Linear Methods for Classification

Robert Schmidt

Abstract

The following are notes on the key results from the **Elements of Statistical Learning** text. They were primarily derived from course notes and readings in the Stanford STATS 315: *Modern Applied Statistics* series.

Contents

1 Discriminant analysis rule derivations

We need an expression for $\mathbb{P}(G \mid \mathbf{X})$ in order to perform classification. Adopting the notation in ESL, let:

$$\begin{cases} f_k(x) \to \text{class-conditional density of } \mathbf{X} \text{ in class } G = k \\ \pi_k \to \text{ prior probability of class } k \\ \sum_{k=1}^K \pi_k = 1 \end{cases}$$

Via Bayes rule, the desired probability is:

$$\mathbb{P}(G = k \mid \mathbf{X} = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^K f_\ell(x)\pi_\ell}$$

In the case that each class density is multivariate Gaussian, the densities are of the form:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\Bigg(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)\Bigg)$$

Here, we consider the two-class case ($K = 2$). In this case, we examine the log ratio of the conditional probabilities to find the classification boundary:

$$L_{k\ell} = \log\left(\frac{\mathbb{P}(G = k \mid \mathbf{X} = x)}{\mathbb{P}(G = \ell \mid \mathbf{X} = x)}\right) = \log\frac{f_k(x)}{f_\ell(x)} + \log\frac{\pi_k}{\pi_\ell}$$

The degree to which this formula can be simplified depends upon the assumptions on the covariance matrix $\mathbf{\Sigma}_k$. To this end, we will consider LDA and QDA.

1.1 LDA rule derivation

For LDA, we assume the f_k are multivariate normal with common covariance $\mathbf{\Sigma}$ and separate μ_k. In other words:

$$\begin{aligned} f_k &= c \cdot \exp\Bigg(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k)\Bigg) \\ f_\ell &= c \cdot \exp\Bigg(-\frac{1}{2}(x-\mu_\ell)^T \mathbf{\Sigma}^{-1}(x-\mu_\ell)\Bigg) \end{aligned}$$

Here, the leading coefficient terms are the same c since the two classes share a common covariance matrix. This greatly simplifies the following calculations. First, let us examine the ratio of the class densities:

$$\begin{aligned} \frac{f_k}{f_\ell} &= \exp\Bigg(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_\ell)^T \mathbf{\Sigma}^{-1}(x-\mu_\ell)\Bigg) \\ \log\frac{f_k}{f_\ell} &= -\frac{1}{2}\Bigg[(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k) - \frac{1}{2}(x-\mu_\ell)^T \mathbf{\Sigma}^{-1}(x-\mu_\ell)\Bigg] \\ &= -\frac{1}{2}\Bigg[x^T\mathbf{\Sigma}^{-1}x - 2\mu_k^T\mathbf{\Sigma}^{-1}x + \mu_k^T\mathbf{\Sigma}^{-1}\mu_k - x^T\mathbf{\Sigma}^{-1}x + 2\mu_\ell^T\mathbf{\Sigma}^{-1}x - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell\Bigg] \\ &= -\frac{1}{2}\Bigg[-2\mu_k^T\mathbf{\Sigma}^{-1}x + \mu_k^T\mathbf{\Sigma}^{-1}\mu_k + 2\mu_\ell^T\mathbf{\Sigma}^{-1}x - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell\Bigg] \\ &= x^T\mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell) - \frac{1}{2}\Bigg[\mu_k^T\mathbf{\Sigma}^{-1}\mu_k - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell\Bigg] \end{aligned}$$

For a moment, let us consider in more detail the $-\frac{1}{2}\Bigg[\mu_k^T\mathbf{\Sigma}^{-1}\mu_k - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell\Bigg]$ term.

This should simplify to the $-\frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)$ term found in ESL (4.9).
Let us prove that the two are equivalent:

$$\begin{aligned} -\frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell) &= -\frac{1}{2}[\mu_k^T\mathbf{\Sigma}^{-1}\mu_k - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell + \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_k - \mu_k^T\mathbf{\Sigma}^{-1}\mu_\ell] \\ &= -\frac{1}{2}[\mu_k^T\mathbf{\Sigma}^{-1}\mu_k - \mu_\ell^T\mathbf{\Sigma}^{-1}\mu_\ell] \end{aligned}$$

Hence, we arrive at the final LDA expression from ESL equation (4.9):

$$\boxed{L_{k\ell} = \log\frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell) + x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)}$$

1.2 QDA rule derivation

2 Discriminant analysis computations

2.1 LDA computation: sphering the data

2.2 Computations for reduced-rank LDA

2.3 Rayleigh quotient and canonical discriminant analysis

3 LDA vs. least squares fit

3.1 Least squares regression coefficient is identical to LDA coefficient, up to scalar multiple

3.2 Difference between LDA and OLS coefficient

4 Logistic regression

4.1 Derivation of logistic rule

4.2 Two-class algorithm for logistic regression

4.3 Newton-Raphson IRLS algorithm