Generative models for classification

Week 08

Bayes theorem for classification with a generative model

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X = x \mid Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

Discriminant Analysis

Here the approach is to model the distribution of X in each of the classes separately, and then use *Bayes theorem* to flip things around and obtain $\Pr(Y \mid X)$.

When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.

However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions.

Bayes theorem for classification with a generative model

$$\Pr(Y = k | X = x) = \frac{\Pr(X = x | Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

$$\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\Sigma \pi_l f_l(x)}$$

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

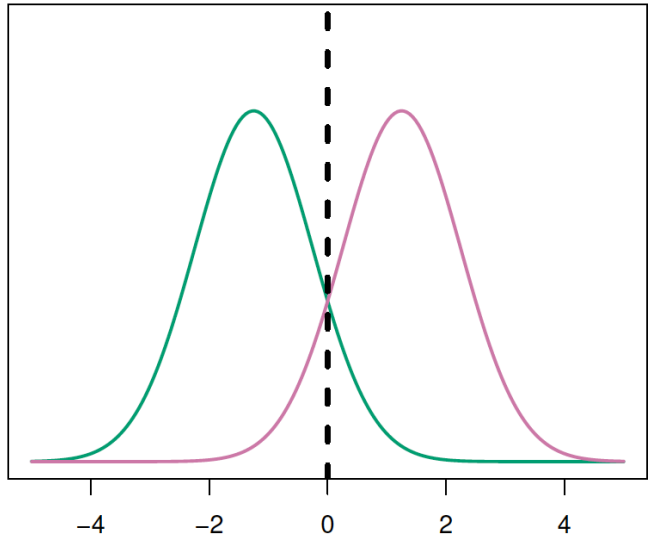
$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\Sigma \pi_l f_l(x)}, \quad \text{where}$$

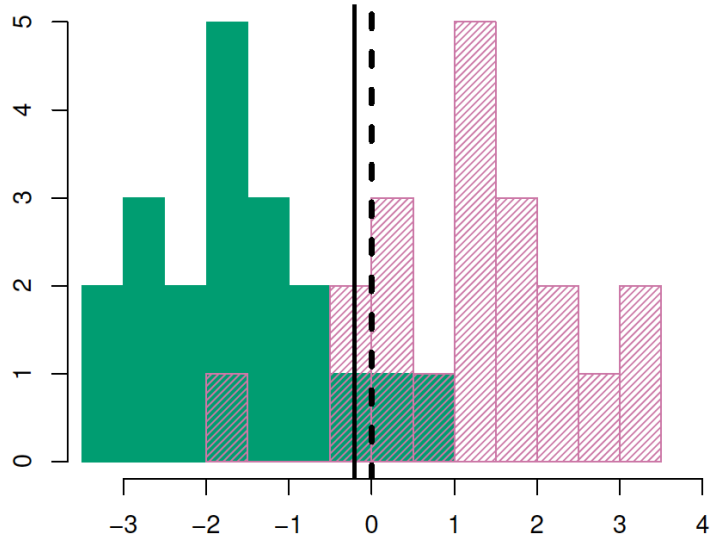
- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Generative models 101

Back to the Week 3: Generative models, likelihood and Bayes Theorem

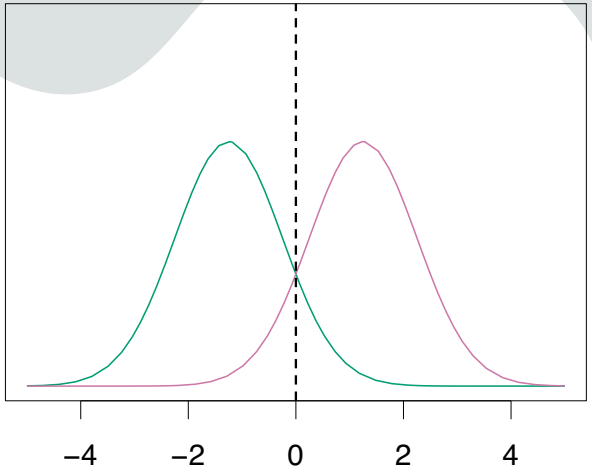
Simplest generative model : 2 groups Gaussian setting (aka LDA p=1)



