# Maximum Likelihood Estimation For The Quadratic Discriminant Analysis

Saffet Gökçen Şen

May 18, 2020

Assume that there is a discrete generative model. A classification is to be made using this model. The probability density function of the output to be in the class $c$ is given as follows:

$$p\left(y=c|\mathbf{x},\boldsymbol{\theta}\right)=\frac{p\left(\mathbf{x}|y=c,\boldsymbol{\theta}\right)p\left(y=c|\boldsymbol{\theta}\right)}{p\left(\mathbf{x}|\boldsymbol{\theta}\right)} \tag{1}$$

If $p\left(\mathbf{x}|y=c,\boldsymbol{\theta}\right)$ is given by a multivariate Gaussian distribution, then this classification is called the quadratic discriminant classification. The maximum likelihood estimator for the quadratic discriminant classifier is to be derived. Hence, the likelihood of the data should be written first. The likelihood of a single data sample $\left(\mathbf{x}_i, y_i\right)$ is as follows:

$$p\left(\mathbf{x}_i, y_i|\boldsymbol{\theta}\right)=p\left(\mathbf{x}_i|y_i,\boldsymbol{\theta}\right)p\left(y_i|\boldsymbol{\theta}\right)=\prod_{c=1}^{C}p\left(y_i|\boldsymbol{\theta}_c\right)^{I(y_i=c)}\prod_{c=1}^{C}p\left(\mathbf{x}_i|y_i,\boldsymbol{\theta}_c\right)^{I(y_i=c)} \tag{2}$$

Assuming that the samples are independent and there are $N$ training samples, the likelihood of the data can now be written as follows:

$$p\left(D|\boldsymbol{\theta}\right)=\prod_{i=1}^{N}p\left(\mathbf{x}_i, y_i|\boldsymbol{\theta}\right) \tag{3}$$

For the sake of avoiding any numerical underflow and computational simplicity, the logarithm of the above expression is taken to obtain the log-likelihood:

$$\log p\left(D|\boldsymbol{\theta}\right)=\sum_{i=1}^{N}\log p\left(\mathbf{x}_i, y_i|\boldsymbol{\theta}\right)= \tag{4}$$

$$\sum_{i=1}^{N}\log\left[\prod_{c=1}^{C}p\left(y_i|\boldsymbol{\theta}_c\right)^{I(y_i=c)}\prod_{c=1}^{C}p\left(\mathbf{x}_i|y_i,\boldsymbol{\theta}_c\right)^{I(y_i=c)}\right]= \tag{5}$$

$$\sum_{i=1}^{N}\left[\sum_{c=1}^{C}\log\left[p\left(y_i|\boldsymbol{\theta}_c\right)^{I(y_i=c)}\right]+\sum_{c=1}^{C}\log\left[p\left(\mathbf{x}_i|y_i,\boldsymbol{\theta}_c\right)^{I(y_i=c)}\right]\right]= \tag{6}$$

$$\sum_{i=1}^{N}\left[\sum_{c=1}^{C}I\left(y_i=c\right)\log p\left(y_i|\boldsymbol{\theta}_c\right)+\sum_{c=1}^{C}I\left(y_i=c\right)\log p\left(\mathbf{x}_i|y_i,\boldsymbol{\theta}_c\right)\right]\Rightarrow \tag{7}$$

$$\log p\left(D|\boldsymbol{\theta}\right) = \sum_{c=1}^{C} N_c \log p\left(y_i = c|\boldsymbol{\theta}_c\right) + \sum_{c=1}^{C} \sum_{i:y_i=c} \log p\left(\mathbf{x}_i|y_i, \boldsymbol{\theta}_c\right) \tag{8}$$

The first term is maximized if the maximum likelihood estimator (MLE) for the class occurrences which is

$$\pi_{c_{MLE}} = \frac{N_c}{N} \tag{9}$$

is plugged in place of $p\left(y_i = c|\boldsymbol{\theta}_c\right)$. $N_c$ is the number of occurrences of the class $c$. The second term is maximized if the maximum likelihood estimators for the means and covariances of the multivariate Gaussian distributions for the probability densities of the input features conditioned on the class label are used. The MLE's for the mean and covariances are as follows:

$$\boldsymbol{\mu}_{c_{MLE}} = \frac{1}{N_c} \sum_{i:y_i=c} x_i \tag{10}$$

$$\boldsymbol{\Sigma}_{c_{MLE}} = \frac{1}{N_c} \sum_{i:y_i=c} \left(\mathbf{x}_i - \boldsymbol{\mu}_{c_{MLE}}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_{c_{MLE}}\right)^T \tag{11}$$

The maximum likelihood estimation measure for $p\left(y = c|\mathbf{x}, \boldsymbol{\theta}\right)$ is calculated without computing $p\left(\mathbf{x}|\boldsymbol{\theta}\right)$ since it is the same for all of the classes.

**References**

Machine Learning A Probabilistic Perspective, Kevin P. Murphy