Naive Bayes with Categorical Variables

Lu Sun

March 30, 2020

Here we discuss the parameter estimation of naive Bayes (NB) with categorical variables based on maximum likelihood estimation (MLE), corresponding to (6) and (8) in Ch.3.2.3 of Machine Learning.

1 Categorical Naive Bayes

In NB, we have a d-dimensional input variable $X^\top = (X_1, X_2, ..., X_d)$, and a output variable Y. Based on the conditional independence assumption, the posterior is formulated by

$$P(Y|X) \propto P(X|Y)P(Y) = \prod_{j=1}^{d} P(X_j|Y)P(Y). \tag{1}$$

Suppose that all the variables are categorical, such that $X_j \in \{1, 2, ..., K\}$ $(j = 1, 2, ..., d)$ and $Y \in \{1, 2, ..., M\}$. We use θ_{jkm} and π_m to denote the probabilities $P(X_j = k|Y = m)$ and $P(Y = m)$, respectively, and thus

$$\begin{aligned} \theta_{jkm} &= P(X_j = k|Y = m), \\ \pi_m &= P(Y = m), \quad \forall j, k, m. \end{aligned} \tag{2}$$

In total, we need to estimate $d(K-1)M$ parameters for θ, and $M-1$ parameters for π.

2 Probability Density Function

In this section, we derive the probability density function $P(X_j|Y, \theta)$ $(\forall j)$ and $P(Y|\pi)$. The first probability density function is defined by

$$\begin{aligned} P(X_j|Y, \theta) = &\left(\theta_{j11}^{\mathbf{1}_{X_j=1}\mathbf{1}_{Y=1}} \cdots \theta_{jK1}^{\mathbf{1}_{X_j=K}\mathbf{1}_{Y=1}}\right) \cdot \\ &\left(\theta_{j12}^{\mathbf{1}_{X_j=1}\mathbf{1}_{Y=2}} \cdots \theta_{jK2}^{\mathbf{1}_{X_j=K}\mathbf{1}_{Y=2}}\right) \cdot \\ &\vdots \\ &\left(\theta_{j1M}^{\mathbf{1}_{X_j=1}\mathbf{1}_{Y=M}} \cdots \theta_{jKM}^{\mathbf{1}_{X_j=K}\mathbf{1}_{Y=M}}\right) \cdot \\ = &\prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{jkm}^{\mathbf{1}_{X_j=k}\mathbf{1}_{Y=m}}, \end{aligned} \tag{3}$$

where $\mathbf{1}_.$ denotes the indicator function. Similarly, the second probability density function is

$$P(Y|\pi) = \pi_1^{\mathbf{1}_{Y=1}} \pi_2^{\mathbf{1}_{Y=2}} \cdots \pi_M^{\mathbf{1}_{Y=M}} = \prod_{m=1}^{M} \pi_m^{\mathbf{1}_{Y=m}}. \tag{4}$$

3 Likelihood Function

Given a training dataset $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$, in which $x_i \in \{1, ..., K\}^d$ and $y \in \{1, ..., M\}$. The log-likelihood function is written as follows:

$$\begin{aligned}
\ell(\theta, \pi) &= \ln P(\mathcal{D}|\theta, \pi) \\
&= \ln P\left((x_1, y_i), ..., (x_N, y_i)|\theta, \pi\right) \\
&= \ln \prod_{i=1}^{N} P(x_i, y_i|\theta, \pi) \\
&= \ln \prod_{i=1}^{N} P(x_i|y_i, \theta) P(y_i|\pi) \\
&= \ln \prod_{i=1}^{N} \prod_{j=1}^{d} P(x_{ij}|y_i, \theta) P(y_i|\pi) \\
&= \ln \prod_{i=1}^{N} \prod_{j=1}^{d} \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{jkm}^{\mathbf{1}_{x_{ij}=k} \mathbf{1}_{y_i=m}} + \ln \prod_{i=1}^{N} \prod_{m=1}^{M} \pi_m^{\mathbf{1}_{y_i=m}} \\
&= \sum_{j=1}^{d} \sum_{m=1}^{M} \sum_{k=1}^{K} \ln \theta_{jkm}^{\sum_{i=1}^{N} \mathbf{1}_{x_{ij}=k} \mathbf{1}_{y_i=m}} + \sum_{m=1}^{M} \ln \pi_m^{\sum_{i=1}^{N} \mathbf{1}_{y_i=m}} \\
&= \sum_{j=1}^{d} \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{jkm} \ln \theta_{jkm} + \sum_{m=1}^{M} \alpha_m \ln \pi_m,
\end{aligned} \tag{5}$$

