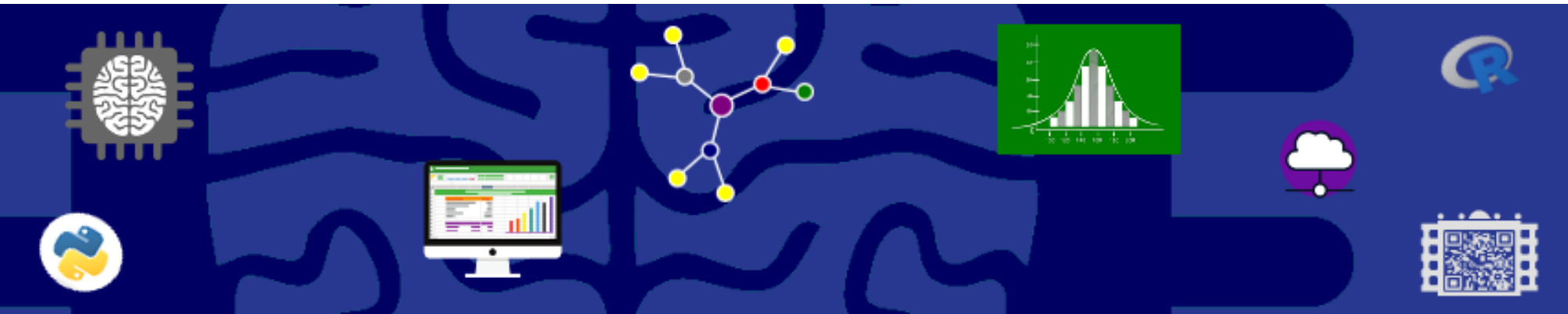
Introduction to Data Science

Winter Semester 2018/19

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

Contents II

Contents III

Contents

Classification

- **Classification**: response variable is **qualitative** or **categorical**.
- Involves assigning a predictor observation to a finite number of **classes** or **categories**.
- Likely more fundamental to human experience than regression.
  Examples: military triage, spam classification, fraud detection, tumor diagnostics, friend-foe distinction ...
- Common formulation: perform a linear regression, view (continuous) response result as probability of belonging to each class, choose class with larges probability.
- This chapter: 3 widely used classifiers:
  - **logistic regression**
  - **linear discriminant analysis** (LDA)
  - K-**nearest neighbors**

Contents

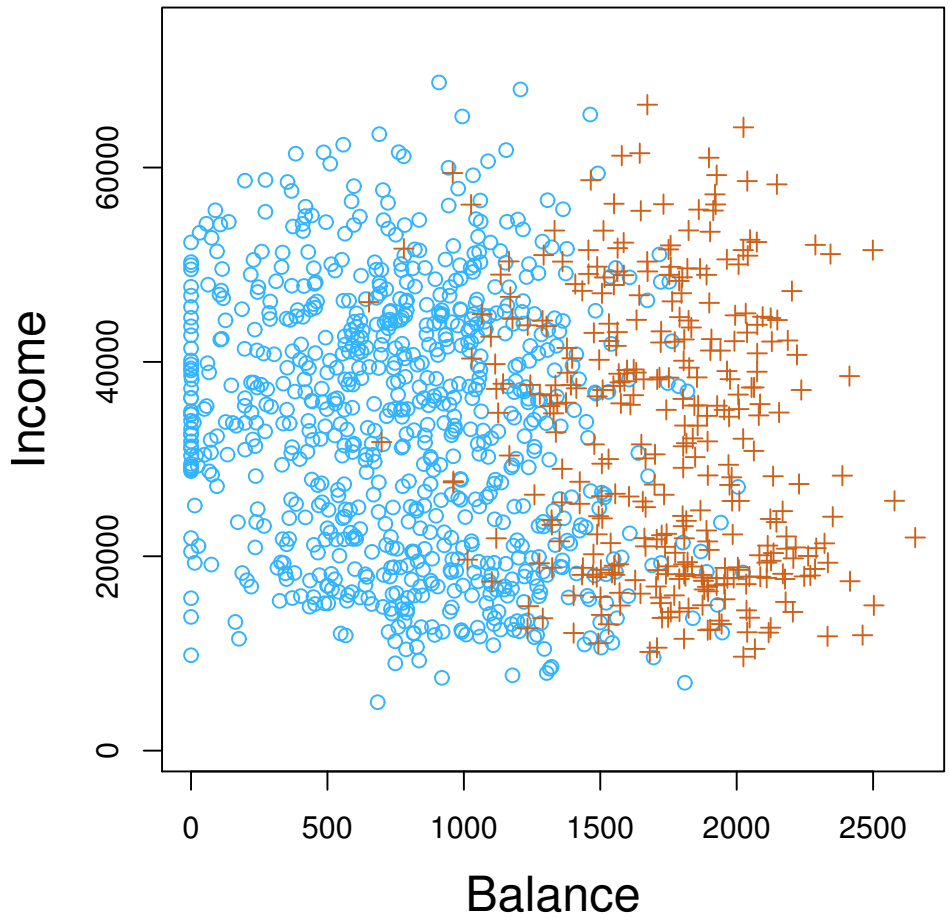
Overview of Classification

Setting

As for regression: use training observations $\{(x_i, y_i)_{i=1}^n\}$, to construct **classifier** able to perform classification also for test data not used in training.