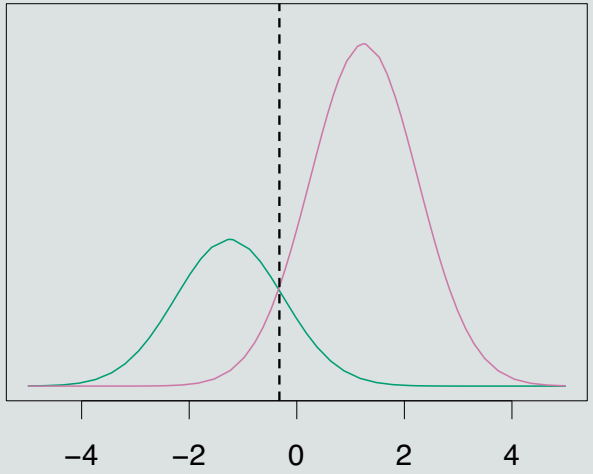


Classify to the highest density *prior

π_1=.5, π_2=.5



π_1=.3, π_2=.7



We classify a new point x according to which density* prior is highest $\pi_k f_k(x)$

From $\delta_k(x)$ to $f_k(x)$: the LDA p=1 case

The Gaussian density has the form

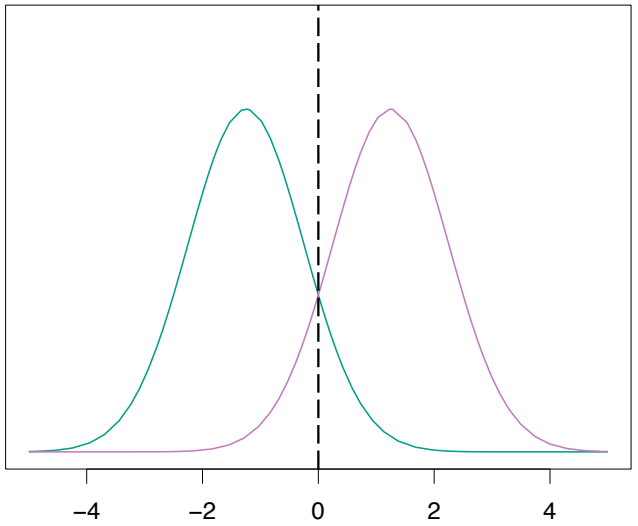
$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

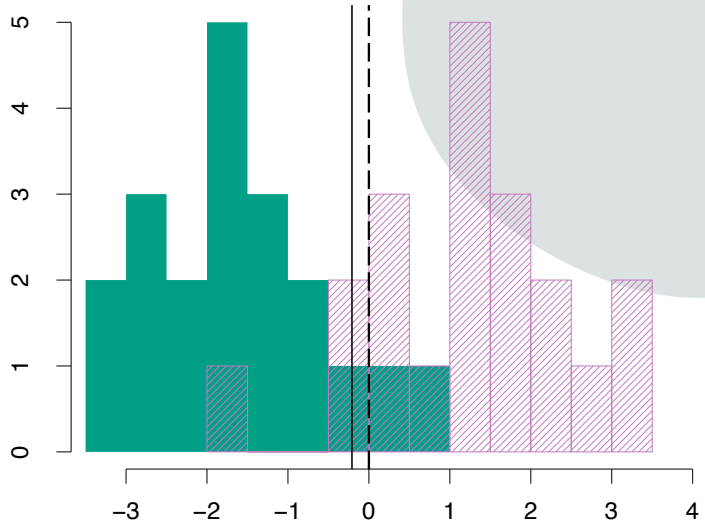
Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

THEORY



(simulated) DATA



$\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.

Discriminant functions aka δk(x)

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.

Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest *discriminant score (see Friday conceptual exercise)*:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

