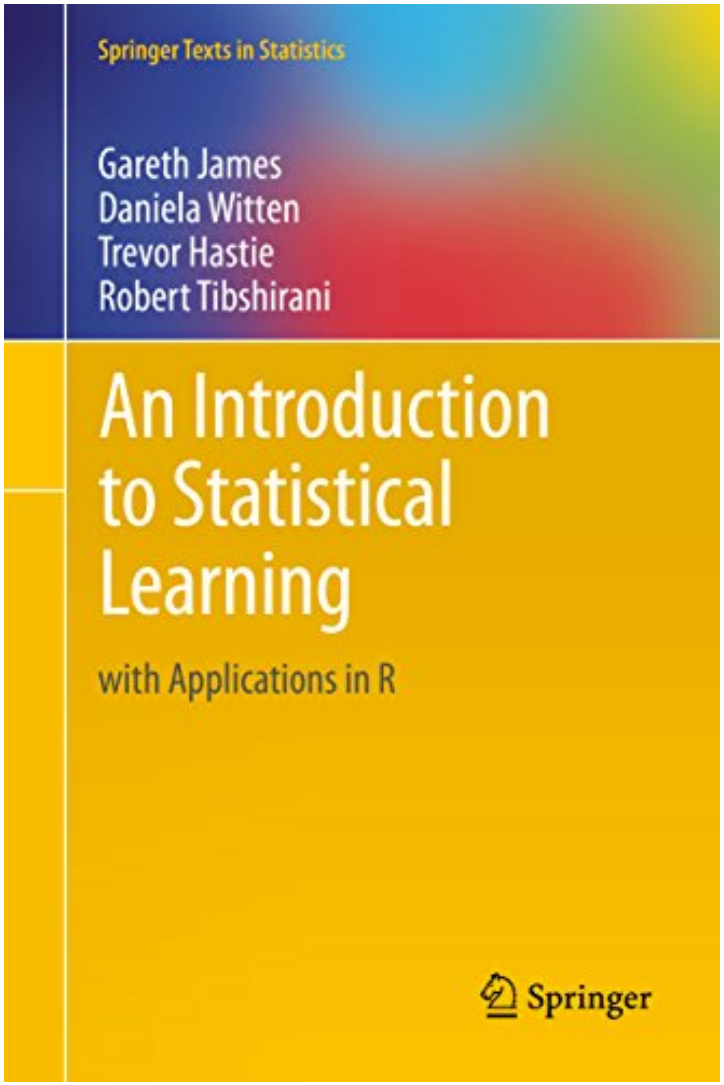
Classification

chapter 4 of *An Introduction to Statistical Learning* (ISL)

Henrique Laureano
http://leg.ufpr.br/~henrique

What we read (long description)



Now in a shorter way

What we read (short description)

At chapter 4 are discussed three of the most widely-used classifiers.

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

What we didn't read

More computer-intensive methods are discussed in later chapters, such as

- Generalized Additive Models (GAM)
- Trees
- Random Forests
- Boosting
- Support Vector Machines (SVM)

On the Agenda

We could consider encoding the response, Y, as a quantitative variable, e.g.,

Predict the medical condition of a patient on the basis of her symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

We could consider encoding the response, Y, as a quantitative variable, e.g.,

Predict the medical condition of a patient on the basis of her symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

Unfortunately, this coding implies an ordering on the outcomes.

Each possible coding would produce a fundamentally different linear model that would ultimately lead to different sets of predictions.

That leads us to other questions,

- What if the response variable values did take on a natural ordering, `such as mild`, `moderate`, and `severe`?
- For a `binary` (two level) qualitative response, the situation is better.
  - However, if we use linear regression, some of our estimates might be outside the [0, 1] interval.
  - However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

That leads us to other questions,

- What if the response variable values did take on a natural ordering, `such as mild`, `moderate`, and `severe`?
- For a `binary` (two level) qualitative response, the situation is better.
  - However, if we use linear regression, some of our estimates might be outside the [0, 1] interval.
  - However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.

For these reasons, it is preferable to use a classification method that is truly suited for qualitative response values, such as the ones presented next.

Curiously,

it turns out that the classifications that we get if we use linear regression to predict a binary response will be the same as for the linear discriminant analysis (LDA) procedure we discuss later.

On the Agenda

A classic 'book example dataset relationship'

