Discriminant analysis

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on **ISLR book** and its accompanying slides

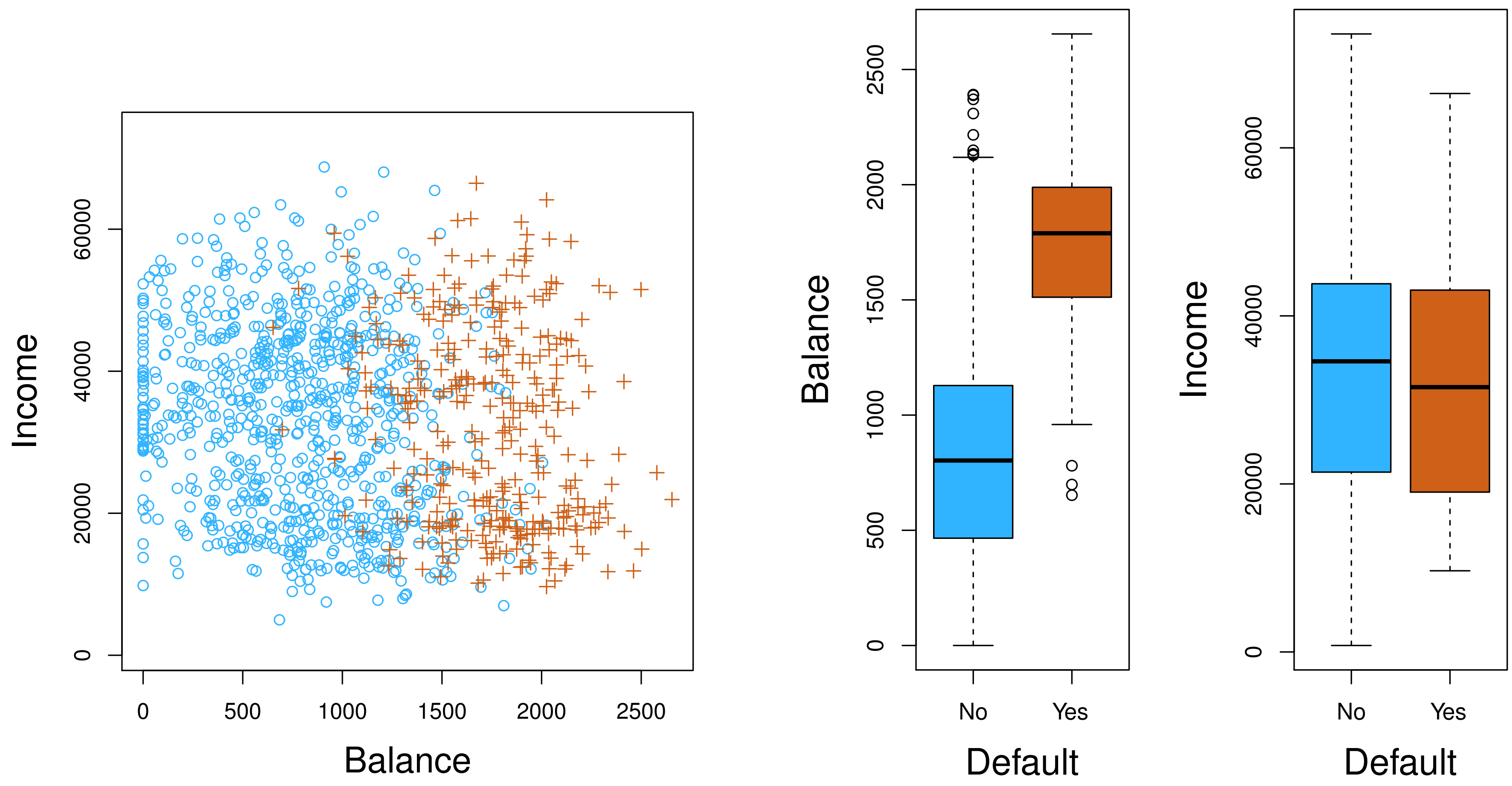
http://cw.felk.cvut.cz/wiki/courses/b4m36san/start

Introduction

- Study multivariate relationships with categorical dependent variable
  - independent variables are continuous,
  - but can be categorical too,
  - **nominal** dependent variables take values in an unordered set $\mathcal{C}$
    * eye color$\in${brown, blue,green}, email$\in${spam, ham},
- the main goals are to
  - **classify** into the target categories
    * given a feature vector $\mathbf{X} \in \mathcal{X}$ and a nominal response Y taking values in $\mathcal{C}$, the goal is to build a function $f : \mathcal{X} \to \mathcal{C}$,
    * often, the mapping is probabilistic $f_p : \mathcal{X} \times \mathcal{C} \to [0, 1]$,
  - **understand** the role of the individual independent variables
    * assess the strength of their relationships with the target variable.

Example: Credit Card Default

- Simulated dataset, an individual may default on his credit card payment.



Can we use linear regression?

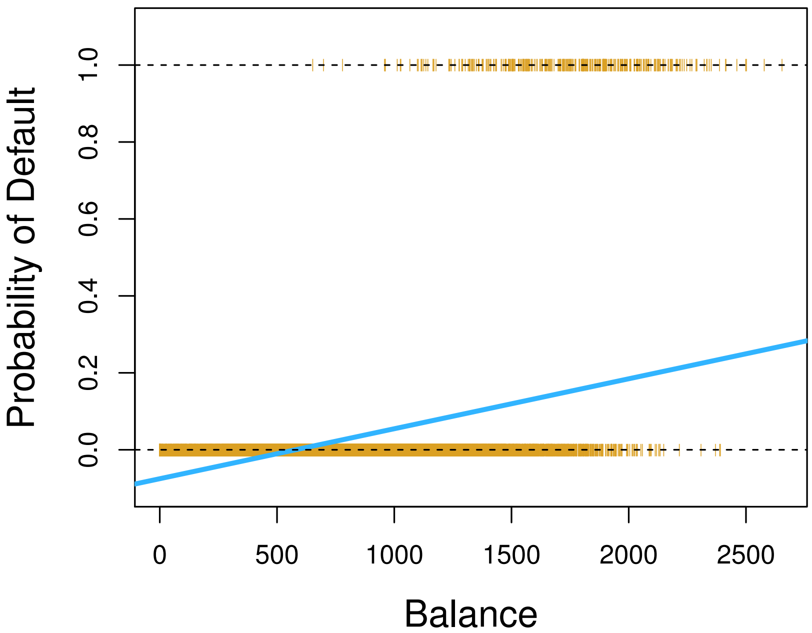
- the target variable Y expressing default can be coded

