Lecture 9: $K \geq 2$ Classes and k-Nearest Neighbours

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Review of Linear Discriminant Analysis
2. $K \geq 2$ Classes
3. k-Nearest Neighbours

Normal Model for $\mathbf{X}|Y = c$

- Model the conditional distribution of $\mathbf{X} \mid Y$ instead of $Y \mid \mathbf{X}$
- For example, we can assume that

$$\mathbf{X} \mid Y = c_j \sim \text{MN}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}\right)$$

(MN = multivariate normal, with dimension the number of variables in $\mathbf{X}$)

- The classes differ in their $\mathbf{X}$ mean vectors
- The class distributions are estimated by

$$\hat{f}_j(\mathbf{X}) \sim \text{MN}\left(\hat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}\right)$$

using the sample mean of each group and the pooled sample covariance matrix

- We can then find, for a given $\mathbf{x}$, the class j that has the largest $\hat{f}_j(\mathbf{x})\, p_j$

Fisher's Linear Discriminant Analysis for **NORMAL** Populations

Writing f_1 for MN $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and f_2 for MN $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ then

$$f_1(\mathbf{x})\, p_1 \; > \; f_2(\mathbf{x})\, p_2 \quad \Leftrightarrow \quad \ln\left(\frac{f_1(\mathbf{x})\, p_1}{f_2(\mathbf{x})\, p_2}\right) > 0 \quad \Leftrightarrow \quad \mathbf{a}^T\mathbf{x} + b \; > \; 0$$

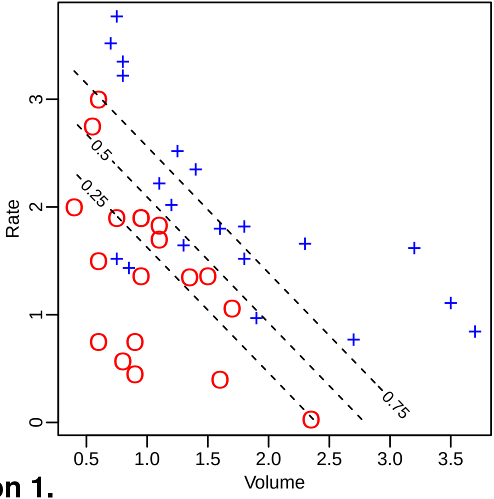
for some vector $\mathbf{a} \in \mathbb{R}^p$ and number $b \in \mathbb{R}$. In other words, boundaries between classes are **linear.** Furthermore, we can estimate this linear boundary because

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$$

and

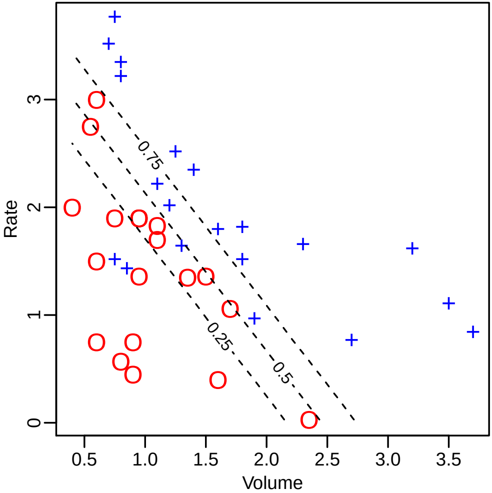
$$b = -\frac{1}{2} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right) - \ln\left(\frac{p_2}{p_1}\right)$$

Vaso Constriction Data: LDA Fit



Clicker question 1.

Vaso Constriction Data: Logistic Fit



LDA Classification Rule for NORMAL Populations

Note that if f_1 is MN $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and f_2 is MN $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ then

$$f_1(\mathbf{x})\, p_1 \; > \; f_2(\mathbf{x})\, p_2 \quad \Leftrightarrow \quad \mathbf{a}^T \mathbf{x} + b \; > \; 0$$

- Boundaries between classes are **linear.**
- Furthermore

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}\, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