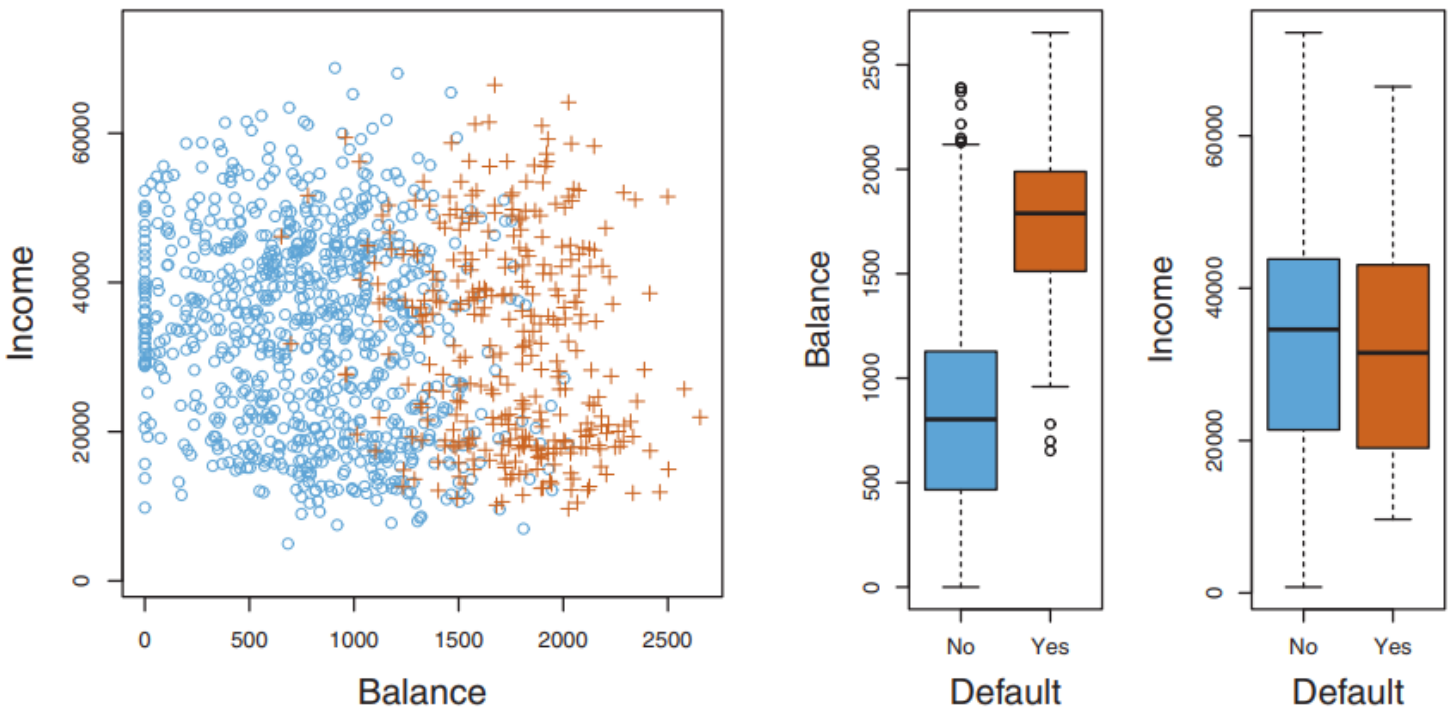


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

. . . a very pronounced relationship between `balance` and `default`.