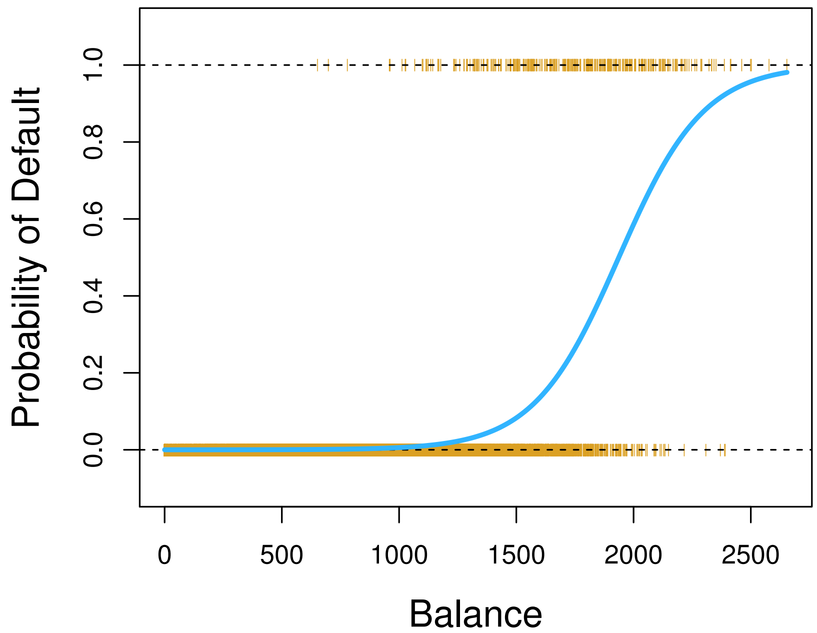
$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

- perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$
  - in this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to linear discriminant analysis (discussed later),
  - since in the population $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task,
  - however, linear regression might in general
    * produce probabilities less than zero or bigger than one,
    * be sensitive to outliers,
    * "mask out" some classes in problems with multinomial targets,
- **logistic regression** is more appropriate.

Linear versus logistic regression

- Consider a simple linear model $Y = \beta_0 + \beta_1\ Balance$ (left),
- introduce a non-linear **logit** transformation (right).





The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Logistic regression

- Let's write $p(\mathbf{X}) = Pr(Y = 1|\mathbf{X})$ for short,
- logistic regression uses the form

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\ldots\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots\beta_p X_p}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots\beta_p X_p)}}$$

- no matter what values β_i or X_i take, $p(\mathbf{X})$ will have values between 0 and 1,
- a bit of rearrangement gives

$$log\Big(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\Big) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$$

- this monotone transformation is called the **log odds** or **logit** transf. of $p(\mathbf{X})$,
- we use maximum likelihood to estimate the parameters β_i

$$\ell(\beta_0, \ldots, \beta_p) = \prod_{\forall i\, y_i=1} p(\mathbf{x_i}) \prod_{\forall i\, y_i=0} (1 - p(\mathbf{x_i}))$$

Logistic regression – motivation

- In linear regression
  - the outcome thresholds the distance to the decision boundary
  - the distance can easily be computed,
- transform this distance to probability $p(X)$ with the following requirements
  - the objects lying on the boundary have $p(X) = 0.5$,
  - distant objects have $p(X) \to 0$ (in one direction) or $p(X) \to 1$ (in the other direction),
  - the transformation is most sensitive around the decision boundary,
- transformation steps
  - start with the linear model, its limitations are known,
  - distance has no ceiling $\to$ turn probability into odds to remove the range restrictions,
  - however, we need to consider direction from the decision boundary too,
  - apply log transform to remove the floor restriction.

Logistic regression – making predictions

- In R `glm` function can be applied to learn logistic models
  - **generalized linear models** allow the linear model to be related to the response variable via a link function, and allow for responses whose error distribution is different from a normal distribution,
- fit the *default* model for *balance*

	Coefficient	Std. error	Z-statistic	p-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- fit another *default* model for *student*

	Coefficient	Std. error	Z-statistic	p-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Logistic regression – making predictions

- Coefficient interpretation in simple models?
  - simpler for a binary predictor such as *student* $(Pr(default = yes|student = yes) = p(s^+))$,
  - compare log-odds for the student and non-student groups,

$$\frac{p(s^+)}{1-p(s^+)} = e^{\hat{\beta}_0+\hat{\beta}_1} \;\&\; \frac{p(s^-)}{1-p(s^-)} = e^{\hat{\beta}_0} \rightarrow \frac{\frac{p(s^+)}{1-p(s^+)}}{\frac{p(s^-)}{1-p(s^-)}} = e^{\hat{\beta}_1}$$

  - $e^{\hat{\beta}_1} = e^{0.4049} = 1.5$ gives the **odds ratio** between the groups,
- where is the decision boundary and what is its shape?

Logistic regression – making predictions

- Coefficient interpretation in simple models?
  - simpler for a binary predictor such as *student* $(Pr(default = yes|student = yes) = p(s^+))$,
  - compare log-odds for the student and non-student groups,

$$\frac{p(s^+)}{1-p(s^+)} = e^{\hat{\beta}_0+\hat{\beta}_1} \; \& \; \frac{p(s^-)}{1-p(s^-)} = e^{\hat{\beta}_0} \rightarrow \frac{\frac{p(s^+)}{1-p(s^+)}}{\frac{p(s^-)}{1-p(s^-)}} = e^{\hat{\beta}_1}$$

  - $e^{\hat{\beta}_1} = e^{0.4049} = 1.5$ gives the **odds ratio** between the groups,
- where is the decision boundary and what is its shape?
  - more clear for a continuous predictor $(Pr(default = y|balance) = p(b))$

$$p(b) = 0.5 \rightarrow \frac{p(b)}{1-p(b)} = 1 \rightarrow \log\left(\frac{p(b)}{1-p(b)}\right) = \beta_0+\beta_1 b = 0 \rightarrow b = -\frac{\beta_0}{\beta_1} = \$1937$$

