Discriminant Analysis (supervised learning)

Marianna Bolla, Prof. DSc.

For the time being, except the regression (where we had a target variable depending on the predictor ones), we have discussed methods of the so-called *unsupervised learning*, when we retrieved information from our data without any preliminary assumption. Now, another method of the *supervised learning* is introduced, when we already have some preliminary knowledge about the classification of our data (made by an expert) and we want to reproduce (imitate) this classification based merely on multivariate measurements. In such situations we have a learning sample to build the artificial intelligence, and we test it on the same or on another sample. In this way, so-called expert systems are constructed, and in case of good performance, they can be used (with care) for automatic classification (for example, for medical diagnosis).

At the beginning, we have a p-dimensional sample of independent observations, but they are not identically distributed, rather form a mixture of k multivariate distributions, which are clearly distinguished by an expert in the so-called learning sample. For example, we have p clinical measurements of patients coming from k different diagnostic groups. If the measurements have something to do with the diagnosis, there is a hope that with some algorithm we are able to assign a patient to one of the groups merely based on his/her measurements. Based on the classes of the learning sample and some intuition, we are provided with the following knowledge:

- p-variate densities $f_1(\mathbf{x}), \dots, f_k(\mathbf{x})$ of the classes (usually these are multivariate Gaussian with estimated parameters);
- the prior probabilities $\pi_1, \dots, \pi_k$ of a randomly selected object belonging to the classes, $\sum_{i=1}^{k} \pi_i = 1$ (they are usually proportional to the sample sizes, but can as well correspond to the expert's intuition).

Our purpose is to find a partition $\mathcal{X}_1, \dots, \mathcal{X}_k$ of the p-dimensional sample space so that the obtained classes would, as much as possible, coincide with the original ones. Equivalently, we have to find a decision rule which decides the membership of an object based on its measurement $\mathbf{x} \in \mathbb{R}^p$.

Our algorithm minimizes the following average loss function:

$$L = \sum_{i=1}^{k} \pi_i L_i.$$

Here the average loss L_i is due to misclassifying objects of $\mathcal{X}_i$, defined as

$$L_i = \sum_{j=1}^{k} \int_{\mathcal{X}_j} r_{ij} f_i(\mathbf{x})\, d\mathbf{x}$$

where the risk $r_{ij} \geq 0$ of classifying an object of class i into class j is given for $i, j = 1, \dots, k$. Note that $r_{ii} = 0$ $(i = 1, \dots, k)$, otherwise they are not necessarily symmetric.

After some simple calculation

$$L = -\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x},$$

where $S_j(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i r_{ij} f_i(\mathbf{x})$ is called jth *discriminant informant*, and for given $\mathbf{x}$, we want to maximize

$$\sum_{j=1}^{k} \int_{\mathcal{X}_j} S_j(\mathbf{x})\, d\mathbf{x} \tag{1}$$

over the set of k-partitions of $\mathcal{X}$. A simple lemma guarantees that the maximum is attained with the following k-partition $\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*$: an object with measurement $\mathbf{x}$ is classified into $\mathcal{X}_i^*$ for which, $i = \operatorname{argmax}_j S_j(\mathbf{x})$ (such an i is not necessarily unique, but we can break ties arbitrarily).

Now, let us make the following simplification: $r_{ij} = 1$ for $i \neq j$ and of course, $r_{ii} = 0$ $(j = 1, \ldots, k)$. This assumption is quite natural: all misclassifications have the same risk, and there is no risk of a correct classification. By this, the discriminant informant simplifies to

$$S_j(\mathbf{x}) = -\sum_{i \neq j} \pi_i f_i(\mathbf{x}) = -\sum_{i=1}^{k} \pi_i f_i(\mathbf{x}) + \pi_j f_j(\mathbf{x}) = c + \pi_j f_j(\mathbf{x})$$

where the constant c does not depend on j, therefore instead $S_j(\mathbf{x})$ we can as well maximize $\pi_j f_j(\mathbf{x})$. That is, an object with measurement $\mathbf{x}$ is placed into the group j for which $\pi_j f_j(\mathbf{x})$ is maximum. Observe that this is nothing else but a *Bayesian decision rule*. Indeed, let Y denote the cluster membership, and $\mathbf{X}$ is the underlying p-variate random vector. Then for a randomly selected object with measurement $\mathbf{x}$ the following conditional probability is maximized in j:

$$\mathbf{P}(Y = j \,|\, \mathbf{X} = \mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\sum_{i=1}^{k} \pi_i f_i(\mathbf{x})}$$

where we used the Bayes rule. The maximization is equivalent to maximizing the numerator with respect to $j = 1, \ldots, k$. Further, if all the prior probabilities are equal, for given $\mathbf{x}$, we maximize $\mathbf{f}_j(\mathbf{x})$ which is just the *maximum likelihood discrimination rule*.