On the Agenda

To start, a comparison with Linear Regression

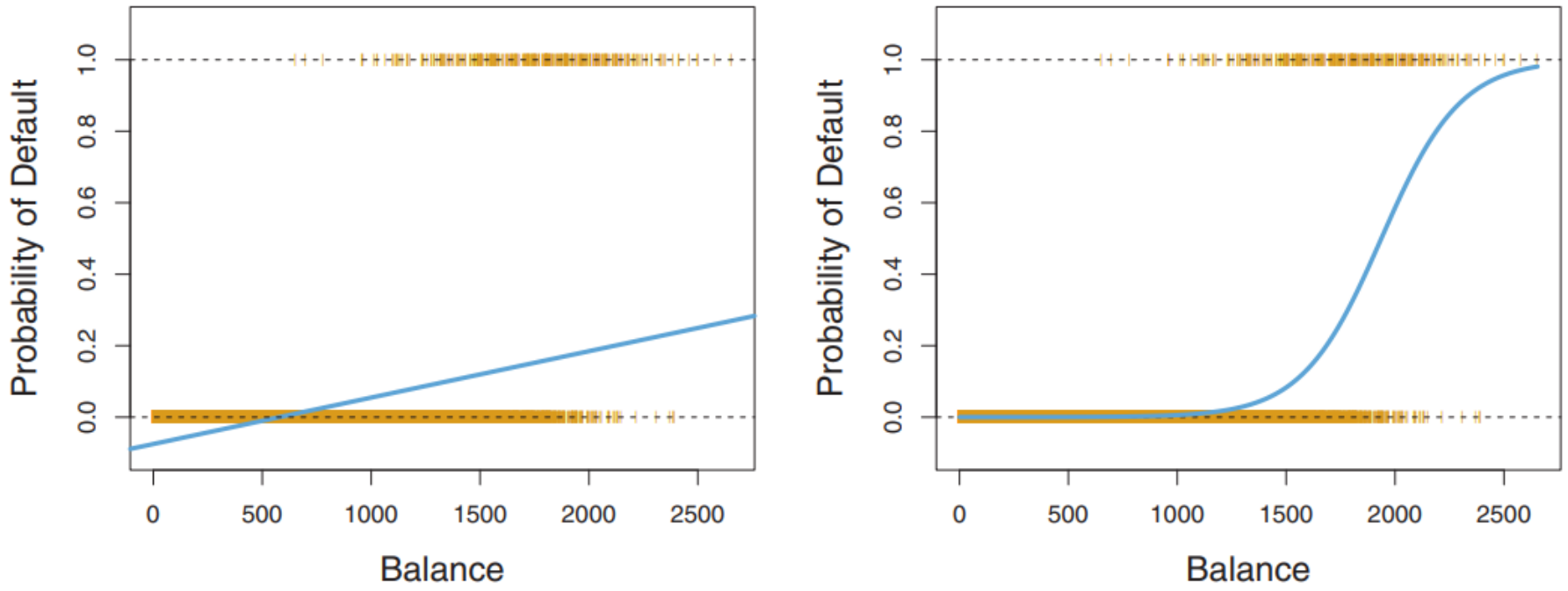


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Logistic regression in two slides

Some math, but with just one predictor

The model and its relations (*showing my LaTeX skills*)

$$p(X) = \underbrace{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}_{\substack{\textit{logistic}\\ \textit{function}\\ \textit{(S-shaped)}}} \Rightarrow \underbrace{\frac{p(X)}{1-p(X)}}_{\textit{odds} \in (0,\infty)} = e^{\beta_0+\beta_1 X} \Rightarrow \underbrace{\log \frac{p(X)}{1-p(X)}}_{\substack{\textit{log-odds}\\ \text{or}\\ \textit{logit}}} = \beta_0+\beta_1 X$$

Some math, but with just one predictor

The model and its relations (*showing my LaTeX skills*)

$$p(X) = \underbrace{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}_{\substack{\text{logistic} \\ \text{function} \\ \text{(S-shaped)}}} \Rightarrow \underbrace{\frac{p(X)}{1-p(X)}}_{\text{odds} \in (0,\infty)} = e^{\beta_0+\beta_1 X} \Rightarrow \underbrace{\log \frac{p(X)}{1-p(X)}}_{\substack{\text{log-odds} \\ \text{or} \\ \text{logit}}} = \beta_0+\beta_1 X$$

For example,

$$p(X) = 0.2 \Rightarrow \frac{0.2}{1-0.2} = \frac{1}{4} \quad \textit{and} \quad p(X) = 0.9 \Rightarrow \frac{0.9}{1-0.9} = 9.$$

Maximum likelihood

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize a math equation called a

likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Maximum likelihood

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize a math equation called a

likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

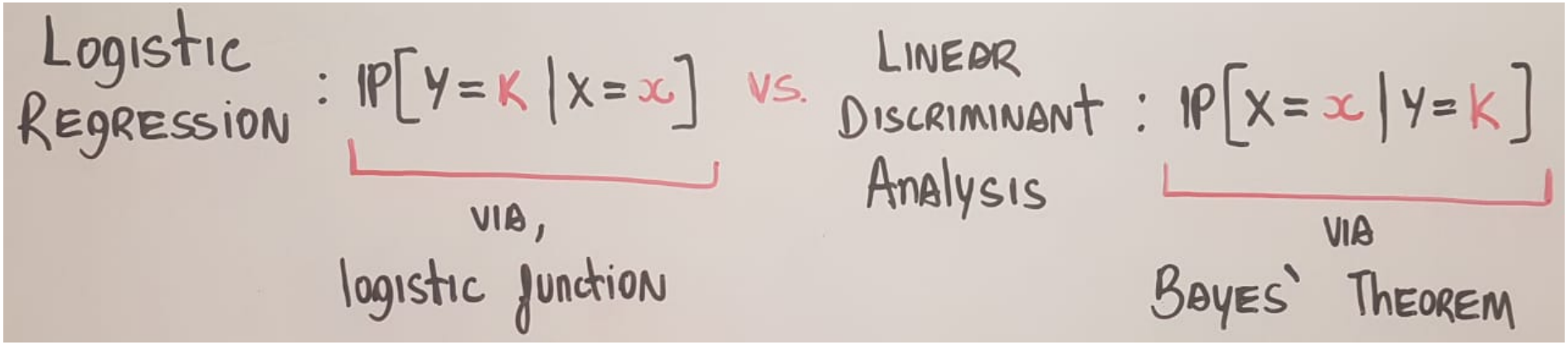
The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown, and must be estimated. The general method of maximum likelihood is preferred, since it has better statistical properties.

Maximum likelihood is a very general approach that is used to fit many of the non-linear models examined throughout the book. In the linear regression setting, the least squares approach is in fact a special case of maximum likelihood.