Logistic regression – making predictions

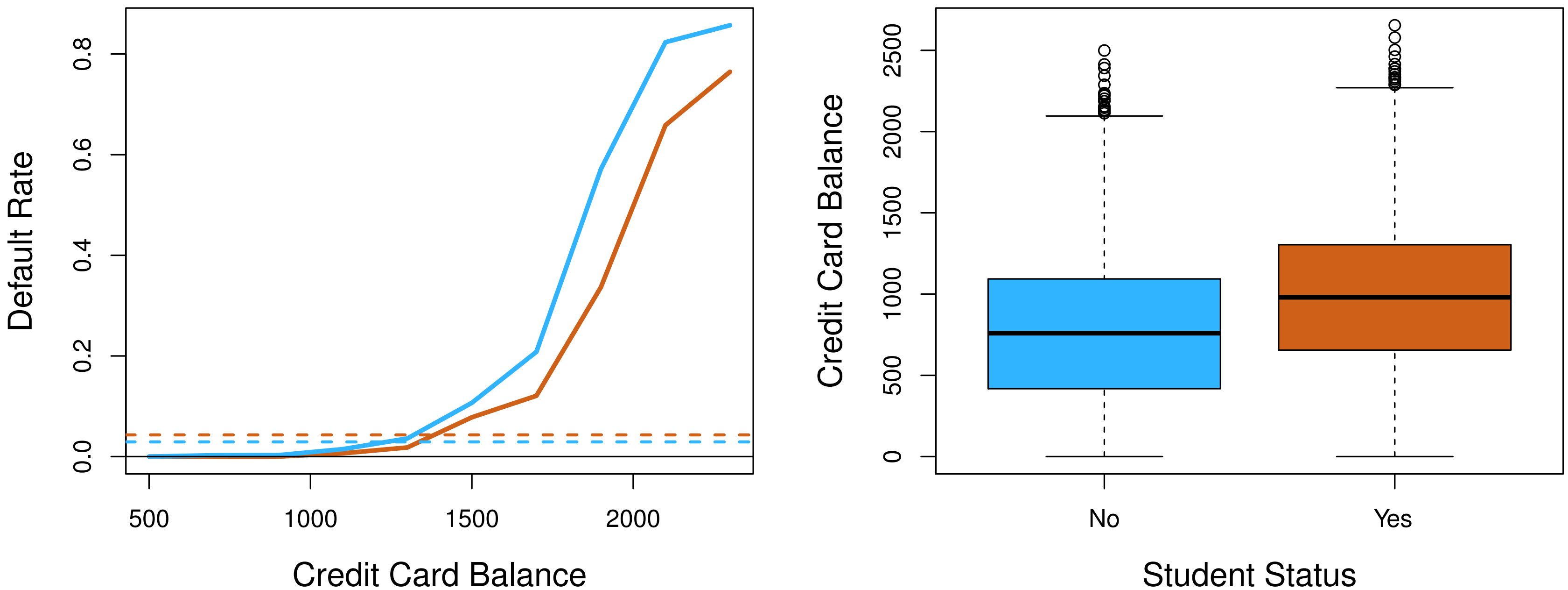
- Now fit the *default* model with several predictors

	Coefficient	Std. error	Z-statistic	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

- why is coefficient for student negative, while it was positive before?

Confounding

- Students tend to have higher balances than non-students (right)
  - so their marginal default rate is higher than for non-students,
- but for each level of balance, students default less than non-students (left),
- multiple logistic regression can tease this out.



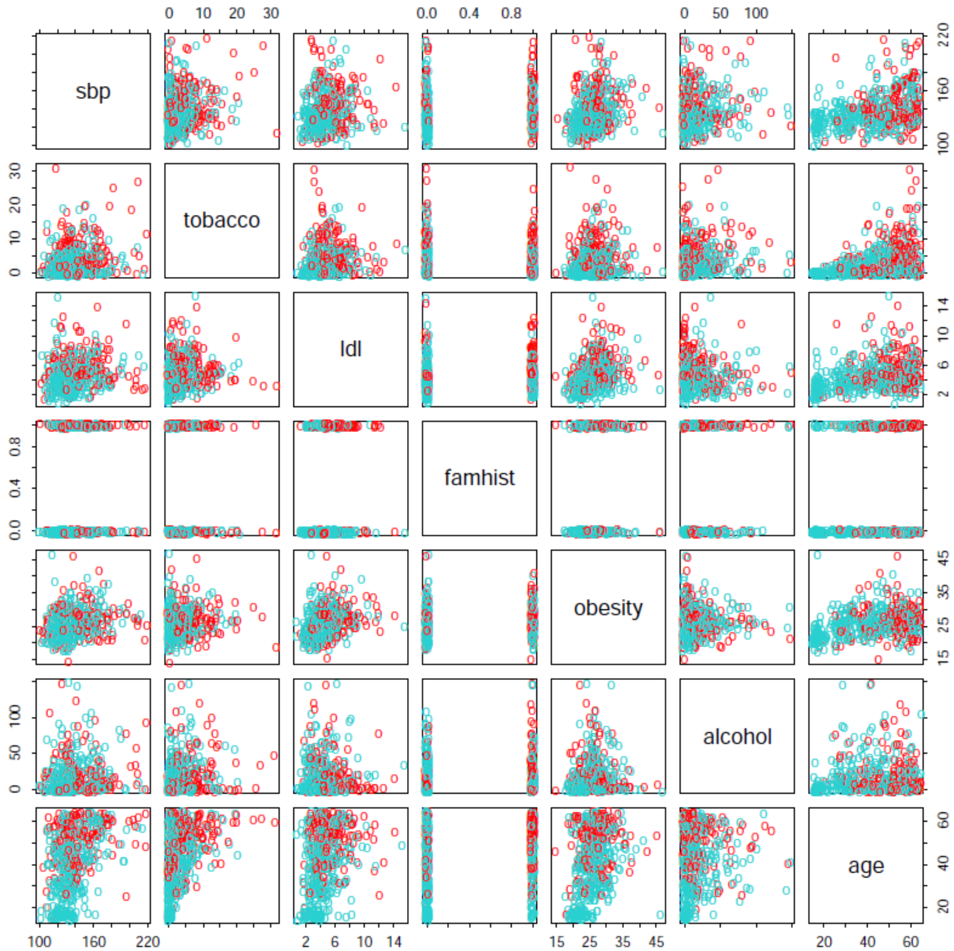
Example: South African Heart Disease

- 160 cases of MI (myocardial infarction) and 302 controls (all male in age range 15-64), from Western Cape, South Africa in early 80s,
- overall prevalence very high in this region: 5.1%,
- measurements on seven predictors (risk factors), shown in scatterplot matrix,
- goal is to identify relative strengths and directions of risk factors,
- part of an intervention study aimed at educating the public on healthier diets,
- **case-control sampling** and logistic regression
  - 160 cases, 302 controls $\rightarrow \hat{\pi} = 0.35$, yet the prevalence is $\pi = 0.051$,
  - with case-control samples, the regression parameter β_j estimates are accurate (if our model is correct),
  - only the constant term β_0 is incorrect, simple transformation helps

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\hat{\pi}}{1-\hat{\pi}}$$

Example: South African Heart Disease

- Scatterplot matrix, the response is color coded (MI=red,controls=turquoise),
- *famhist* is a binary variable, with 1 indicating family history of MI.