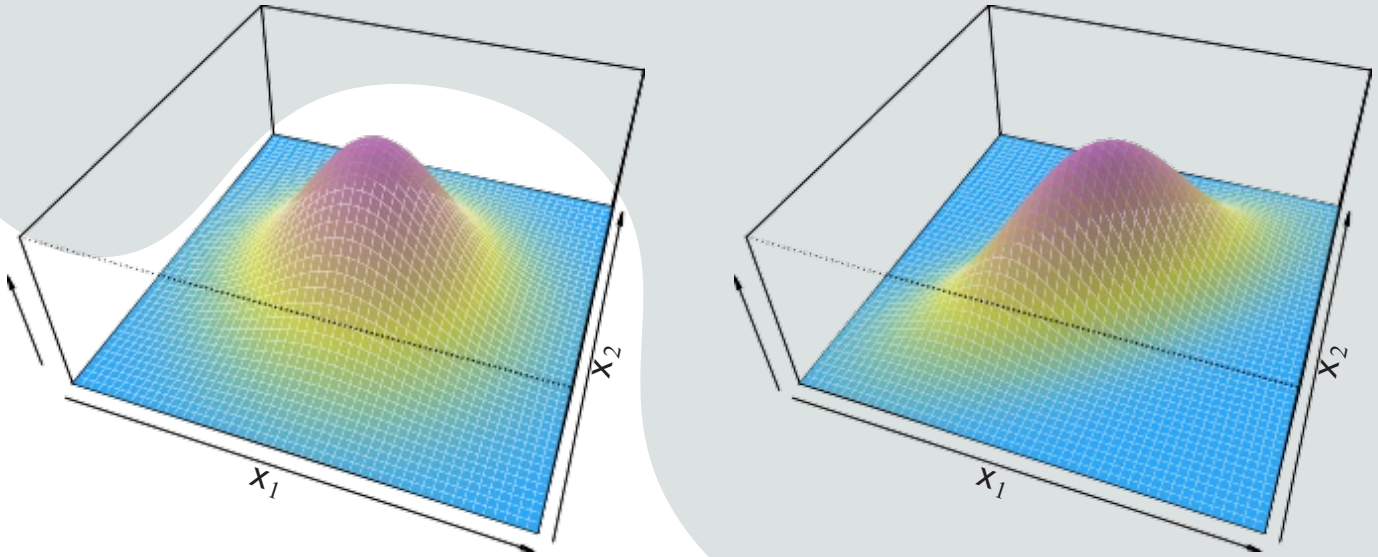
Note that $\delta_k(x)$ is a *linear* function of x.

Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable.

- If n is small and the $f_k(X)$ is approximately normal in each of the classes, LDA is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also also provides low-dimensional views of the data (**a supervised PCA)**.