On the Agenda

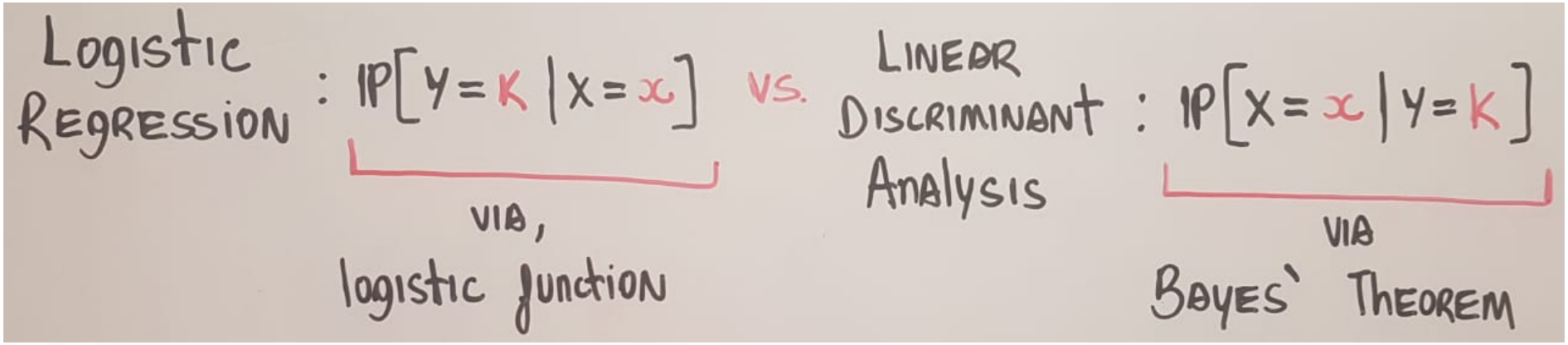
Different ideas, sometimes the same results

Different ideas,



With LDA we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for $\mathbb{P}[Y = k|X = x]$.

Different ideas,



With LDA we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for $\mathbb{P}[Y = k|X = x]$.

Sometimes the same results

When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression.

But, ok... why not continue with logistic regression?

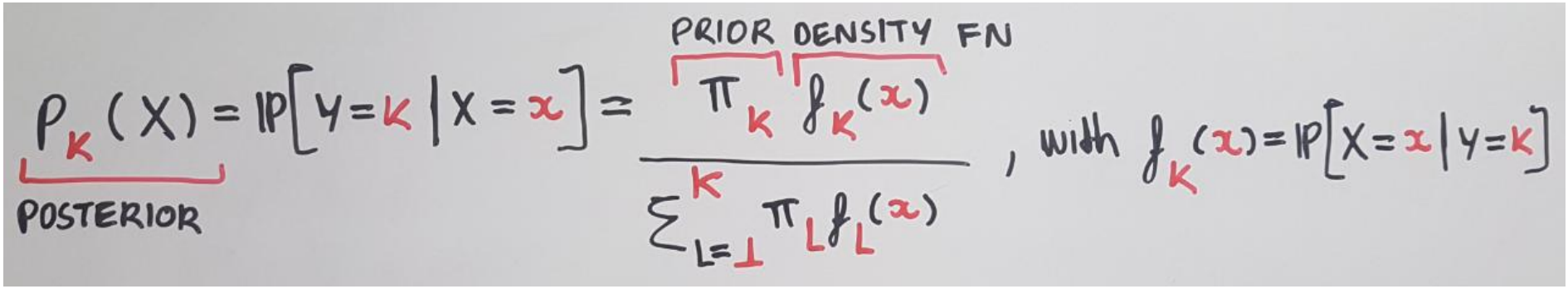
But, ok... why not continue with logistic regression?

Simple, LDA is popular when we have more than two response classes.

Now, a reason more serious: stability

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.