Example: South African Heart Disease

```
Call: glm(formula = chd ~ ., family = binomial, data = heart)

Coefficients:    Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp             0.0057607  0.0056326   1.023  0.30643
tobacco         0.0795256  0.0262150   3.034  0.00242 **
ldl             0.1847793  0.0574115   3.219  0.00129 **
famhistPresent  0.9391855  0.2248691   4.177 2.96e-05 ***
obesity        -0.0345434  0.0291053  -1.187  0.23529
alcohol         0.0006065  0.0044550   0.136  0.89171
age             0.0425412  0.0101749   4.181 2.90e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

Logistic regression with more than two classes

- So far, logistic regression with two classes only,
- it is easily generalized to more than two classes
  - in symmetric form, there is a linear function for each class

$$Pr(Y = k|\mathbf{X}) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\dots\beta_{pk}X_p}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_{1j}X_1+\cdots+\beta_{pj}X_p}}$$

  - this option is used e.g., in the R package *glmnet*,
  - in asymmetric form, one of the outcomes is selected as a pivot,
  - K-1 models are trained

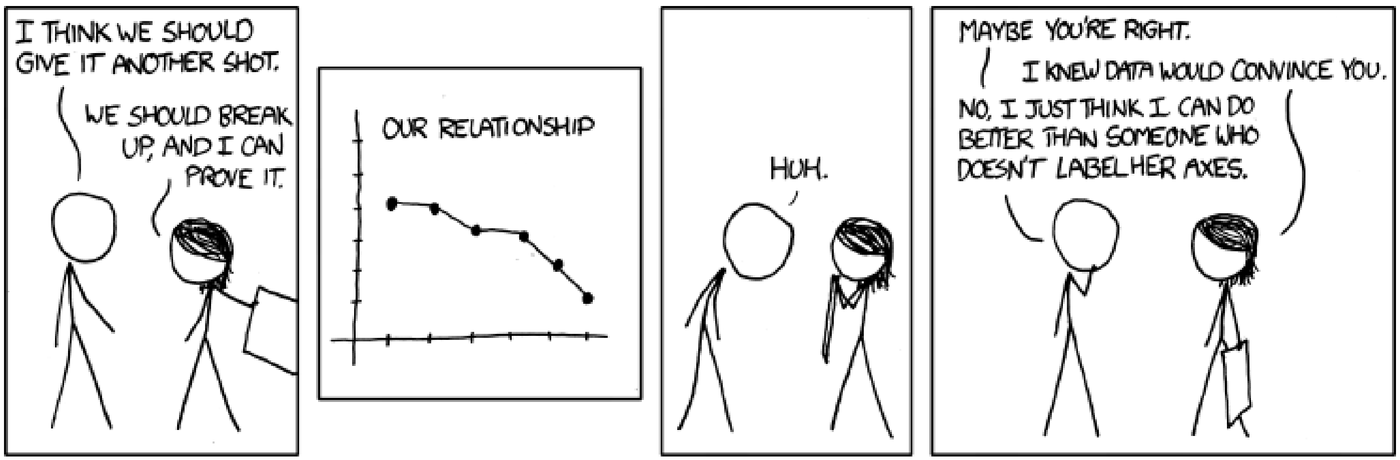
$$\forall i = 1 \dots K-1 \quad \frac{Pr(Y = i|\mathbf{X})}{Pr(Y = K|\mathbf{X})} = e^{\beta_{0i}+\beta_{1i}X_1+\cdots+\beta_{pi}X_p}$$

  - it can easily be shown that

$$Pr(Y = i|\mathbf{X}) = \frac{e^{\beta_\mathbf{i}\cdot\mathbf{X}}}{1+\sum_{j=1}^{K-1} e^{\beta_\mathbf{j}\cdot\mathbf{X}}} \quad Pr(Y = K|\mathbf{X}) = \frac{1}{1+\sum_{j=1}^{K-1} e^{\beta_\mathbf{j}\cdot\mathbf{X}}}$$

- multiclass logistic regression is also referred to as **multinomial regression**.

Scientific comics . . .



Taken from https://xkcd.com

Discriminant analysis

- The distribution of $\mathbf{X}$ in each of the classes modeled separately,
- Bayes theorem flips things around and helps to obtain $Pr(Y|\mathbf{X})$

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