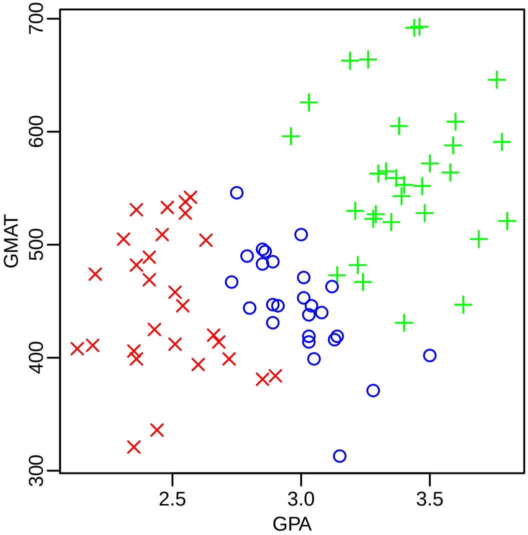
$$b = -\frac{1}{2}\, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\, \boldsymbol{\Sigma}^{-1}\, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\left(\frac{p_2}{p_1}\right)$$

- There is nothing special about having two classes. We can do the same with more than 2 classes

Admissions Data

- Admissions data from a Graduate School of Business
- From Applied Mulivariate Statistical Analysis by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
- The data are in the file T11-6.DAT
- 3 classes
  - $y = 1$ “admit student” (plotted as $+$)
  - $y = 2$ “do not admit student” (plotted as $\times$)
  - $y = 3$ “borderline” (plotted as $\circ$)
- 2 explanatory variables: GPA and GMAT, i.e., $\mathbf{x} = (\text{GPA}, \text{GMAT})$

Admissions Data



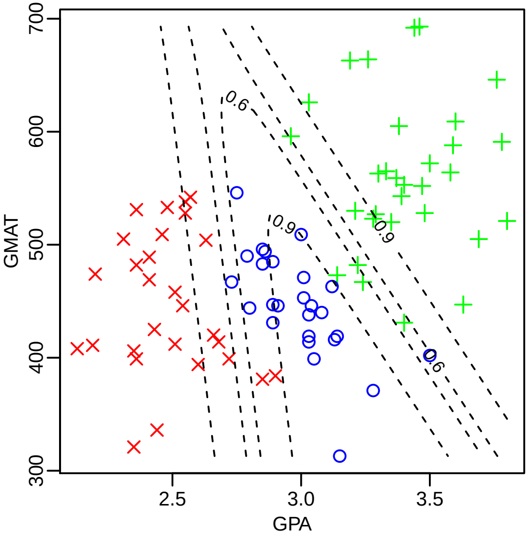
Admissions Data: LDA

- 3 classes, so 3 multivariate normal distributions f_j for $\mathbf{X}$, conditional on the class
- As before we have prior class probabilities p_j
- The 3 class codes are $c_1 = 1, c_2 = 2, c_3 = 3$
- Math is essentially the same as for 2 classes
  - $\Pr(Y = c_j \,|\, \mathbf{X} = \mathbf{x}) \propto f_j(\mathbf{x})\, p_j$
  - LDA therefore gives posterior probabilities

$$\Pr(Y = c_j \,|\, \mathbf{X} = \mathbf{x}) = \frac{f_j(\mathbf{x})\, p_j}{\sum_{m=1}^{3} f_m(\mathbf{x})\, p_m}$$

  - $\hat{y}(\mathbf{x})$ is the c_j with the largest estimated $\Pr(Y = c_j \,|\, \mathbf{X} = \mathbf{x})$

Admissions Data: LDA Maximum Probability



LDA: Drawbacks

- Assuming normality leads to a certain class of estimators for the boundaries and the conditional class probabilities

$$\hat{\mathbf{a}}^T \mathbf{x} + \hat{b} > 0$$

where

$$\hat{\mathbf{a}} = \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

and

$$\hat{b} = -\frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \hat{\mathbf{\Sigma}}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) - \ln\left(\frac{p_2}{p_1}\right)$$

- Works well for many applications, but ...
normality may not be reasonable, covariance matrix estimation is not robust to outliers, ...