Model framework

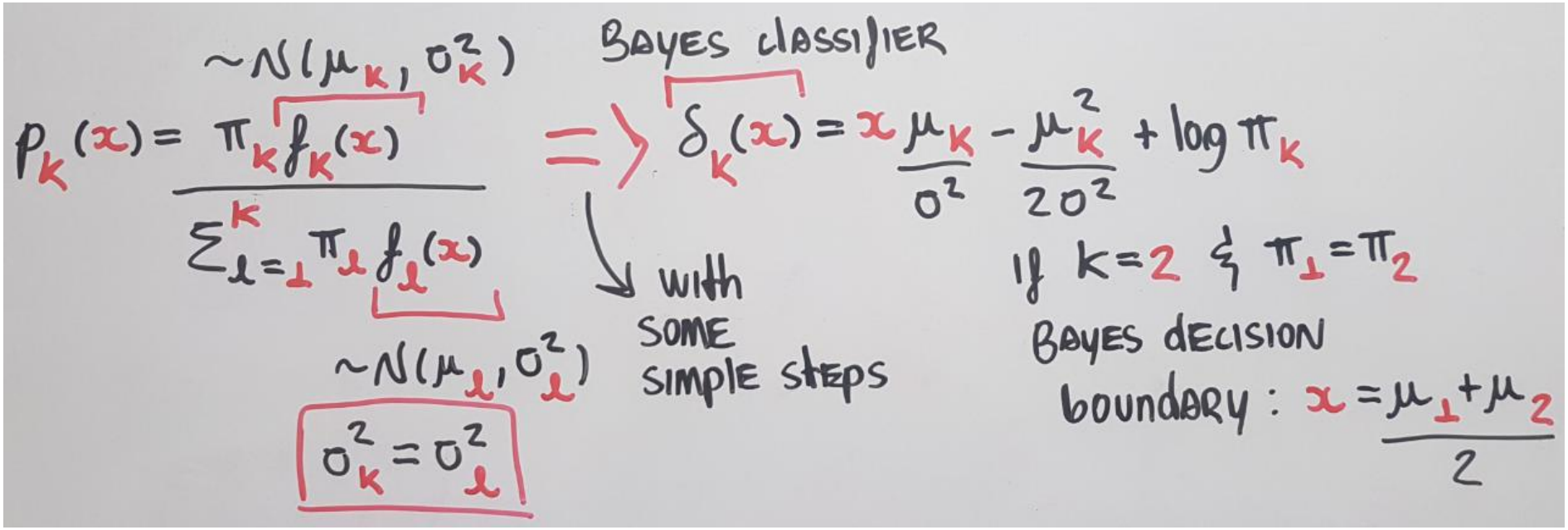


- π_k is the overall or prior prob. that a chosen obs. comes from k.
- In general, estimating π_k is easy if we have a sample of Ys: we simply compute the fraction of observations that belong to the kth class. However, estimating $f_k(x)$ tends to be more challenging, unless we assume some simple forms for these densities.

Remember from Chap. 2 that the Bayes classifier has the lowest possible error rate out of all classifiers.

Dealing with just one predictor

Assumptions: $f_k(x)$ is normal with equal variance for the kth classes.



Putting a hat (simple average and a weighted average of the sample variances for each class) in everything, the LDA approx. this Bayes classifier.

Ok, nice! But... why the name linear discriminant analysis?

Ok, nice! But... why the name linear discriminant analysis?

The word linear stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

That is, the LDA decision rule depends on x only through a linear combination of its elements.

Ok, nice! But... why the name linear discriminant analysis?

The word linear stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

That is, the LDA decision rule depends on x only through a linear combination of its elements.

LDA is trying to approximate the Bayes classifier, which has the lowest total error rate out of all classifiers (if the Gaussian model is correct).

Getting bigger

More than one predictor $\Rightarrow$ Multivariate normal distribution, with a class-specific mean vector and a common covariance matrix

Getting bigger

More than one predictor $\Rightarrow$ Multivariate normal distribution, with a class-specific mean vector and a common covariance matrix

---

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu_k)^\top \mathbf{\Sigma}^{-1}(x-\mu_k)\right\}$$

$$\Rightarrow$$

$$\hat{\delta}_k(x) = x^\top \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^\top \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k + \log \hat{\pi}_k$$