- this approach is quite general,
- when we use normal (Gaussian) distributions for each class
  - this option leads to **linear or quadratic discriminant analysis**

$$Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x})}$$

  - where $f_k(\mathbf{x}) = Pr(\mathbf{X} = \mathbf{x}|Y = k)$ is the density for $\mathbf{X}$ in class k,
  - where $\pi_k = Pr(Y = k)$ is is the marginal or prior probability for class k.

Linear discriminant analysis for p=1

- Plug the Gaussian density model into Bayes formula $(p_k(x) = Pr(Y = k|X = x))$

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

- note, that we assume $\forall k \;\; \sigma_k = \sigma$ here,
- happily, there are simplifications and cancellations,
- maximize the **discriminant score** instead

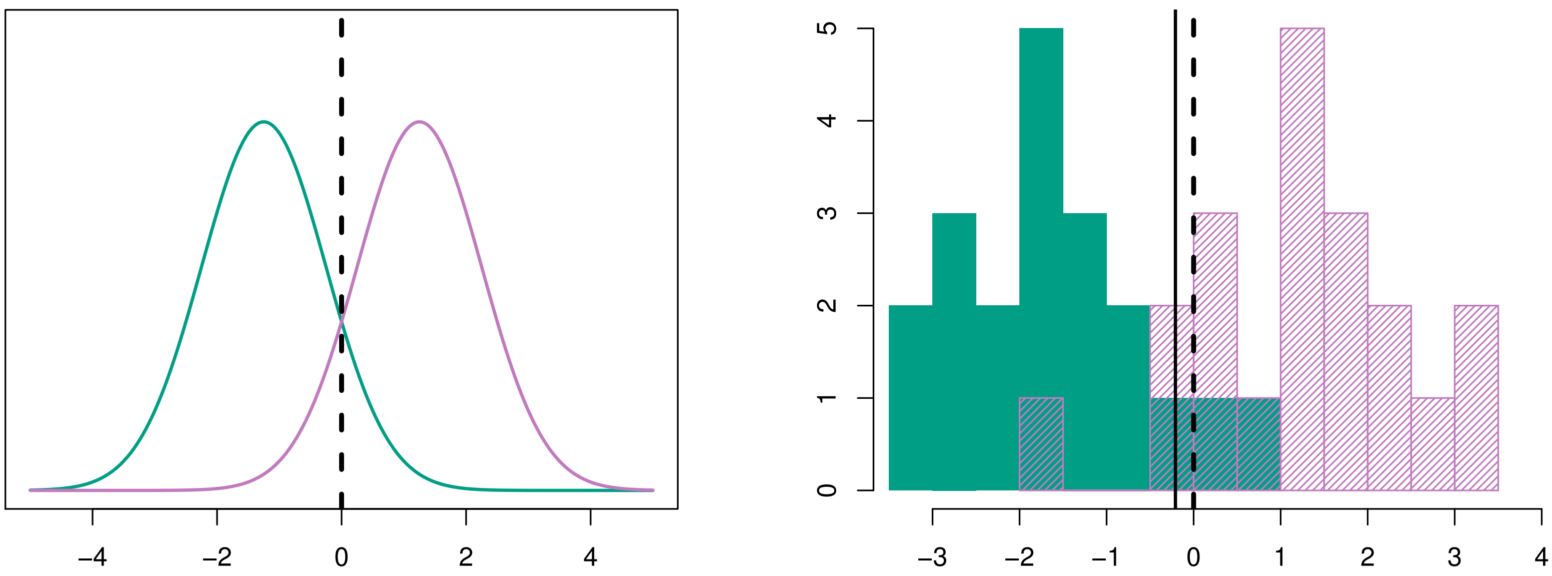
$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

- this is a linear function of $\mathbf{x}$,
- for $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, the decision boundary is at

$$x = \frac{\mu_1 + \mu_2}{2}$$

Estimating the parameters

- Typically these parameters are unknown, we estimate them from data,
- example below with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$ and $\sigma^2 = 1$



Linear discriminant analysis for $p > 1$

- Density: $f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$
- discriminant function: $\delta_k(\mathbf{x}) = \mathbf{x}^T\mathbf{\Sigma}^{-1}\mu_k^T - \frac{1}{2}\mu_k\mathbf{\Sigma}^{-1}\mu_k + log(\pi_k)$,
- despite its complex form a linear function of $\mathbf{x}$.

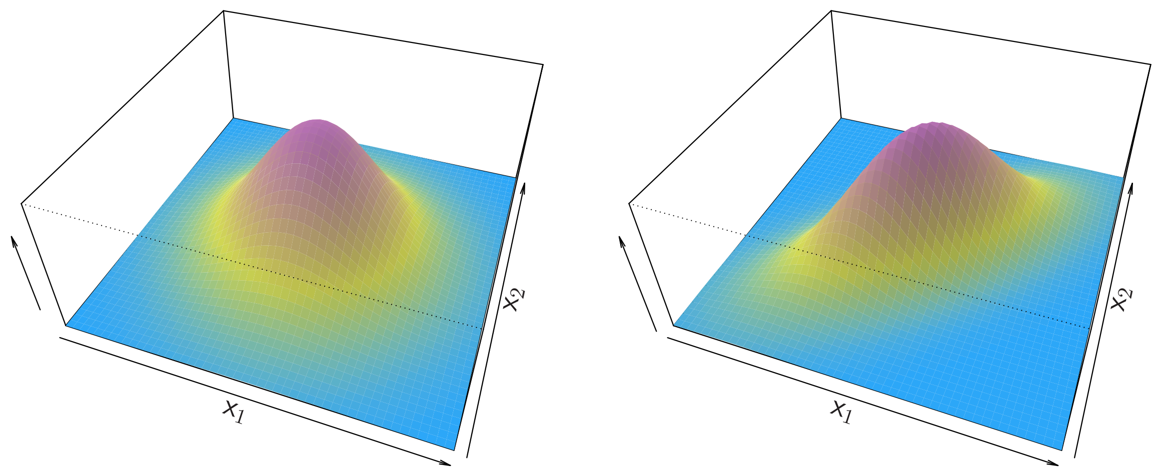
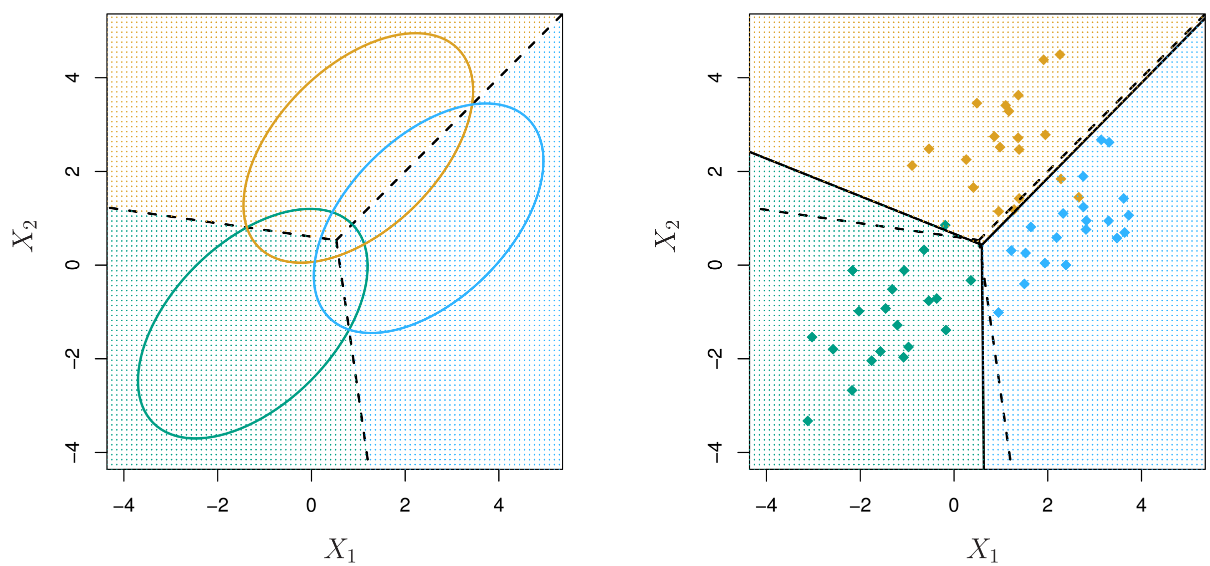


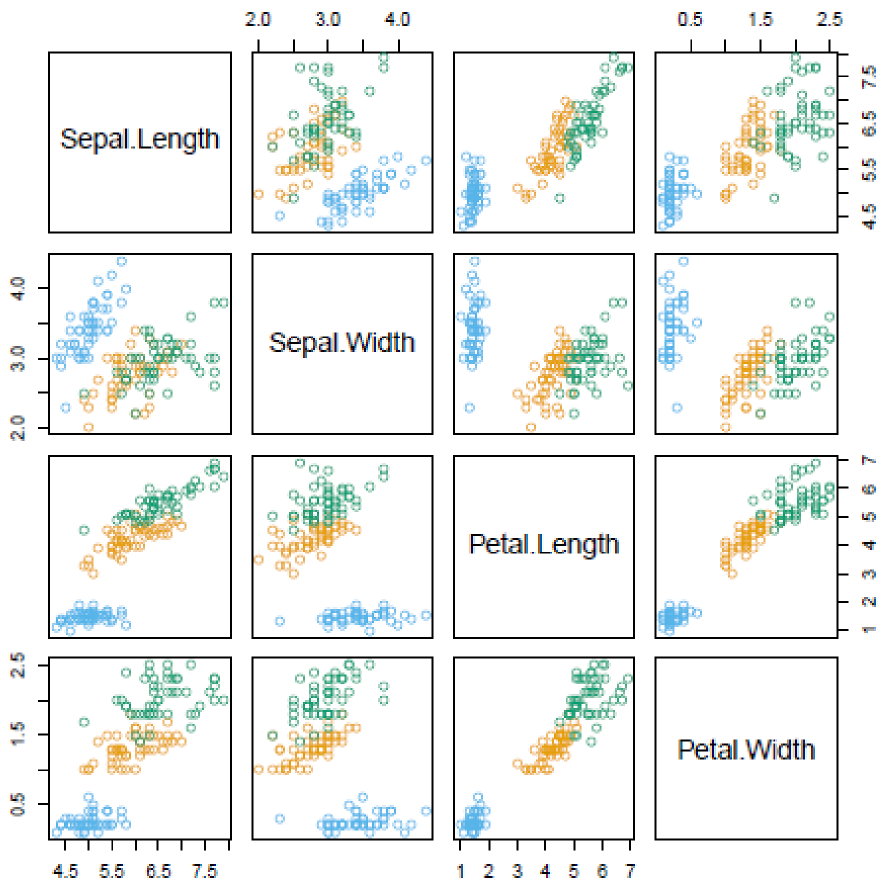
Illustration: $p = 2$ and $K = 3$ classes

- Three classes with the same priors, class-specific mean vectors and a **common covariance matrix**,
- ellipses in the left represent 95% confidence regions for each of the classes, dashed lines Bayes optimal decision boundaries,