Logistic Classification

- When we have $K > 2$ classes we propose
  - For classes $j = 1, \ldots, K-1$

$$\eta_j(\mathbf{x}_i) = \beta_{0j} + \boldsymbol{\beta}_j^T \mathbf{x}_i$$

  - For class K

$$\eta_K(\mathbf{x}_i) = 1$$

- Furthermore we take

$$\Pr\left(Y = c_j \,|\, \mathbf{x}\right) = \frac{\exp\left(\eta_j(\mathbf{x}_i)\right)}{1 + \sum_{m=1}^{K-1} \exp\left(\eta_m(\mathbf{x}_i)\right)} \quad \text{for } j = 1, \ldots, K$$

Logistic Classification

- Under this model, again the boundary between any two classes is always linear

$$\ln\left(\frac{\Pr\left(Y = c_j \,|\, \mathbf{x}\right)}{\Pr\left(Y = c_\ell \,|\, \mathbf{x}\right)}\right) = \left(\beta_{0j} - \beta_{0\ell}\right) + \left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\ell\right)^T \mathbf{x}$$

- Hence

$$\Pr\left(Y = c_j \,|\, \mathbf{x}\right) = \Pr\left(Y = c_\ell \,|\, \mathbf{x}\right) \Leftrightarrow \left(\beta_{0j} - \beta_{0\ell}\right) + \left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_\ell\right)^T \mathbf{x} = 0$$

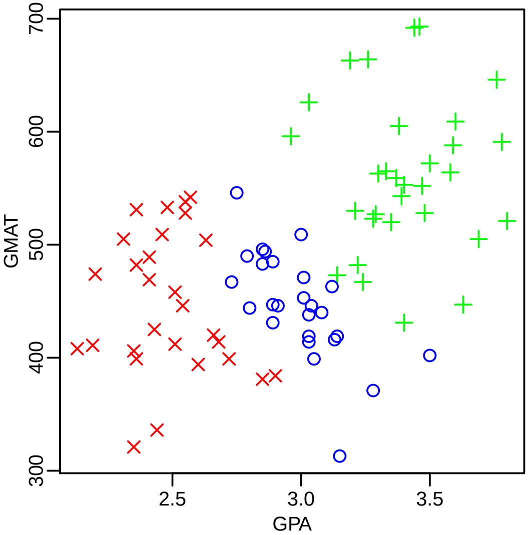
- How do we estimate the parameters $\left(\beta_{0j}, \boldsymbol{\beta}_j\right)$ for $j = 1, \ldots, K - 1$?

Logistic Classification

Maximum likelihood with a multinomial model

$$
\begin{aligned}
& L\left(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n, \beta_{0,1}, \ldots, \beta_{0,K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}\right) \\
= & \prod_{i=1}^{n} \Pr\left(Y_i = y_i \,|\, \mathbf{x}_i \,;\, \beta_{0,1}, \ldots, \beta_{0,K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}\right)
\end{aligned}
$$

Admissions Data



Admissions Data: Logistic Regression in R

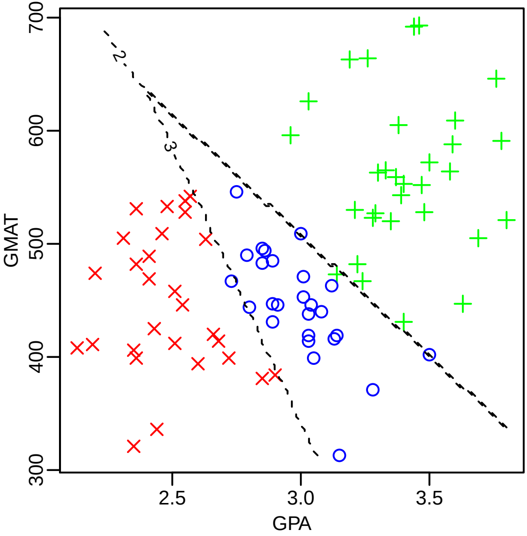
```
> # Logistic regression using multinom in nnet
> library(nnet)
> admiss.logistic <- multinom(y ~ GPA + GMAT, data = adm
   maxit = 10000)
# weights:  12 (6 variable)
initial  value 93.382045
iter  10 value 15.844299
iter  20 value 7.434088
iter  30 value 7.245347
...
iter 620 value 5.395291
iter 630 value 5.393841
final  value 5.392175
converged
```