An example

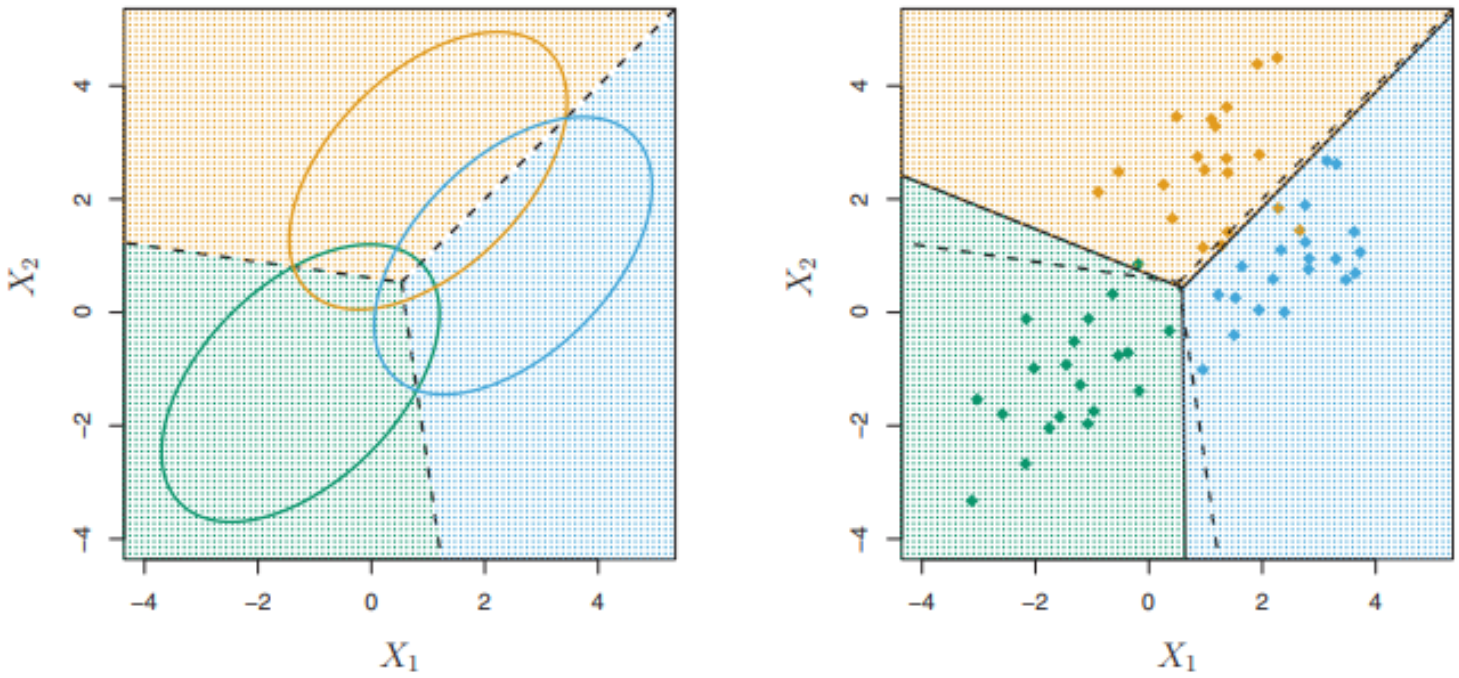


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

Ok, and about what else do we need to talk? (1/2)

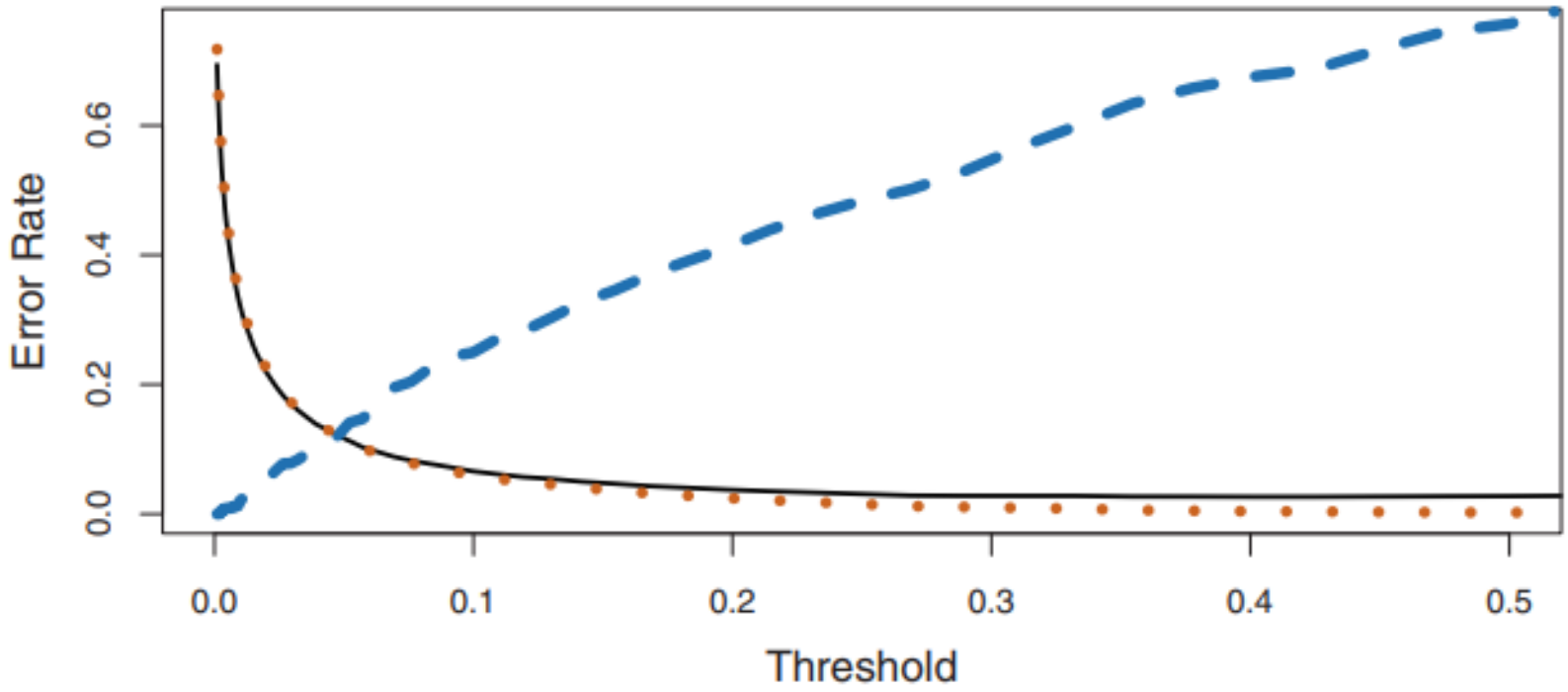


FIGURE 4.7. *For the* `Default` *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.*

Ok, and about what else do we need to talk? (2/2)