- in the right LDA decision boundaries learned from a sample with 20 observations per class.



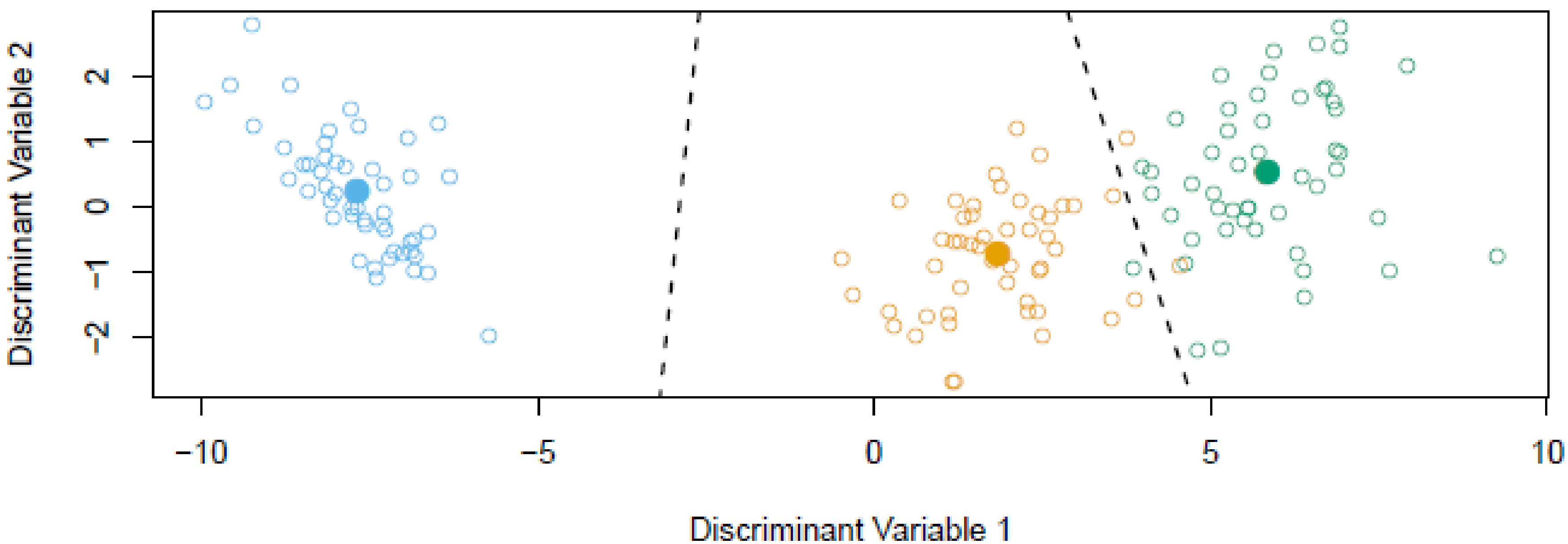
Fisher's Iris data

- Three classes/species: **setosa**, **versicolor**, virginica,
- 4 continuous features, 50 samples per class,
- LDA correctly classifies all but 3 training samples.



Fisher's discriminant plot

- LDA can be viewed in K-1 dimensional discriminant plot,
- it classifies to the closest centroid, they span a K - 1 dimensional plane,
- for $K > 3$ dimensionality reduction to visualize the discriminant rule.



From $\delta_k(\mathbf{x})$ to class probabilities

- Turn discriminant scores into class probability estimates

$$\hat{Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

- classifying to the largest $\delta_k(\mathbf{x})$ amounts to classifying to the class for which $Pr(Y = k|\mathbf{X} = \mathbf{x})$ is largest,
- when K $=$ 2, classify to class 2 if $Pr(Y = 2|X = x) \geq 0.5$, else to class 1,
- **confusion matrix** and classification accuracy can be employed then

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

Evaluation of a discriminative model

- This approach is often insufficient for
  - skewed classes (imbalanced class sizes),
  - unequal losses (different misclassification costs),
- for unequal losses, change the decision threshold from 0.5 to some other value from [0,1]
  - example: when predicting defaults in earlier Credit dataset, we would make nearly 80% error on the true Yes cases,
  - sensitivity is very low $\rightarrow$ changing the threshold adapts to a different loss function.

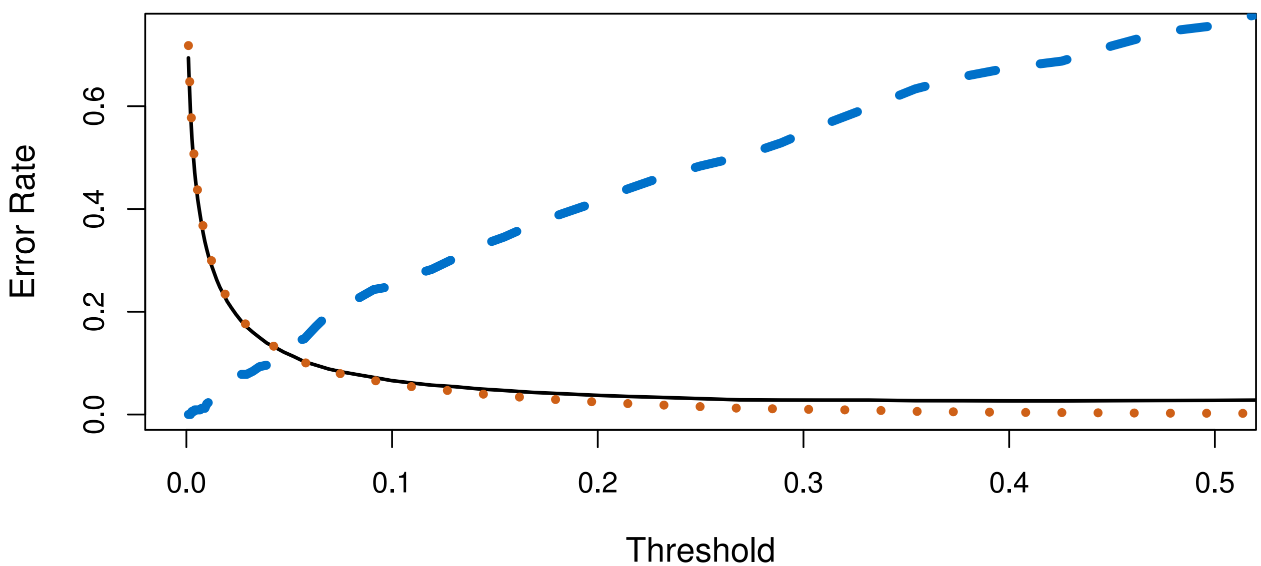
Unequal losses: Credit data

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- Credit data with skewed classes, observations for LDA
  - $\frac{23+252}{10000}$ errors $\rightarrow$ a 2.75% misclassification rate!
  - overfitting not a big concern here since $n = 10000$ and $p = 4$!,
  - if we always classified to class *No* in this case, we would make $\frac{333}{10000}$ errors, or only 3.33%,