LDA ... when $p > 1$

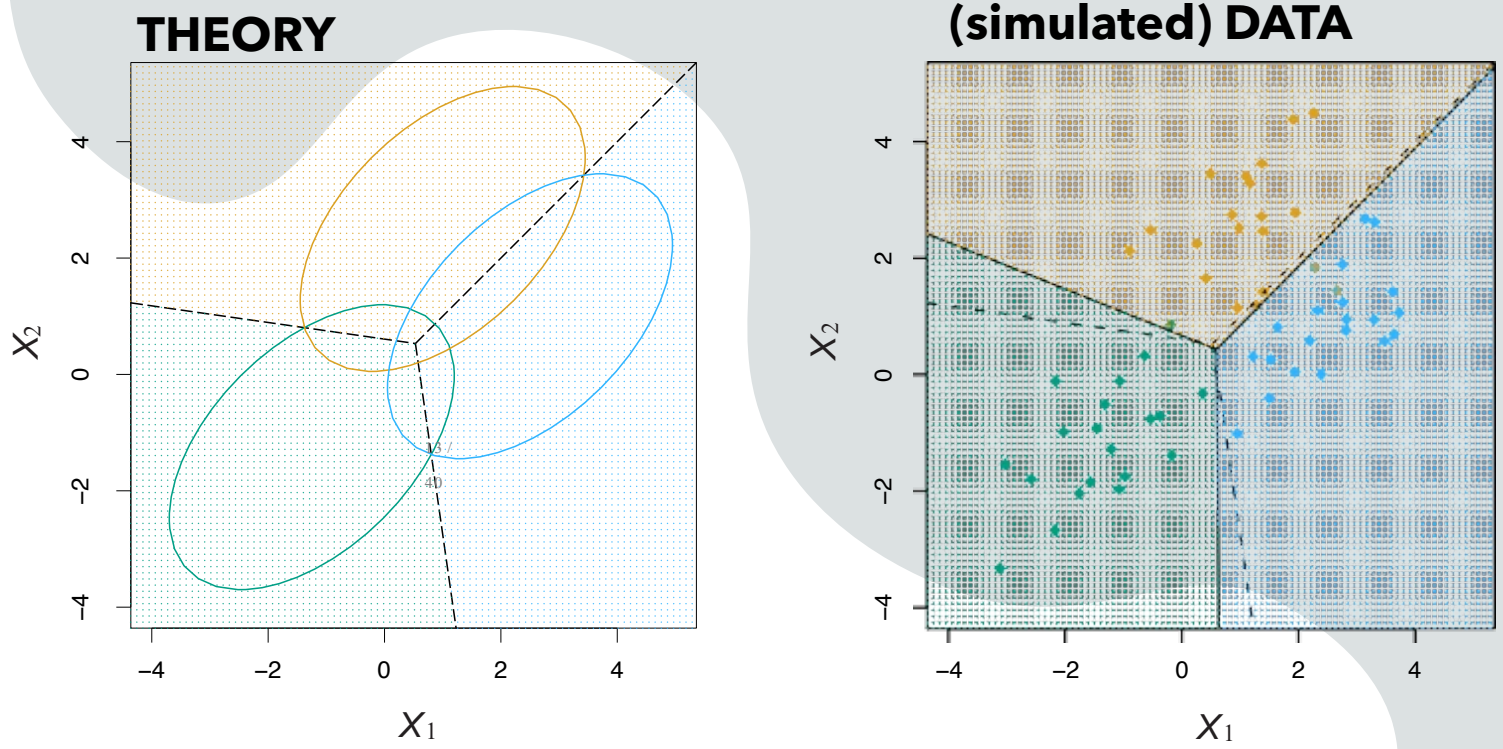


$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1} (x-\mu)}$$

$$\text{Discriminant function: } \delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$$

Despite its complex form,
$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

$p = 2$ and $K = 3$ classes



Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.
The dashed lines are known as the *Bayes decision boundaries*.
The solid lines are the boundaries given data estimation

Q; are X_1 and X_2 correlated here ?

Types of errors

False positive rate: The fraction of negative examples that are classified as positive — 0.2% in example.

False negative rate: The fraction of positive examples that are classified as negative — 75.7% in example.

We produced this table by classifying to class Yes if

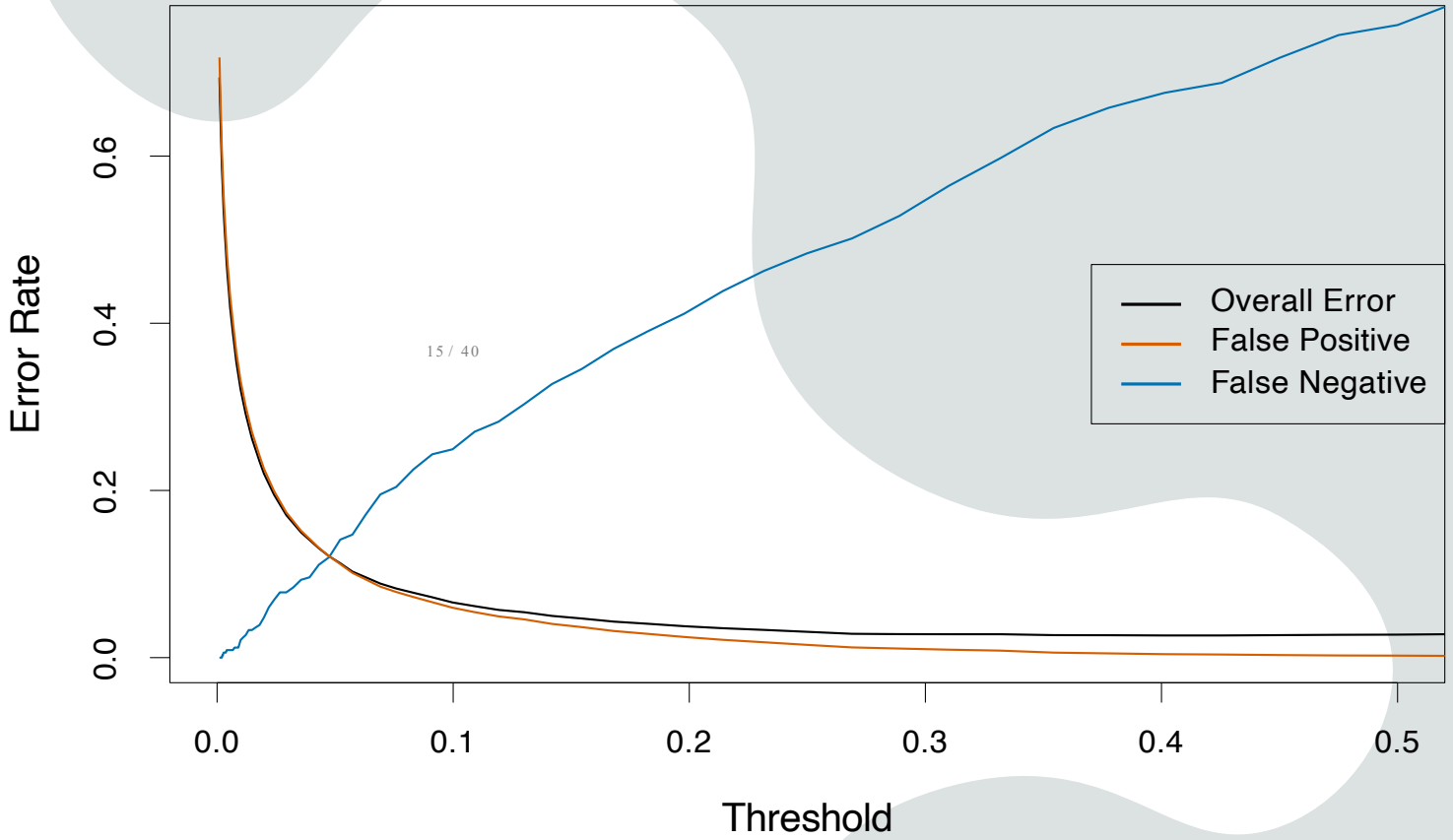
$$\Pr(\texttt{Default} = \texttt{Yes} \mid \texttt{Balance}, \texttt{Student}) \geq 0.5$$