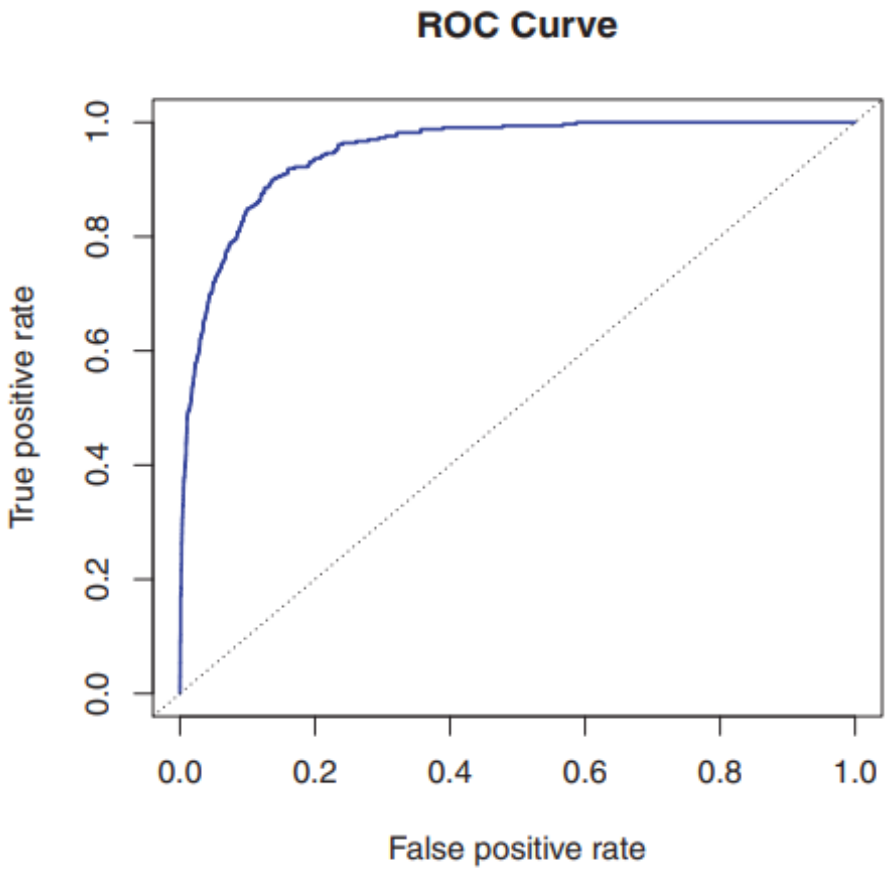


FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold*

On the Agenda

Unlike LDA, QDA assumes that each class has its own covariance matrix.

Under this assumption, the approximation of the Bayes classifier becomes

$$\text{QLA}: \quad \hat{\delta}_k(x) = -\frac{1}{2}(x-\hat{\mu}_k)^\top \hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) - \frac{1}{2}\log|\hat{\Sigma}_k| + \log\hat{\pi}_k.$$

x appears as a quadratic function, this is where QDA gets its name.

Unlike LDA, QDA assumes that each class has its own covariance matrix.

Under this assumption, the approximation of the Bayes classifier becomes

$$\text{QLA}: \quad \hat{\delta}_k(x) = -\frac{1}{2}(x-\hat{\mu}_k)^\top \hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) - \frac{1}{2}\log|\hat{\Sigma}_k| + \log\hat{\pi}_k.$$

x appears as a quadratic function, this is where QDA gets its name.