  - of the true *No*'s, we make $\frac{23}{9667} = 0.2\%$ errors (false positive rate),
  - of the true *Yes*'s, we make $\frac{252}{333} = 75.7\%$ errors (false negative rate)!

Unequal losses: Credit data

- Let us change threshold in: if $Pr(Y = \text{default}|X = x) \geq$ thres then default
  - threshold 0.5 optimizes the overall error rate,
  - lower thresholds better fit the smaller class of defaulting customers (probably most interesting for a credit company).



Error rates as a function of the default threshold value, black solid . . . the overall error, blue dashed . . . the fraction of incorrectly classified defaulting customers (FNR), orange dotted . . . the fraction of incorrectly classified non-defaulting customers (FPR).

Receiver operating characteristics (ROC)

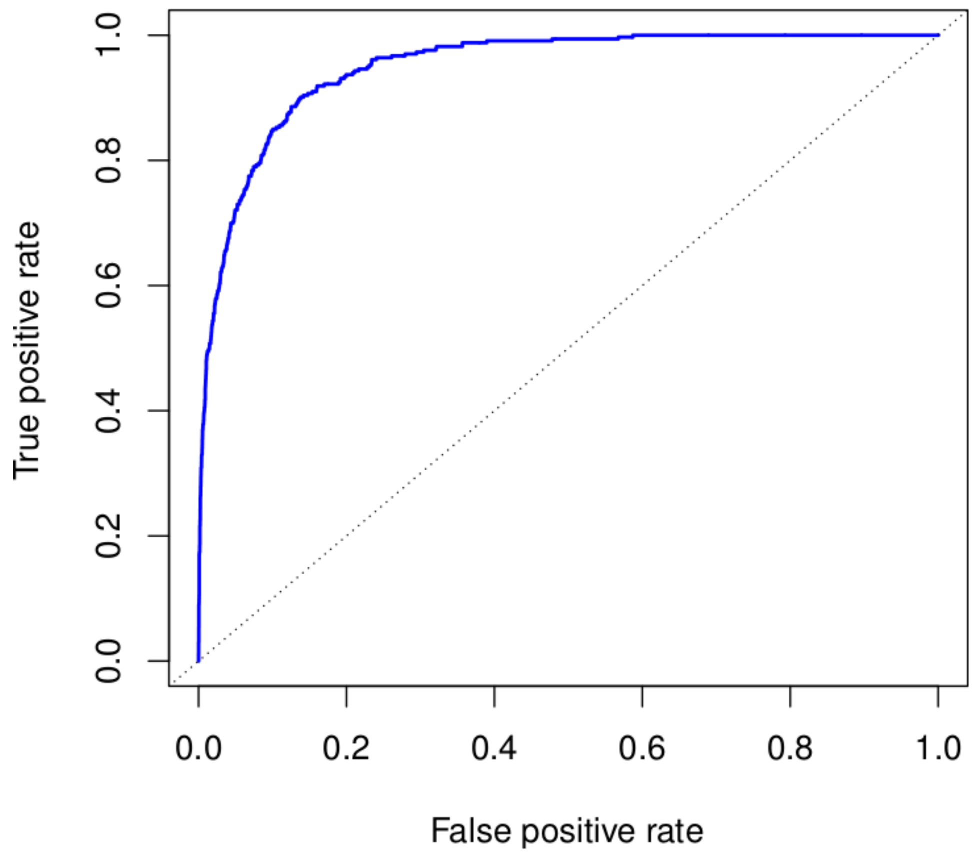
- Displays the errors for all possible thresholds
  - a popular way to evaluate probabilistic classifiers,
  - a better tool for imbalanced datasets than classification accuracy,
  - the overall performance of a classifier is given by the area under the ROC curve (AUC, AUROC), a number from $\langle 0, 1\rangle$, 0.5 for random votes,
  - AUROC represents the probability that a random positive example is positioned to the right of a random negative example on the scale given by the probabilistic classifier.

$$TPR = \text{sensitivity} = \frac{\text{number of true positives}}{\text{total number of positives}} = \frac{TP}{P}$$

$$FPR = 1 - \text{specificity} = \frac{\text{number of false positives}}{\text{total number of negatives}} = \frac{FP}{N}$$

Receiver operating characteristics (ROC): Credit data

- AUC is 0.95, which is close to the maximum of 1,
- the LDA classifier can be considered very good.



A ROC curve traces out TPR and FPR as we vary the threshold value for the posterior probability of default.

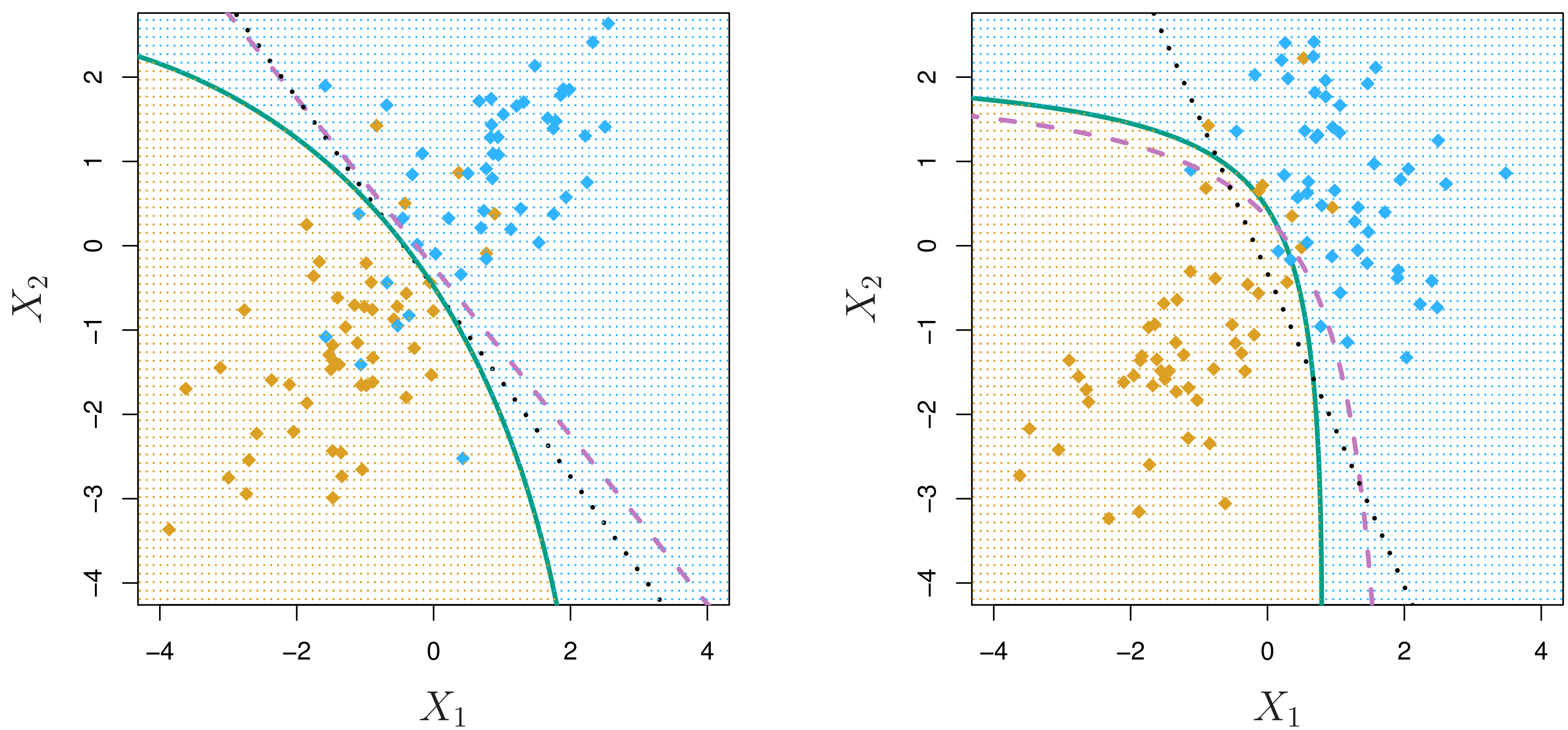
Other forms of discriminant analysis

- Consider different models in the general Bayes formula below

