Linear and Quadratic Discriminant Analysis

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Linear classification generalities

Suppose y takes values in $\mathcal{G} = \{1, 2, \ldots, K\}$, or equivalently that $\mathbf{y}$ is a K-variate set of indicators, $y_k = I[y = k]$. We consider methods of producing prediction/classification rules $\widehat{f}(\mathbf{x})$ taking values in $\mathcal{G}$ (and mostly ones) that have sets $\left\{\mathbf{x} \in \Re^p | \widehat{f}(\mathbf{x}) = k\right\}$ with boundaries that are defined (at least piece-wise) by linear equalities

$$\mathbf{x}'\boldsymbol{\beta} = c \tag{1}$$

We consider several means of choosing those boundaries. The first two of these have classical "statistical" origins (linear discriminant analysis and logistic regression). Then we consider ones that have geometric origins (notions of separating hyperplanes and "support vector" classifiers).

Common Sigma class-conditional models

Suppose that for $(\mathbf{x}, y) \sim P$, $\pi_k = P[y = k]$ and the conditional distribution of $\mathbf{x}$ on $\Re^p$ given that $y = k$ is $\text{MVN}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, i.e. the conditional pdf is

$$g_k(\mathbf{x}) = (2\pi)^{-p/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right)$$

For future reference, note that under these assumptions

$$\begin{aligned} &\ln\left(\frac{P[y = k|\mathbf{x}]}{P[y = l|\mathbf{x}]} \right) \\ &\quad = \ln\left(\frac{\pi_k}{\pi_l} \right) - \frac{1}{2} \boldsymbol{\mu}_k' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \frac{1}{2} \boldsymbol{\mu}_l' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l + \mathbf{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) \end{aligned} \qquad (2)$$

Optimal decision boundaries

Based on the MVN (with common covariance matrix) form of the g_k, a theoretically optimal predictor/classifier/decision rule is

$$f(\mathbf{x}) = \arg\max_k \left[\ln(\pi_k) - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \right]$$

and boundaries between regions in $\Re^p$ where $f(\mathbf{x}) = k$ and $f(\mathbf{x}) = l$ are subsets of the sets

$$\left\{ \mathbf{x} \in \Re^p \,|\, \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) = -\ln\left(\frac{\pi_k}{\pi_l}\right) + \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_l'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_l \right\}$$

i.e. are defined by equalities of the form (1).

K=3 and *p=2* example

The figure below illustrates this in a simple $K = 3$ case where $p = 2$.

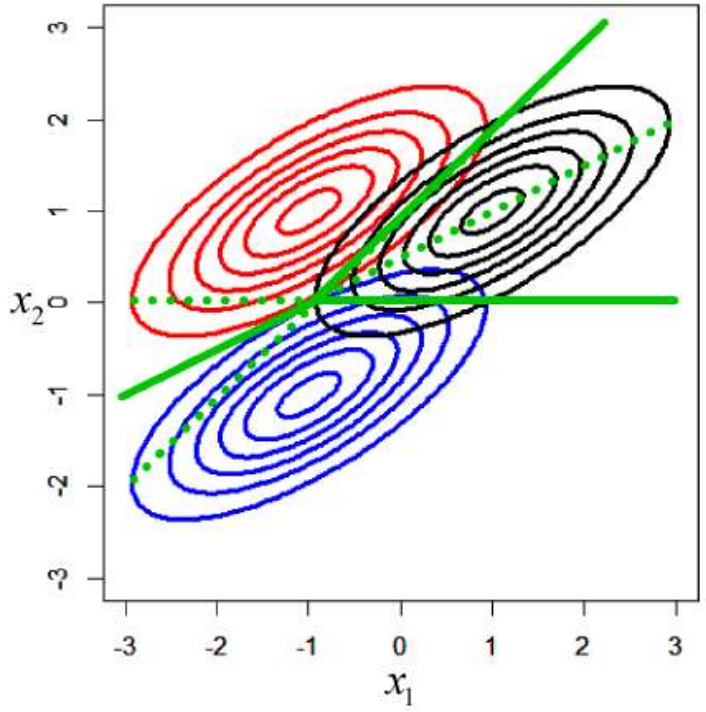

Figure: Contours of $K = 3$ bivariate normal pdfs and corresponding linear (equal class probabilities) classification boundaries.

Quadratic Discriminant Analysis

If class covariance matrices are allowed to vary (class-conditional distribution k with covariance matrix $\mathbf{\Sigma}_k$) a theoretically optimal predictor/decision rule is

$$f(\mathbf{x}) = \arg\max_k \left[\ln(\pi_k) - \frac{1}{2}\ln(\det \mathbf{\Sigma}_k) - -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \right]$$

and boundaries between regions in $\Re^p$ where $f(\mathbf{x}) = k$ and $f(\mathbf{x}) = l$ are subsets of the sets

$$\{\mathbf{x} \in \Re^p \mid \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)' \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)' \mathbf{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) = - \ln\left(\frac{\pi_k}{\pi_l}\right) - \tfrac{1}{2}\ln(\det \mathbf{\Sigma}_k) + \tfrac{1}{2}\ln(\det \mathbf{\Sigma}_l)\}$$

Unless $\mathbf{\Sigma}_k = \mathbf{\Sigma}_l$ this kind of set is a quadratic surface in $\Re^p$, not a hyperplane. One gets (not linear, but) Quadratic Discriminant Analysis.

LDA-QDA "compromise"

In order to use LDA or QDA, one must estimate the vectors μ_k and the covariance matrix Σ or matrices Σ_k from the training data. Estimating K potentially different matrices Σ_k requires estimation of a very large number of parameters. So thinking about QDA versus LDA, one is again in the situation of needing to find the level of predictor complexity that a given data set will support. QDA is a more flexible/complex method than LDA, but using it in preference to LDA increases the likelihood of over-fit.

One idea that has been offered as a kind of continuous compromise between LDA and QDA is for $\alpha \in (0,1)$ to use

$$\widehat{\Sigma}_k(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\widehat{\Sigma}_{\text{pooled}}$$

in place of $\widehat{\Sigma}_k$ in QDA.

Other LDA-QDA "compromises"

This kind of "compromise" thinking even suggests as an estimate of a covariance matrix common across k

$$\widehat{\mathbf{\Sigma}}(\gamma) = \gamma\widehat{\mathbf{\Sigma}}_{\text{pooled}} + (1-\gamma)\widehat{\sigma^2}\mathbf{I}$$

for $\gamma \in (0,1)$ and $\widehat{\sigma^2}$ an estimate of variance pooled across groups k and then across coordinates of $\mathbf{x}$, j, in LDA. Combining these two ideas, one might even invent a two-parameter set of fitted covariance matrices

$$\widehat{\mathbf{\Sigma}}_k(\alpha,\gamma) = \alpha\widehat{\mathbf{\Sigma}}_k + (1-\alpha)\left(\gamma\widehat{\mathbf{\Sigma}}_{\text{pooled}} + (1-\gamma)\widehat{\sigma^2}\mathbf{I}\right)$$

for use in QDA. Employing these in LDA or QDA provides the flexibility of choosing a complexity parameter or parameters and potentially improving classification performance.