where $\alpha_{jkm} = \sum_{i=1}^{N} \mathbf{1}_{x_{ij}=k} \mathbf{1}_{y_i=m}$ and $\alpha_m = \sum_{i=1}^{N} \mathbf{1}_{y_i=m}$, $\forall j, k, m$. It is worth noting that

$$\sum_{k=1}^{K} \alpha_{jkm} = \sum_{i=1}^{N} \left(\sum_{k=1}^{K} \mathbf{1}_{x_{ij}=k} \right) \mathbf{1}_{y_i=m} = \sum_{i=1}^{N} \mathbf{1}_{y_i=m} = \alpha_m, \quad \forall j, m. \tag{6}$$

4 MLE

Based on the fact that $\sum_{k=1}^{K} \theta_{jkm} = 1$, there are $K-1$ independent parameters in $P(X_j|Y = m)$. Thus we can treat $\theta_{jKm} = 1 - \sum_{k=1}^{K-1} \theta_{jkm}$ as the dependent parameter. Similarly, there are $M - 1$ indpendent parameters in $P(Y)$, and we treat $\pi_M = 1 - \sum_{m=1}^{M-1} \pi_m$ as the dependent parameter.

Since the log-likelihood is a concave function w.r.t. θ and π, its global maximum is obtained by setting its derivative as 0, leading to

$$
\begin{aligned}
\frac{\partial \ell(\theta,\pi)}{\partial \theta_{jkm}} &= \frac{\alpha_{jkm}}{\theta_{jkm}} - \frac{\alpha_{jKm}}{1-\sum_{k=1}^{K-1}\theta_{jkm}} = \frac{\alpha_{jkm}}{\theta_{jkm}} - \frac{\alpha_{jKm}}{\theta_{jKm}} = 0 \\
\frac{\partial \ell(\theta,\pi)}{\partial \pi_{m}} &= \frac{\alpha_{m}}{\pi_{m}} - \frac{\alpha_{M}}{1-\sum_{m=1}^{M-1}\pi_{m}} = \frac{\alpha_{m}}{\pi_{m}} - \frac{\alpha_{M}}{\pi_{M}} = 0.
\end{aligned} \tag{7}
$$

Obviously,

$$
\hat{\theta}_{jkm} = \frac{\alpha_{jkm}}{\alpha_{jKm}}\hat{\theta}_{jKm}, \quad \hat{\pi}_m = \frac{\alpha_m}{\alpha_M}\hat{\pi}_M. \tag{8}
$$

Substituting (8) into the facts $\sum_{k=1}^{K}\theta_{jkm}=1$ and $\sum_{m=1}^{M}\pi_m=1$, gives rise to

$$
\hat{\theta}_{jKm} = \frac{\alpha_{jKm}}{\sum_{k=1}^{K}\alpha_{jkm}}, \quad \hat{\pi}_M = \frac{\alpha_M}{\sum_{m=1}^{M}\alpha_m}. \tag{9}
$$

By combing (6), (8) and (9), we reach our conclusion:

$$
\begin{aligned}
\hat{\theta}_{jkm} &= \frac{\alpha_{jkm}}{\sum_{k=1}^{K}\alpha_{jkm}} = \frac{\sum_{i=1}^{N}\mathbf{1}_{x_{ij}=k}\mathbf{1}_{y_i=m}}{\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbf{1}_{x_{ij}=k}\mathbf{1}_{y_i=m}} = \frac{\sum_{i=1}^{N}\mathbf{1}_{x_{ij}=k}\mathbf{1}_{y_i=m}}{\sum_{i=1}^{N}\mathbf{1}_{y_i=m}}, \quad k=1,2,...,K, \\
\hat{\pi}_m &= \frac{\alpha_m}{\sum_{m=1}^{M}\alpha_m} = \frac{\sum_{i=1}^{N}\mathbf{1}_{y_i=m}}{\sum_{m=1}^{M}\sum_{i=1}^{N}\mathbf{1}_{y_i=m}}, \quad m=1,2,...,M.
\end{aligned} \tag{10}
$$