Solutions to Quizzes in Lectures 7 and 8

Lu Sun

March 30, 2020

1 Solution to Quiz in Lecture 7

1.1 Probability Density Function

Suppose that we have a categorical random variable X with K states, i.e., $X \in \{1, 2, ..., K\}$. Let θ_k denote the probability of $X = k$ ($k = 1, 2, ..., K$), the probability density function is defined by

$$P(X|\theta) = \theta_1^{\mathbf{1}_{X=1}} \theta_2^{\mathbf{1}_{X=2}} \cdots \theta_K^{\mathbf{1}_{X=K}}, \tag{1}$$

where $\theta = \{\theta_1, \theta_2, ..., \theta_K\}$, and $\mathbf{1}_{(\cdot)}$ is the indicator function.

1.2 Likelihood Function

Given a training dataset $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, in which each sample x_i is an observation of X, the likelihood function becomes

$$\begin{aligned} L(\theta) &= P(\mathcal{D}|\theta) \\ &= P(x_1, x_2, ..., x_N|\theta) \\ &= \prod_{i=1}^{N} P(x_i|\theta) \\ &= \prod_{i=1}^{N} \theta_1^{\mathbf{1}_{x_i=1}} \theta_2^{\mathbf{1}_{x_i=2}} \cdots \theta_K^{\mathbf{1}_{x_i=K}} \\ &= \theta_1^{\sum_{i=1}^{N} \mathbf{1}_{x_i=1}} \theta_2^{\sum_{i=1}^{N} \mathbf{1}_{x_i=2}} \cdots \theta_K^{\sum_{i=1}^{N} \mathbf{1}_{x_i=K}} \\ &= \theta_1^{\alpha_1} \theta_2^{\alpha_2} \cdots \theta_K^{\alpha_K}, \end{aligned} \tag{2}$$

where α_k denotes the number of $X = k$ in the training dataset $\mathcal{D}$, thus $\alpha_k = \sum_{i=1}^{N} \mathbf{1}_{x_i=k}$, $\forall k$.

1.3 Prior Probability

If the prior of θ are from the Dirichlet$(\beta_1, \beta_2, ..., \beta_K)$, we have

$$P(\theta) = \frac{\theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \cdots \theta_K^{\beta_K-1}}{B(\beta_1, \beta_2, ..., \beta_K)}. \tag{3}$$

In (3), β_k ($\forall k$) is the hyperparameter of Dirichlet distribution, and $B(\cdot)$ denotes the beta distribution, that is irrelevant with θ.

1.4 Posterior Probability

By combining (2) and (3), log-posterior is formulated as follows:

$$\begin{aligned}
\ln P(\theta|\mathcal{D}) &\propto \ln\left(P(\mathcal{D}|\theta)P(\theta)\right) \\
&\propto \ln\left(\theta_1^{\alpha_1+\beta_1-1}\theta_2^{\alpha_2+\beta_2-1}\cdots\theta_K^{\alpha_K+\beta_K-1}\right) \\
&\propto \sum_{k=1}^{K}(\alpha_k+\beta_k-1)\ln\theta_k. \qquad (4)
\end{aligned}$$

Based on the fact that $\sum_{k=1}^{K}\theta_k = 1$, there are $K-1$ independent parameters in $\{\theta_1, \theta_2, ..., \theta_K\}$. Thus we can treat $\theta_K = 1 - \sum_{k=1}^{K-1}\theta_k$ as the dependent parameter. As the log-posterior is a concave function w.r.t. θ, its global maximum is obtained by setting its derivative equal to 0, leading to

$$\begin{aligned}
\frac{\partial \ln P(\theta|\mathcal{D})}{\partial \theta_k} &= \frac{\alpha_k+\beta_k-1}{\theta_k} - \frac{\alpha_K+\beta_K-1}{1-\sum_{k=1}^{K-1}\theta_k} \\
&= \frac{\alpha_k+\beta_k-1}{\theta_k} - \frac{\alpha_K+\beta_K-1}{\theta_K} \\
&= 0. \qquad (5)
\end{aligned}$$

Obviously,

$$\hat{\theta}_k = \frac{\alpha_k+\beta_k-1}{\alpha_K+\beta_K-1}\hat{\theta}_K. \qquad (6)$$

Substituting (6) into $\sum_{k=1}^{K}\theta_k = 1$, gives rise to

$$\hat{\theta}_K = \frac{\alpha_K+\beta_K-1}{\sum_{k=1}^{K}\alpha_k+\beta_k-1}. \qquad (7)$$

By combing (6) and (7), we reach our conclusion:

$$\hat{\theta}_k = \frac{\alpha_k+\beta_k-1}{\sum_{k=1}^{K}\alpha_k+\beta_k-1}, \quad k = 1, 2, ..., K. \qquad (8)$$

2 Solution to Quiz in Lecture 8

The solution is the MLE version of the above one, by replacing X and θ by Y and π, respectively.