Modern Multivariate Statistical Learning-29

Linear Methods of Classification Part 1: Linear and Quadratic Discriminant Analysis

Stephen Vardeman

Analytics Iowa LLC

April 2014

"Linear" Classification Rules

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

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

A potential method here is to regress the K indicator variables y_k onto $\mathbf{x}$ (producing least squares regression vector coefficients $\widehat{\boldsymbol{\beta}}_k$) and to employ

$$\widehat{f}(\mathbf{x}) = \arg\max_k \widehat{f}_k(\mathbf{x}) = \arg\max_k \mathbf{x}'\widehat{\boldsymbol{\beta}}_k$$

This fails miserably because of the possibility of "masking" if $K > 2$, and so we consider three smarter alternatives: Linear (and Quadratic) Discriminant Analysis, Logistic Regression, and direct search for separating hyperplanes. The first two of these have classical "statistical" origins.

Linear Discriminant Analysis

MVN Class-Conditional Modeling With Common Covariance Matrix

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

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

Linear Discriminant Analysis

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).

"Quadratic" Discriminant Analysis

General MVN Class-Conditional Modeling and Classification

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

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

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 and QDA Classifier Complexity

Compromise Methods

In order to use LDA or QDA, one must estimate the vectors $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}$ or matrices $\boldsymbol{\Sigma}_k$ from the training data. Estimating K potentially different matrices $\boldsymbol{\Sigma}_k$ 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

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

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

LDA and QDA Classifier Complexity

Compromise Methods

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.

LDA Detail

"Sphering"

Returning specifically to LDA, let

$$\bar{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_k$$

and note that one is free to replace $\mathbf{x}$ and all K means $\boldsymbol{\mu}_k$ with respectively

$$\mathbf{x}^* = \boldsymbol{\Sigma}^{-1/2} \left(\mathbf{x} - \bar{\boldsymbol{\mu}} \right) \quad \text{and} \quad \boldsymbol{\mu}_k^* = \boldsymbol{\Sigma}^{-1/2} \left(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}} \right)$$

This produces

$$\ln \left(\frac{P\left[y = k | \mathbf{x}^* = \mathbf{u}^* \right]}{P\left[y = l | \mathbf{x}^* = \mathbf{u}^* \right]} \right) = \ln \left(\frac{\pi_k}{\pi_l} \right) - \frac{1}{2} \left\| \mathbf{u}^* - \boldsymbol{\mu}_k^* \right\|^2 + \frac{1}{2} \left\| \mathbf{u}^* - \boldsymbol{\mu}_l^* \right\|^2$$

LDA Detail

1st Sphered Form of LDA

In sphered form, the theoretically optimal (LDA) classifier/decision rule can be described as

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

That is, in terms of $\mathbf{x}^*$, optimal decisions are based on ordinary Euclidian distances to the transformed means $\boldsymbol{\mu}_k^*$. Further, this form can often be made even simpler.

LDA Detail

2nd Sphered Form of LDA

The μ_k^* typically span a subspace of $\Re^p$ of dimension $\min(p, K-1)$. For

$$\underset{p\times K}{\mathbf{M}} = (\mu_1^*, \mu_2^*, \ldots, \mu_K^*)$$

let $\mathbf{P_M}$ be the $p \times p$ projection matrix projecting onto the column space of $\mathbf{M}$ in $\Re^p$ $(C(\mathbf{M}))$. Then because $(\mathbf{P_M}\mathbf{x}^* - \mu_k^*) \in C(\mathbf{M})$ and $(\mathbf{I} - \mathbf{P_M})\mathbf{x}^* \in C(\mathbf{M})^\perp$

$$\begin{aligned}\|\mathbf{x}^* - \mu_k^*\|^2 &= \|(\mathbf{P_M}\mathbf{x}^* - \mu_k^*) + (\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2 \\ &= \|\mathbf{P_M}\mathbf{x}^* - \mu_k^*\|^2 + \|(\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2\end{aligned}$$

Since $\|(\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2$ doesn't depend upon k, an optimal classifier/ decision rule can be described as

$$f(\mathbf{x}) = \arg\max_k \left[\ln(\pi_k) - \frac{1}{2}\|\mathbf{P_M}\mathbf{x}^* - \mu_k^*\|^2\right]$$

Optimal classification can thus be described in terms of the projection of $\mathbf{x}^*$ onto $C(\mathbf{M})$ and its distances to the μ_k^*.