We can further simplify the maximization if the distribution of class j is $\mathcal{N}_p(\mathbf{m}_j, \mathbf{C}_j)$ with positive definite covariance matrix $\mathbf{C}_j$ $(j = 1, \ldots, k)$. Using the multivariate Gaussian density for the densities of the classes, for given $\mathbf{x}$, instead of $\pi_j f_j(\mathbf{x})$, we can maximize its natural logarithm. After leaving out the terms which do not depend on j, one can easily see that the following *quadratic informant* (quadratic function of the coordinates of $\mathbf{x}$) has to be maximized with respect to j:

$$Q_j(\mathbf{x}) = -\frac{1}{2} \ln |\mathbf{C}_j| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_j)^T \mathbf{C}_j^{-1} (\mathbf{x} - \mathbf{m}_j) + \ln \pi_j.$$

In the case of $\mathbf{C}_1 = \cdots = \mathbf{C}_k = \mathbf{C}$, we can disregard the terms which do not depend on $\mathbf{m}_j$, and the following *linear informant* will decide the group memberships:

$$L_j(\mathbf{x}) = \mathbf{m}_j^T \mathbf{C}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_j^T \mathbf{C}^{-1} \mathbf{m}_j + \ln \pi_j.$$

If $k = 2$, then we put an object with measurement $\mathbf{x}$ into the first group if $L_1(\mathbf{x}) \geq L_2(\mathbf{x})$ and to the second one, otherwise. That is, the sample space is separated into two parts by means of a hyperplane. It can be shown that in case of k groups, $k - 1$ hyperplanes will do this job.

We remark that the sample means and covariance matrices are usually estimated from the sub-samples after checking for multivariate normality.

In another approach, R. A. Fisher looked for the linear function $\mathbf{a}^T\mathbf{x}$, the coefficients of which maximize the ratio of the between-groups sum of squares to the within-groups sum of squares. That is, with the notation of the MANOVA,

$$\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \tag{2}$$

has to be maximized with respect to $\mathbf{a}$. Since the scale of $\mathbf{a}$ does not affect the above ratio, $\|\mathbf{a}\| = 1$ can be assumed. It can be proved that the vector $\hat{\mathbf{a}}$ which maximizes the above ratio is the unit-norm eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$, which is of rank at most $k - 1$ (since $\text{rank}(\mathbf{B}) = k-1$ in the general case). The function $\hat{\mathbf{a}}^T\mathbf{x}$ is called *Fischer's linear discriminant function* or the *first canonical variate*. Based on this, we allocate $\mathbf{x}$ into the group i if

$$\|\hat{\mathbf{a}}^T\mathbf{x} - \hat{\mathbf{a}}^T\bar{\mathbf{x}}_i\| \leq \|\hat{\mathbf{a}}^T\mathbf{x} - \hat{\mathbf{a}}^T\bar{\mathbf{x}}_j\| \quad \forall j \neq i$$

where $\bar{\mathbf{x}}_j$ is the sample mean of group j. In the $k = 2$ case, this rule is identical to that given by the linear informants, where $\mathbf{m}_1$ and $\mathbf{m}_2$ are estimated by the group means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, respectively. Note that this is not true in the $k > 2$ case.

Note that in the $k = 2$ case, $\hat{\mathbf{a}}$ is normal to the hyperplane discriminating the two groups.

Remark that successively, number of $\text{rank}(\mathbf{W}^{-1}\mathbf{B})$ canonical variates can be computed, which gives rise to differentiate between the groups in dimension $k-1$. Canonical variates also have important relation to the canonical correlations.

In practice, first we process the discrimination on the learning sample, and in case of good performance, we can apply the algorithm for a test sample of newcoming objects. In lack of a test sample, we can randomly select objects from the same learning sample with some resampling method, called *bootstrapping*. The performance itself is evaluated by the cross-classification of the objects: we calculate the $k \times k$ *confusion matrix* the ij-th entry of which is the number of objects classified into class i by the expert, and into class j by the algorithm.

Possible applications in artificial intelligence: image recognition, medical diagnostic systems; but can be used in market research and bankruptcy prediction or whether a customer will be default or not (based on price and other economic patterns).