Linear Discriminant Analysis, Part I

Patrick Breheny

September 15

Introduction

- As mentioned previously, we are interested in estimating $\Pr(G|\mathbf{x})$
- An obvious way to proceed is via Bayes' Rule:

$$\Pr(G = k|\mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_l f_l(\mathbf{x})\pi_l},$$

where

  - f_k is the density of the explanatory variables among the elements of class k
  - π_k is the marginal (or prior) probability of being in class k

Normal density model

- If we are going to apply this idea, we are going to have to estimate all the class densities $\{f_k\}$
- A straightforward way to proceed is to assume that f_k is multivariate normal:

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

- To simplify things, we will begin by assuming equal variances across the classes: *i.e.*, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$
- In later lectures, we will consider relaxing this assumption

Log probability ratio

- We will now derive our main result for today: the discriminant function and classification rules for the preceding approach
- The proof is simplified by first stating the following lemma: for any symmetric matrix $\mathbf{A}$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{y}^T\mathbf{A}\mathbf{y} = (\mathbf{x}+\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y})$$

- **Theorem:** Suppose the class densities $\{f_k\}$ are multivariate normal with common variance; then

$$\log\frac{\Pr(G=k|\mathbf{x})}{\Pr(G=l|\mathbf{x})} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l) + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)$$

Discriminant function

- **Corollary:** Suppose the class densities $\{f_k\}$ are multivariate normal with common variance; then the discriminant function for the above approach is

$$\delta_k(\mathbf{x}) = \log \pi_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k$$

- Note that this function is linear in $\mathbf{x}$; the above function is therefore a *linear discriminant function*, hence the name *linear discriminant analysis* (LDA) for this approach to modeling $\Pr(G|\mathbf{x})$
- The linearity of $\delta_k(\mathbf{x})$ means that all decision boundaries between any two classes k and l are linear in $\mathbf{x}$ (in p dimensions, a *hyperplane*)

Estimation

Of course, we do not know π_k, $\boldsymbol{\mu}_k$, or $\boldsymbol{\Sigma}$, so we will have to estimate them:

$$\hat{\pi}_k = \frac{n_k}{n}, \text{ where } n \text{ is the number of observations}$$

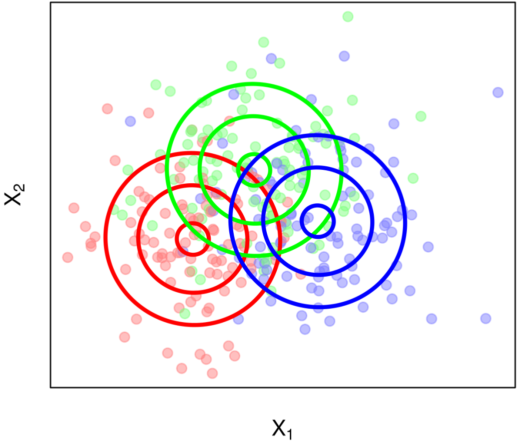
$$\text{and } n_k \text{ is the number of observations in class } k$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{\{i:g_i=k\}} \mathbf{x}_i$$

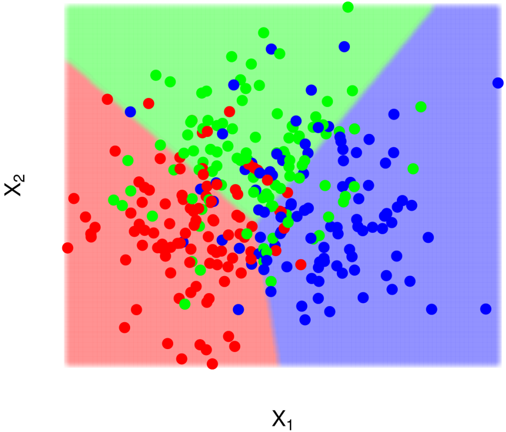
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-K} \sum_k \sum_{\{i:g_i=k\}} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T,$$

where the factor of $n - K$ in the denominator ensures that $\hat{\boldsymbol{\Sigma}}$ is an unbiased estimator of $\boldsymbol{\Sigma}$

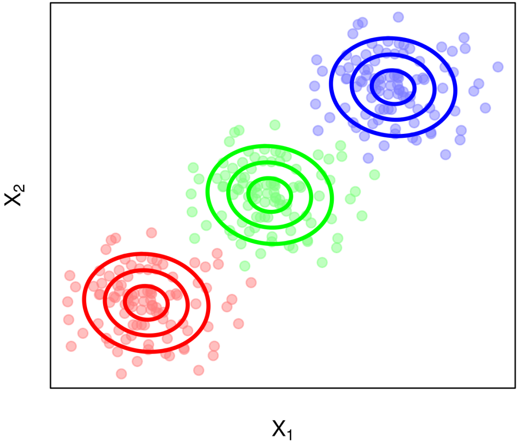
Example #1: Class densities



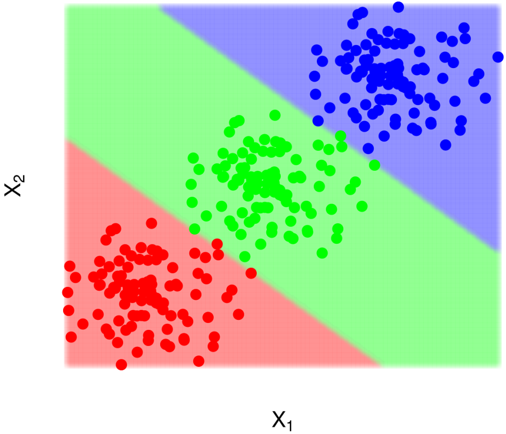
Example #1: Decision boundaries



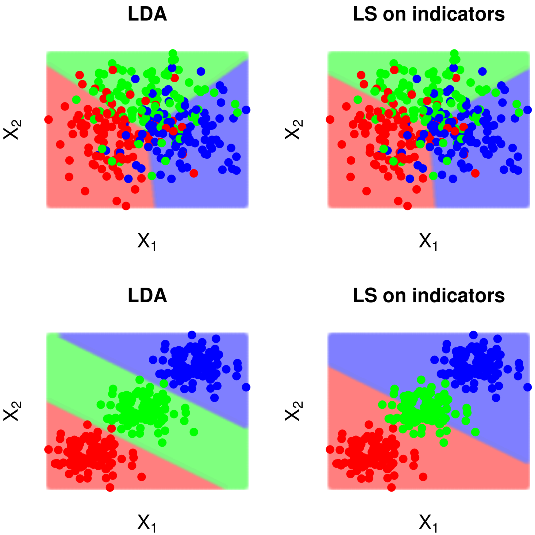
Example #2: Class densities



Example #2: Decision boundaries



Comparison with Linear Regression of Indicators



Remarks

- In the first example, the two approaches give quite similar results
- The second example illustrates that, unlike linear regression, LDA does not suffer from the masking problem
- Next time, we will apply LDA to some real data and compare it with logistic/multinomial regression