We can change the two error rates by changing the threshold from 0.5 to some other value in [0, 1]:

$$\Pr(\texttt{Default} = \texttt{Yes} \mid \texttt{Balance}, \texttt{Student}) \geq \textit{threshold},$$

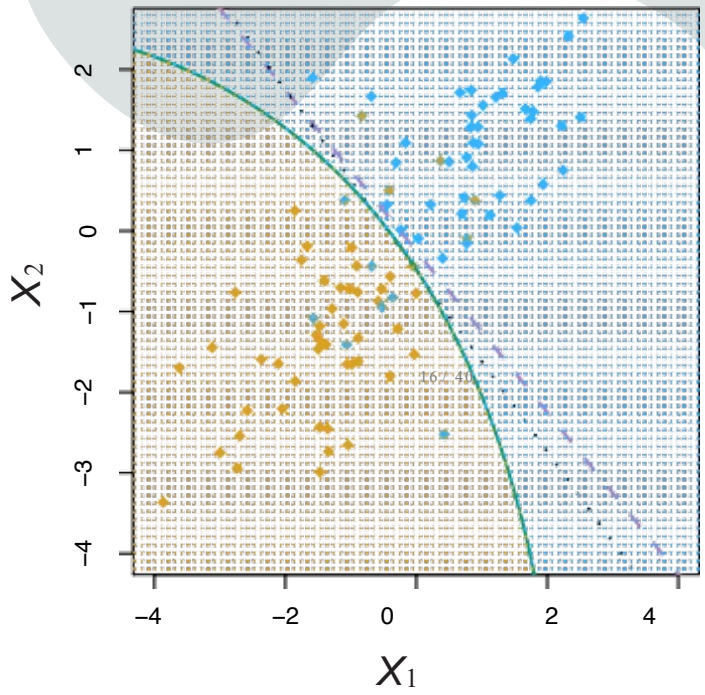
and vary *threshold*.

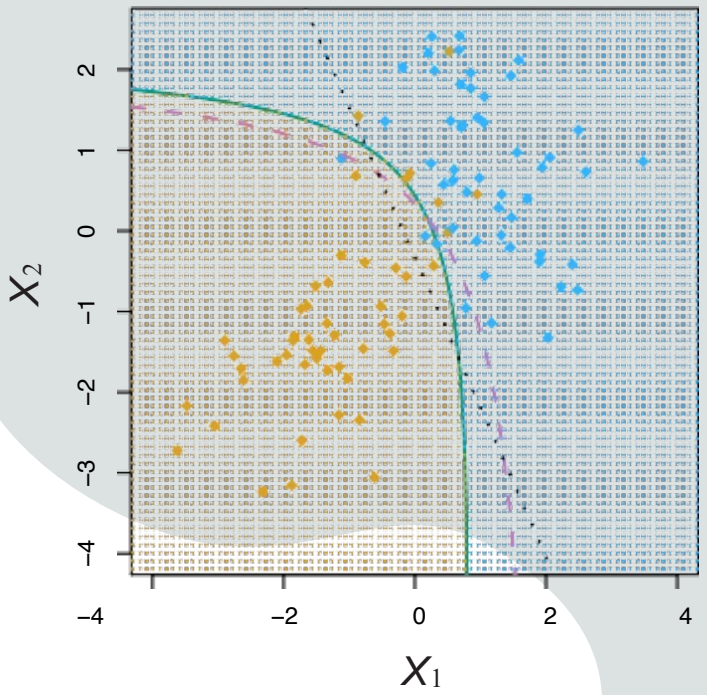
Varying the *threshold*



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

Discriminant Analysis





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

Because the $\mathbf{\Sigma}_k$ are different, the quadratic terms matter.