Admissions Data: Logistic Regression in R

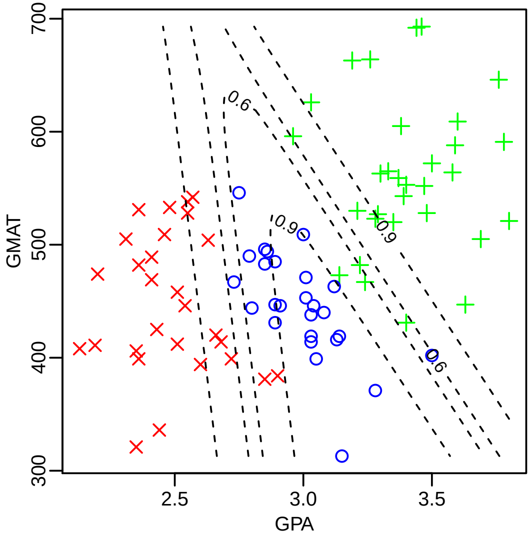
```
> print(summary(admiss.logistic))
Call:
multinom(formula = y ~ GPA + GMAT, data = admiss, maxit = 10000)

Coefficients:
  (Intercept)        GPA       GMAT
2    485.9823 -117.37076 -0.3227173
3    167.3553  -31.06165 -0.1458875

Std. Errors:
  (Intercept)      GPA      GMAT
2   0.7093202 2.648911 0.0182661
3   0.2417068 1.733160 0.0123214

Residual Deviance: 10.78435
AIC: 22.78435
```

Clicker question 2.

Admissions Data: Logistic Classification Boundaries



Admissions Data: LDA Maximum Probability



Logistic Classification Versus LDA

- Both use MLE estimates for the parameters
- Both estimate $\Pr\left(Y = c_j \,|\, \mathbf{x}\right)$
- However, the results might be different
- Which class decision boundaries do you prefer for the Admissions Data? **Clicker question 3.**

Nearest Neighbours

- If we knew the class probabilities $\Pr(Y = y \mid \mathbf{x})$ the optimal classification rule is

$$\hat{y}(\mathbf{x}) = \arg\max_h \Pr(Y = h \mid \mathbf{x})$$

- So we need to estimate $\Pr(Y = y \mid \mathbf{x})$. That's what we have been doing. Another way is to estimate it locally for each $\mathbf{x}$

- Nearest neighbours is a natural way to do so. For each $\mathbf{x}$ let

$$\widehat{\Pr}(Y = y \mid \mathbf{x}) = \text{proportion of points of class } y \text{ “near” } \mathbf{x}$$

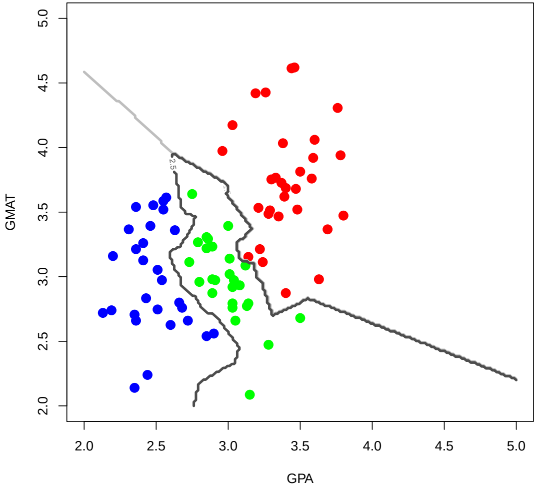
k-Nearest Neighbours

Algorithm:

1. Choose a fixed predetermined number of neighbours, k (usually $k = 1, 3, 5, \ldots$)
2. For each $\mathbf{x}$
   1. Find the k nearest neighbours of $\mathbf{x}$ in the training data (nearest defined by Euclidean distance, say)
   2. n_j is the number of neighbours of class j
   3. $\hat{y}(\mathbf{x}) =$ the class j with the largest n_j

No parametric model for $\Pr(Y \mid \mathbf{x})$! Very flexible!

Admissions Data: 1-NN



Admissions Data: 1-NN