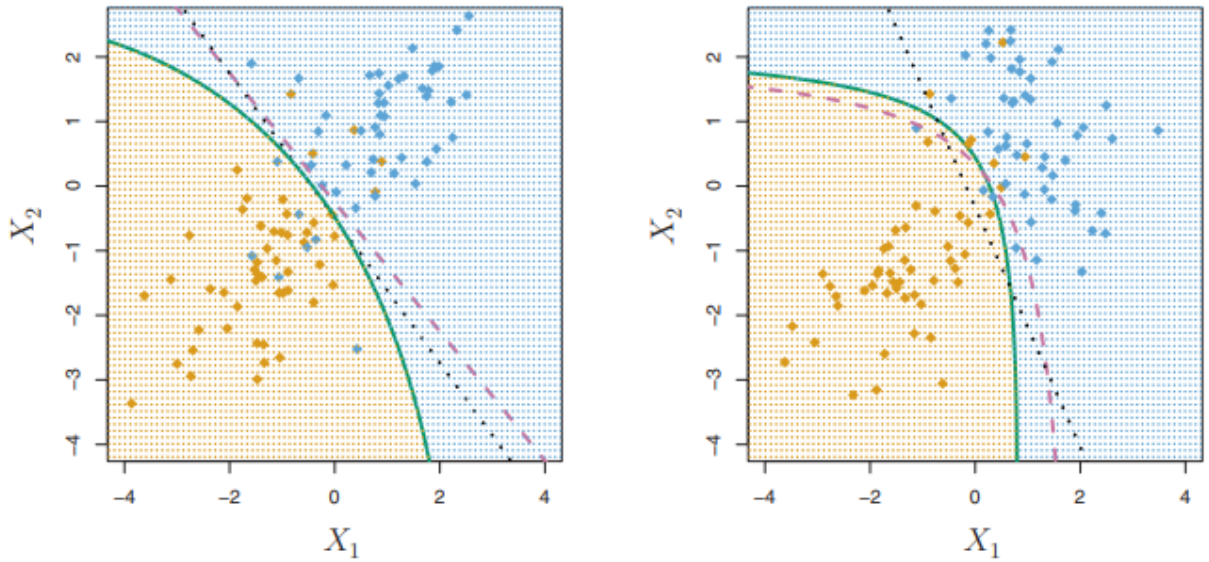
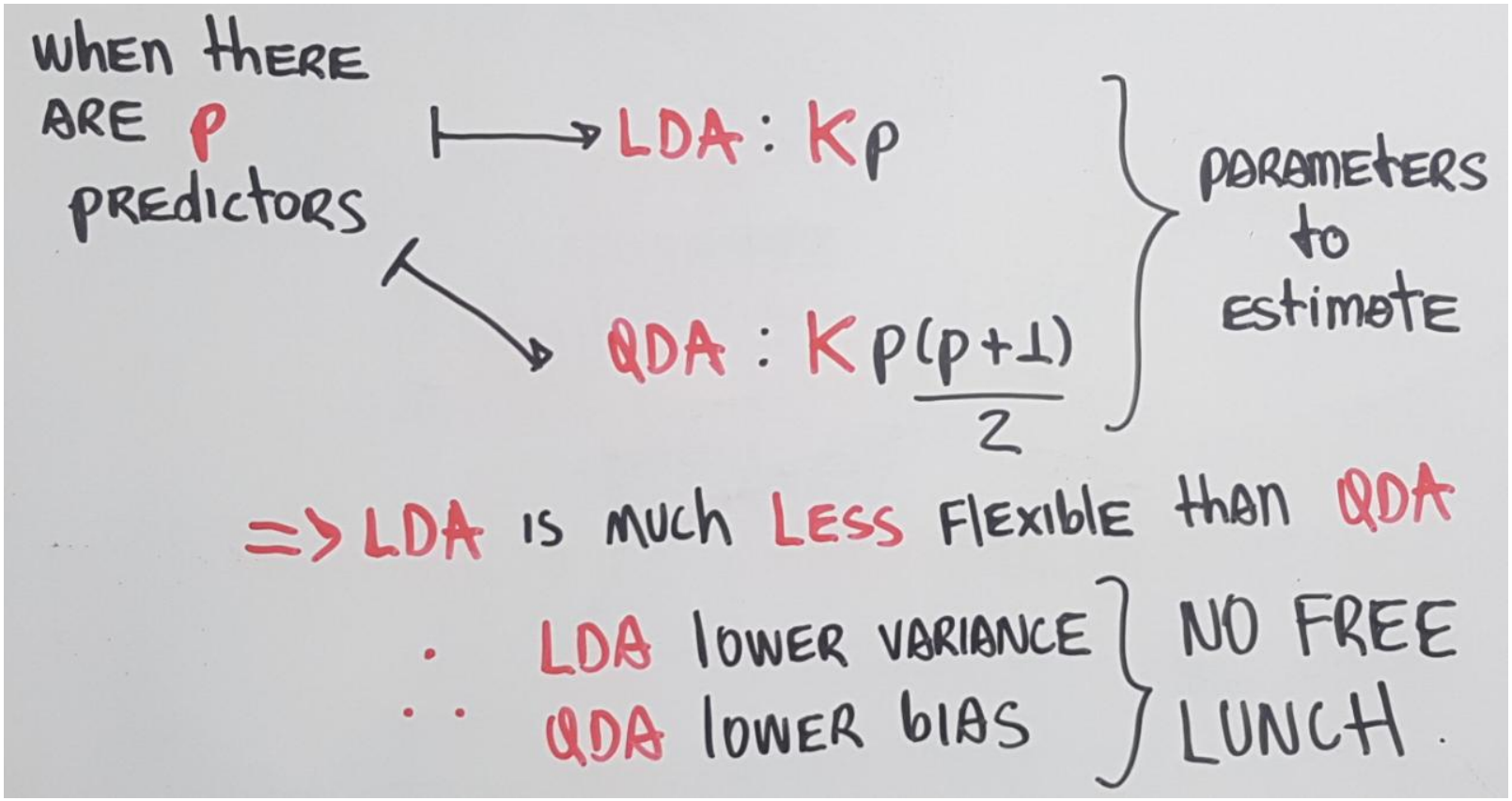


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Ok, but... in practice, what's the difference?

Why does it matter whether or not we assume that the K classes share a common covariance matrix?

The answer lies in the bias-variance trade-off.



Concluding...

LDA tends to be a better bet than QDA if there are relatively few observations and so reducing variance is crucial.

In contrast, QDA is recommended if the data set is very large, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the K classes is clearly untenable.

On the Agenda

- The logistic regression and LDA methods are closely connected, since both produce linear decision boundaries.

To make a nicer comparison, we may mencion the KNN.

- KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. Nevertheless, KNN does not tell us which predictors are important.

- When the true decision boundaries are linear, the LDA and logistic regression approaches will tend to perform well. When the boundaries are moderately non-linear, QDA may give better results. Finally, for much more complicated decision boundaries, a non-parametric approach such as KNN can be superior. But the level of smoothness for a non-parametric approach must be chosen carefully.

and...

laureano@ufpr.br