$$Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x})}$$

- when $f_k(\mathbf{x})$ are Gaussian densities, with the same covariance matrix $\boldsymbol{\Sigma}$ in each class
  - linear discriminant analysis,
- with Gaussians but different $\boldsymbol{\Sigma}_\mathbf{k}$ in each class
  - quadratic discriminant analysis,
- with $f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence assumption) in each class
  - naïve Bayes classifier,
  - for Gaussian this means the $\boldsymbol{\Sigma}_\mathbf{k}$ are diagonal.

Quadratic discriminant analysis



The Bayes optimal decision boundary in purple dashed line, LDA black dotted, QDA green solid. Left: the covariance matrices truly match, LDA is close to optimal solution, QDA suffers from higher variance. Right: the orange class has a positive correlation between predictors, the blue class negative, class covariance matrices differ, the optimal boundar is quadratic, LDA suffers from higher bias.

Summary

- Logistic regression
  - linear decision boundary, direct outcome on feature importance,
  - very popular especially when $K = 2$,
- LDA has a linear decision boundary too, it is useful when
  - the number of samples is small, or the classes are well separated, and Gaussian assumptions are reasonable,
  - $K > 2$, because it also provides low-dimensional views of the data,
- QDA constructs a non-linear (quadric) decision boundary
  - applies to a wider range of problems, more parameters, easier to overfit,
- naïve Bayes is useful when the dimension is very large,
- other classification algorithms
  - kNN, SVM, decision trees, neural networks,
- none of the methods dominates the others in every situation.

The main references

:: Resources (slides, scripts, tasks) and reading

- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R.** Springer, 2014.
- K. Markham: **In-depth Introduction to Machine Learning in 15 hours of Expert Videos**. Available at R-bloggers.