Model-based classification (LDA, QDA, etc.)

Matias Salibian

29 October, 2018

Model-based classification

Consider the following setting for a classification problem: there are p explanatory variables (features), collected in a vector $\mathbf{X} \in \mathbb{R}^p$, along with a categorical variable G that takes one of K possible values in the set $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$. Given a training set $(g_1, \mathbf{x}_1)$, $(g_2, \mathbf{x}_2)$, …, $(g_n, \mathbf{x}_n)$ we are interested in constructing a classifier (i.e. a function $\hat{g} : \mathbf{R}^p \to \mathcal{C}$) with good prediction properties.

As we discussed in class, given a point $\mathbf{x}_0$, the optimal classification (with respect to the 0-1 loss) picks the class c_j with the highest conditional probability of occurring given $\mathbf{x}_0$. In symbols:

$$\hat{g}(\mathbf{x}_0) \, = \, \arg \max_{c_\ell \in \mathcal{C}} \, P\left(G = c_\ell | \, \mathbf{X} = \mathbf{x}_0 \right) \, .$$

If we know (can model reasonably) and can estimate all the conditional probabilities
$P(G = c_\ell | \mathbf{X} = \mathbf{x}_0)$ for all possible values of the features $\mathbf{x}_0 \in \mathbb{R}^p$ then we can estimate the optimal (with respect to the 0-1 loss) classifier. One way to model (and estimate) the above conditional probabilities is to do it indirectly by modeling the distribution of the vector of features $\mathbf{X}$ for each class, in symbols, the (conditional) distribution $\mathbf{X} | G = c_\ell$. If $f(\mathbf{x} | G = c_j)$ is the conditional density (or pmf) of the vector of explanatory variables given that the response is c_j, then we always have:

$$P\left(G = c_j | \, \mathbf{X} = \mathbf{x}_0 \right) \, = \, \frac{f(\, \mathbf{x}_0 \, | \, G = c_j) P(G = c_j)}{\sum_{\ell=1}^K \, f(\, \mathbf{x}_0 | G = c_\ell) P(G = c_\ell)} \, .$$

Hence, to find the value c_i with the largest $P(G = c_i | \mathbf{X} = \mathbf{x}_0)$ we only need to compute the numerators in the right hand side above for each $c_j \in \mathcal{C}$ and select the class with the largest value.

We now look at a simple example of the above general approach.

An example: LDA

One of the simplest models for $\mathbf{X} | G = c_\ell$ is a (multivariate) normal (Gaussian) distribution with mean $\mu_\ell \in \mathbb{R}^p$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ (the same for all $c_\ell \in \mathcal{C}$). In symbols:

$$\mathbf{X} | G = c_i \, \sim \, \mathcal{N}\left(\mu_\ell, \mathbf{\Sigma} \right) \, .$$

The density function for a $\mathcal{N}\left(\mu, \mathbf{\Sigma} \right)$ distribution is given by

$$f(\mathbf{x}) \, = \, \frac{1}{(2\pi)^p} \, \frac{1}{|\mathbf{\Sigma}|^{p/2}} \, \exp\Big((\mathbf{x} - \mu)^\top \, \mathbf{\Sigma}^{-1} \, (\mathbf{x} - \mu) \Big) \, .$$

In order to use all this machinery we need to estimate the density functions for all the $\mathcal{N}\left(\mu_j, \mathbf{\Sigma}\right)$ distributions, and then, given a point $\mathbf{x}_0$, evaluate them and choose the class with the largest value of $f(\,\mathbf{x}_0 \,|\, G = c_j)P(G = c_j)$.

To estimate the above densities we need to estimate the following parameters and probabilities:

- $\mu_i, i = 1, \ldots, K$;
- $\mathbf{\Sigma}$; and
- $P(G = c_j), j = 1, \ldots, K.$

The MLE estimators for the mean vectors are the sample means of the features in each class in the training set, and the common covariance matrix (across classes) can be estimated by taking a weighted average of the sample covariance matrix in each class using the data in the training set. Finally, the probabilities of each class can be estimated with the observed frequency of each class in the training set. In symbols we have

- For class $c_i \in \mathcal{C}$ we have

  $$\hat{\mu}_i \;=\; \frac{1}{n_i} \sum_{j \in \mathcal{N}_i} \mathbf{x}_j \,,$$

  where $\mathcal{N}_i$ is the set of indices of observations in the training set with $g = c_i$, and n_i is the cardinal of that set.
- The common matrix $\mathbf{\Sigma}$ can be estimated with

  $$\widehat{\mathbf{\Sigma}} \;=\; \frac{1}{n-K} \sum_{\ell=1}^{K} (n_\ell - 1) \widehat{\mathbf{\Sigma}}_\ell \,,$$

  where

  $$\widehat{\mathbf{\Sigma}}_\ell \;=\; \frac{1}{n_\ell - 1} \sum_{j \in \mathcal{N}_\ell} \left(\mathbf{x}_j - \bar{\mathbf{x}}_\ell\right) \left(\mathbf{x}_j - \bar{\mathbf{x}}_\ell\right)^\top \;.$$

  and

  $$\bar{\mathbf{x}}_\ell \;=\; \frac{1}{n_\ell} \sum_{i \in \mathcal{N}_\ell} \mathbf{x}_i \,.$$