From Lin Discriminant Analysis, to … QDA, Naive Bayes etc.

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

By **altering the forms for $f_k(x)$**, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \Pi f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

Comparison of methods

Multivariate Normal uncorrelated

Multivariate Normal Neg. correlated

Multivariate t dis. uncorrelated

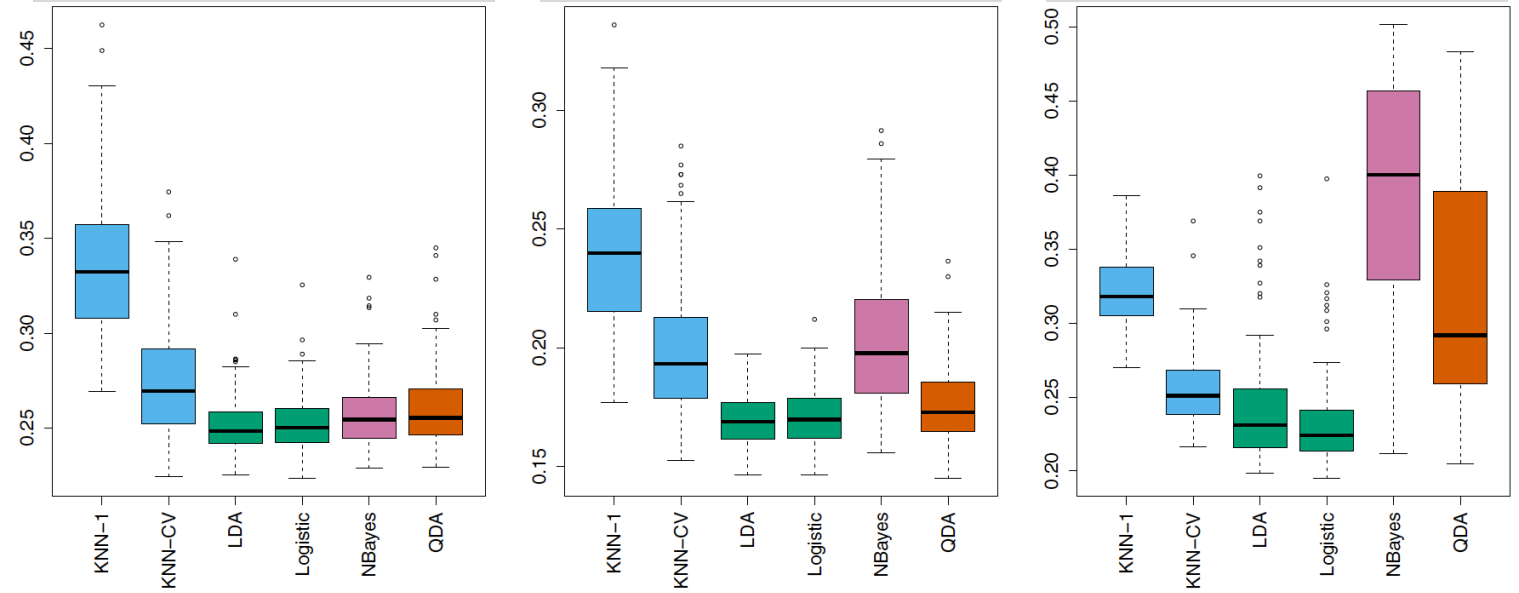
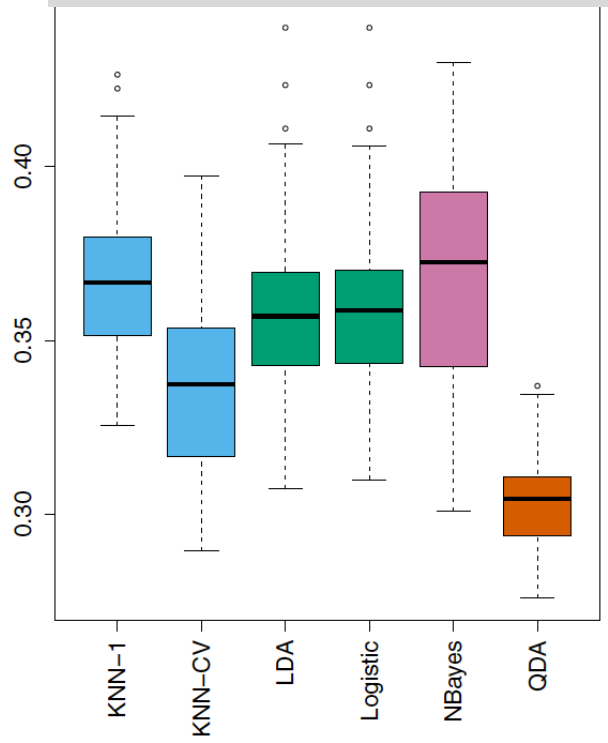
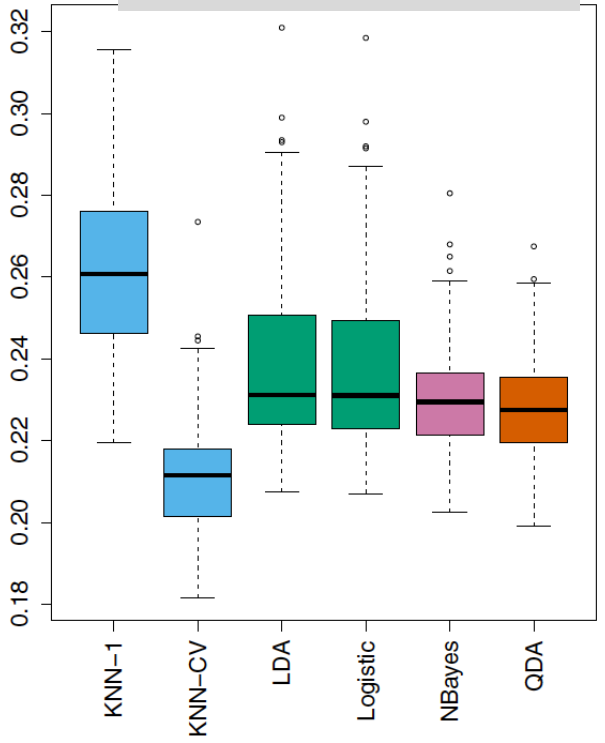


