.

This is equivalent to maximize log-likelihood:

$$logL = \sum_i \sum_k \log(P(x_{ik}|y_i)) + \sum_i \sum_j \mathbb{1}(y_i = C_j) \log(P(C_j)) \tag{2}$$

Now we can see that if we want to estimate $P(C_j)$, the first part of the log-likelihood function is irrelevant, as it does not contain $P(C_j)$. Note that, we have a constraint on $P(C_j)$, which is $\sum_j P(C_j) = 1$. We can use the method of Lagrange multipliers to solve the problem, which makes us to maximize the following Lagrange function:

$$J = \sum_i \sum_j \mathbb{1}(y_i = C_j) \log(P(C_j)) + \lambda(\sum_j P(C_j) - 1) \tag{3}$$

By taking the first derivative respective to $P(C_j)$ and set it to 0, we have

$$\nabla_{P(C_j)} J = \sum_i \frac{\mathbb{1}(y_i = C_j)}{P(C_j)} + \lambda = 0 \tag{4}$$

We can then get $-\lambda = \sum_i \sum_j \mathbb{1}(y_i = C_j) = \sum_i = |D|$, the total number of objects in the dataset. By plugging in λ, we can get $P(C_j) = \frac{\sum_i \mathbb{1}(y_i = C_j)}{|D|} = \frac{|C_{j,D}|}{|D|}$, where $|C_{j,D}|$ denotes the total number of objects in D that belong to class C_j.

Other notes on parameter derivation:

- http://www.cs.columbia.edu/~mcollins/em.pdf
- http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/NB.pdf