"Reduced Rank" LDA

Eigen Analysis for Covariance Matrix of Sphered Class Means

The matrix $\frac{1}{K}\mathbf{M}\mathbf{M}'$ is the (typically rank $\min(p, K-1)$) sample covariance matrix of the $\boldsymbol{\mu}_k^*$ and has an eigen decomposition as

$$\frac{1}{K}\mathbf{M}\mathbf{M}' = \mathbf{V}\mathbf{D}\mathbf{V}'$$

for

$$\mathbf{D} = \mathbf{diag}\left(d_1, d_2, \ldots, d_p\right)$$

where

$$d_1 \geq d_2 \geq \cdots \geq d_p$$

are the eigenvalues and the columns of $\mathbf{V}$ are orthonormal eigenvectors corresponding in order to the (successively smaller) eigenvalues. These $\mathbf{v}_k$ with $d_k > 0$ specify linear combinations of the coordinates of the $\boldsymbol{\mu}_l^*$, $\langle \mathbf{v}_k, \boldsymbol{\mu}_l^* \rangle$, with the largest sample variances subject to the constraints that $\|\mathbf{v}\| = 1$ and $\langle \mathbf{v}_l, \mathbf{v}_k \rangle = 0$ for all $l < k$. (These $\mathbf{v}_k$ are $\perp$ vectors in successive directions of most important unaccounted-for spread of the $\boldsymbol{\mu}_k^*$.) This suggests the possibility of "reduced rank" LDA.

Reduced Rank LDA

Dimension Reduction in LDA

That is, for $l \leq \text{rank}\left(\mathbf{MM}'\right)$ define

$$\mathbf{V}_l = \left(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_l\right)$$

let

$$\mathbf{P}_l = \mathbf{V}_l \mathbf{V}_l'$$

be the matrix projecting onto $C\left(\mathbf{V}_l\right)$ in $\Re^p$. A possible "reduced rank" approximation to the theoretically optimal LDA classification rule is

$$f_l\left(\mathbf{x}\right) = \arg\max_k \left[\ln\left(\pi_k\right) - \frac{1}{2}\left\|\mathbf{P}_l\mathbf{x}^* - \mathbf{P}_l\boldsymbol{\mu}_k^*\right\|^2\right]$$

and l becomes a complexity parameter that one might optimize to regularize the method.

Reduced Rank LDA

Graphical Representation

Note also that for $\mathbf{w} \in \Re^p$

$$\mathbf{P}_l \mathbf{w} = \sum_{k=1}^{l} \langle \mathbf{v}_k, \mathbf{w} \rangle \mathbf{v}_k$$

For purposes of graphical representation of what is going on in these computations, one might replace the p coordinates of $\mathbf{x}$ and the means $\boldsymbol{\mu}_k$ with the l coordinates of

$$\left(\langle \mathbf{v}_1, \mathbf{x}^* \rangle, \langle \mathbf{v}_2, \mathbf{x}^* \rangle, \ldots, \langle \mathbf{v}_l, \mathbf{x}^* \rangle \right)' \tag{3}$$

and of the

$$\left(\langle \mathbf{v}_1, \boldsymbol{\mu}_k^* \rangle, \langle \mathbf{v}_2, \boldsymbol{\mu}_k^* \rangle, \ldots, \langle \mathbf{v}_l, \boldsymbol{\mu}_k^* \rangle \right)' \tag{4}$$

It seems to be essentially ordered pairs of these coordinates that are plotted in HTF in their Figures 4.8 and 4.11.

Reduced Rank LDA

Graphical Representation of Reduced Rank LDA

Regarding this graphical method, we need to point out that since any eigenvector $\mathbf{v}_k$ could be replaced by $-\mathbf{v}_k$ without any fundamental effect in the above development, the vector (3) and all of the vectors (4) could be altered by multiplication of any particular set of coordinates by -1. (Whether a particular algorithm for finding eigenvectors produces $\mathbf{v}_k$ or $-\mathbf{v}_k$ is not fundamental, and there seems to be no standard convention in this regard.) It appears that the pictures in HTF might have been made using the `R` function `lda` and its choice of signs for eigenvectors.

Use of Features in LDA

Basis Functions/Transforms

The form $\mathbf{x}'\boldsymbol{\beta}$ is (of course and by design) linear in the coordinates of $\mathbf{x}$. An obvious natural generalization of this discussion is to consider discriminants that are linear in some (non-linear) functions of the coordinates of $\mathbf{x}$. This is simply choosing some M basis functions/ transforms/features $h_m(\mathbf{x})$ and replacing the p coordinates of $\mathbf{x}$ with the M coordinates of $(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_M(\mathbf{x}))$ in the development of LDA.

Of course, upon choosing basis functions that are all coordinates, squares of coordinates, and products of coordinates of $\mathbf{x}$, one produces *linear* (in the basis functions) discriminants that are general *quadratic* functions of $\mathbf{x}$. Or, data-dependent basis functions that are $K(\cdot, \mathbf{x}_i)$ for some kernel function K produces "LDA" that depends upon the value of some "basis function network." The possibilities opened here are myriad and (as always) "the devil is in the details."