FIGURE 4.11. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

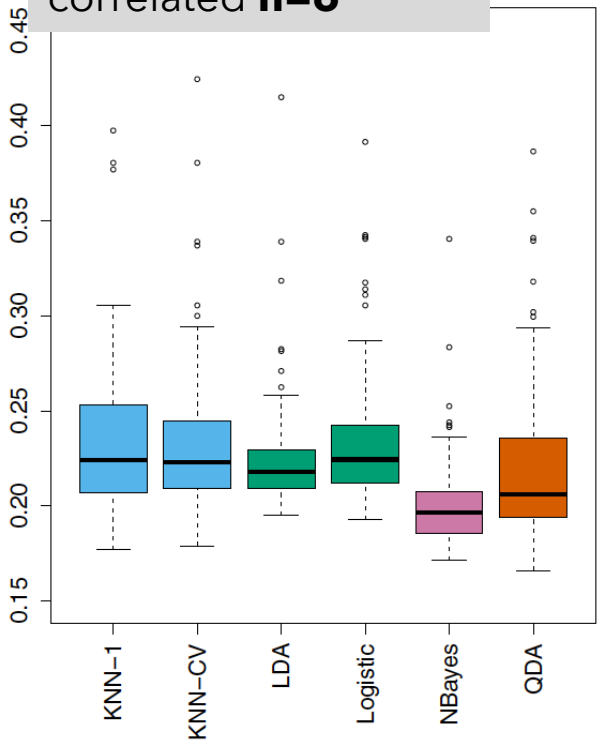
Multivariate Normal
Group speciific correlated



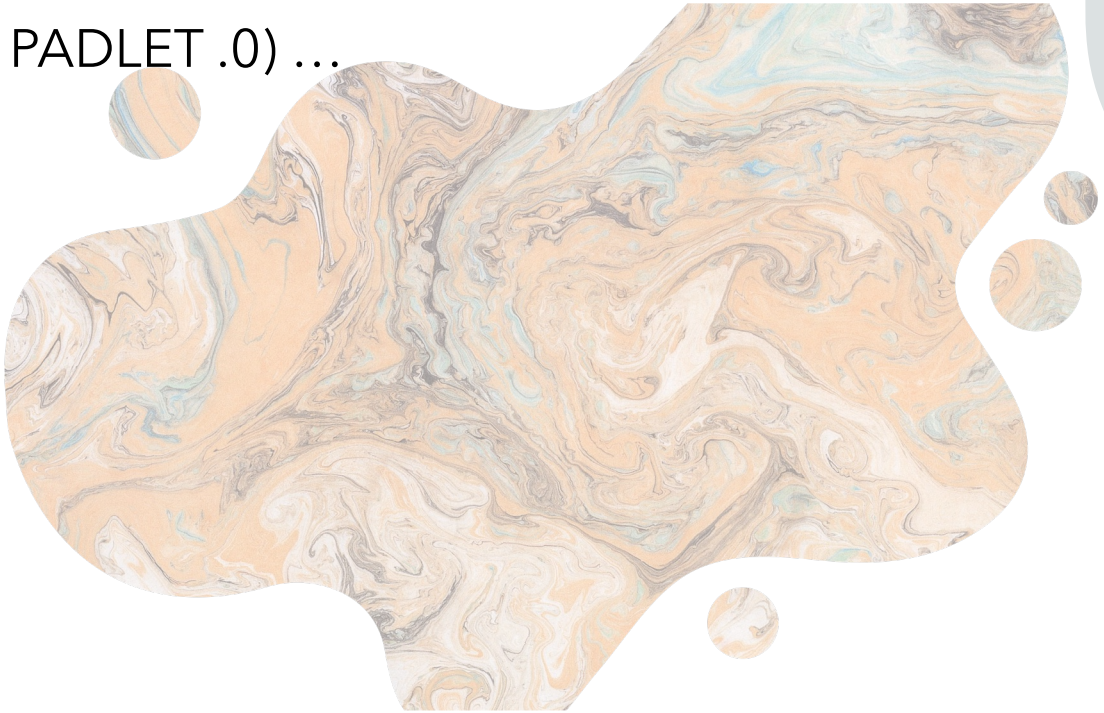
Multivariate Normal
But highly non linear



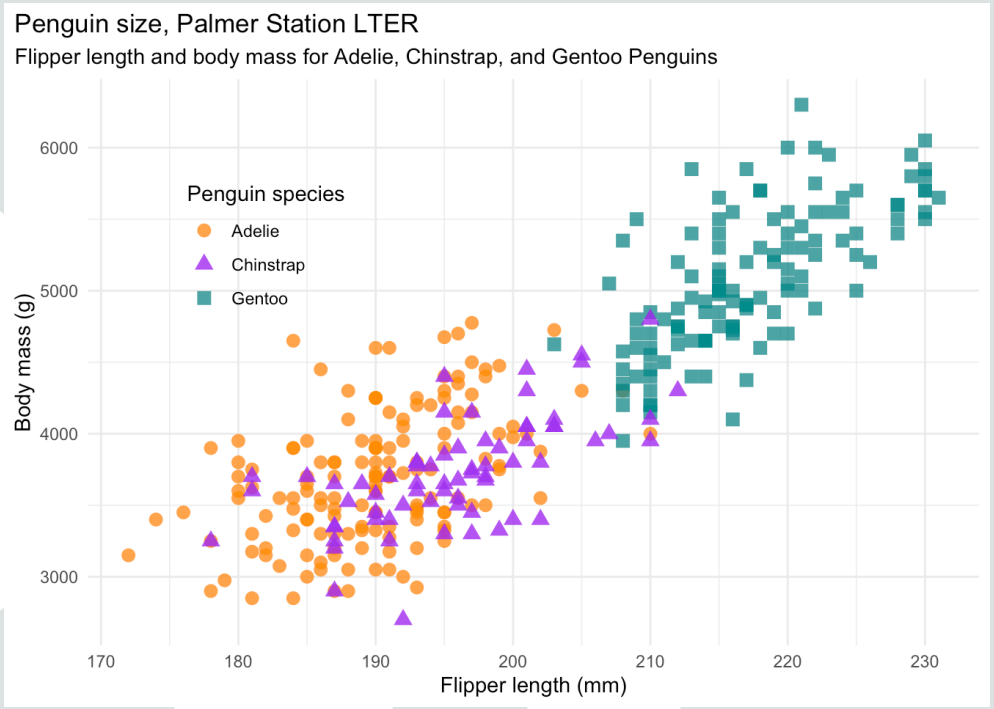
Multivariate Normal
Group speciific
correlated **n=6**



Penguin data



PADLET .0) …



The last word ?

None of these methods uniformly dominates the others: in any setting, the choice of method will depend on the true distribution of the predictors in each of the K classes, as well as other considerations, such as the values of n and p. The latter ties into the bias-variance trade-off.