`Default` data set: 10,000 individuals'

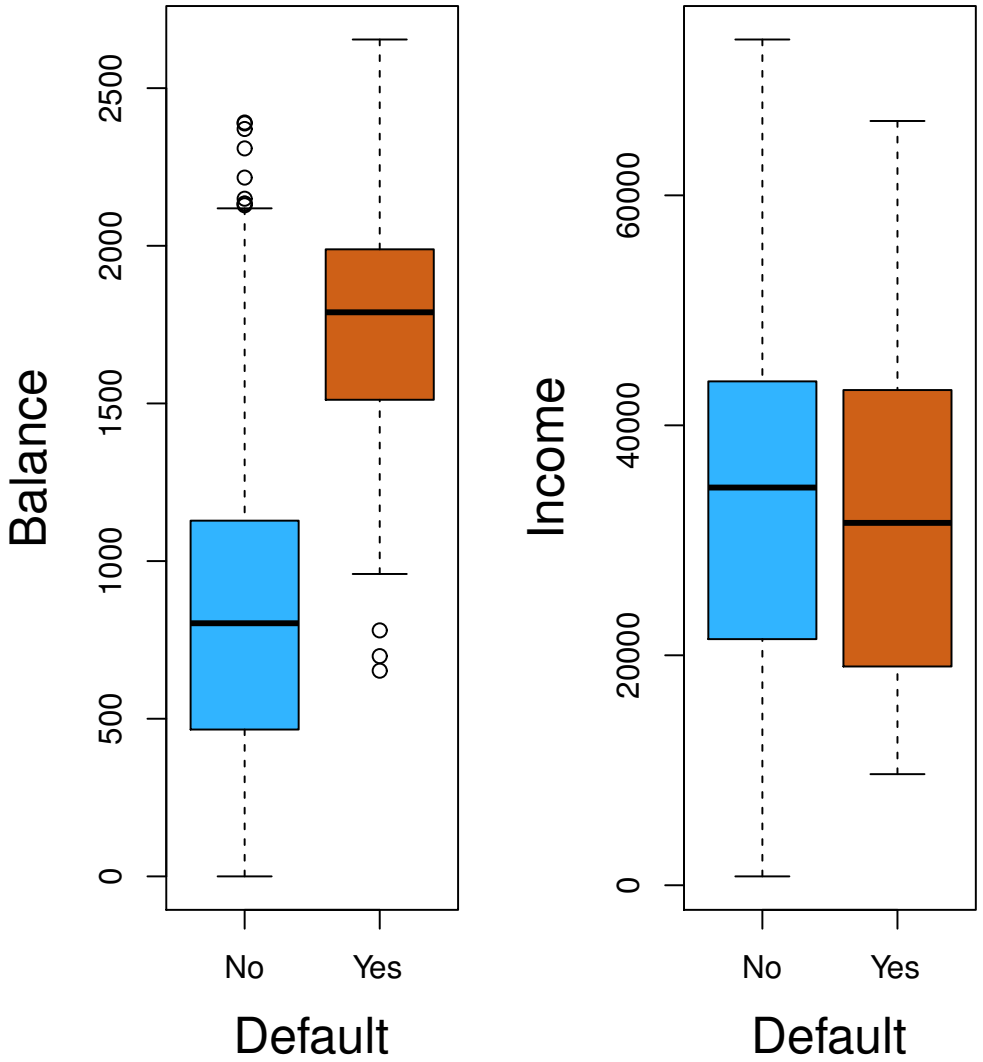
- annual `income` and
- monthly credit card `balance`.
- Response: binary `default` variable, i.e., whether or not person defaulted on their credit card payment in a given month.
- Overall default rate: 3%.



Overview of Classification

Setting

- **Box plots**: distributions of `balance` and `income` split by binary `default` variable.
- Objective: predict `default` (Y) for any pair of `balance` (X_1) and `income` (X_2) values.
- In this data: pronounced relationship between predictor `balance` and response `default`.



Contents

Why Not Linear Regression?

Ordering problem

Simplified model for predicting condition of incoming emergency room patients with possible diagnoses `stroke`, `drug overdose` or `epileptic seizure`.

Possible coding:

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

- Could perform linear regression based on available predictors $X_1, \ldots, X_p$.
- Coding implies (unnatural) ordering in outcome: places `drug overdose` between `stroke` and `epileptic seizure`.
- Also assumes distance between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.
- Different (equally reasonable) coding would lead to different linear model (and different predictions) for same data.

Why Not Linear Regression?

Ordering problem

- Sometimes underlying natural ordering exists (*mild*, *moderate*, *severe*).
- In general no way to map qualitative variable with > 2 values to quantitative response variable amenable to linear regression.
- For *binary* respone, e.g., only `stroke` and `drug overdose`, could use dummy variable approach and code

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}, \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

Following linear regression, could predict `drug overdose` if $\hat{Y} > 0.5$ and `stroke` otherwise.
Here flipping coding gives same results.
$X\hat{\beta}$ from linear regression yields estimate of probability

$$\mathbf{P}(\texttt{drug overdose}|X).$$

- For qualitative responses with > 2 values another approach is needed.

Contents

Logistic Regression

Idea

- `Default` data set, response variable `default` $\in \{$`Yes`, `No`$\}$.
- Logistic regression models *probability* of Y belonging to a particular class.
- Here: probability of default given `balance` denoted by

$$p(\texttt{balance}) := \mathbf{P}(\texttt{default} = \texttt{Yes} | \texttt{balance}) \in [0, 1].$$

- Predict `default` $=$ `Yes` whenever, e.g., $p($`balance`$) > 0.5$.
- More conservative credit card company might prefer lower threshold, e.g., $p($`balance`$) > 0.1$.

Logistic Regression

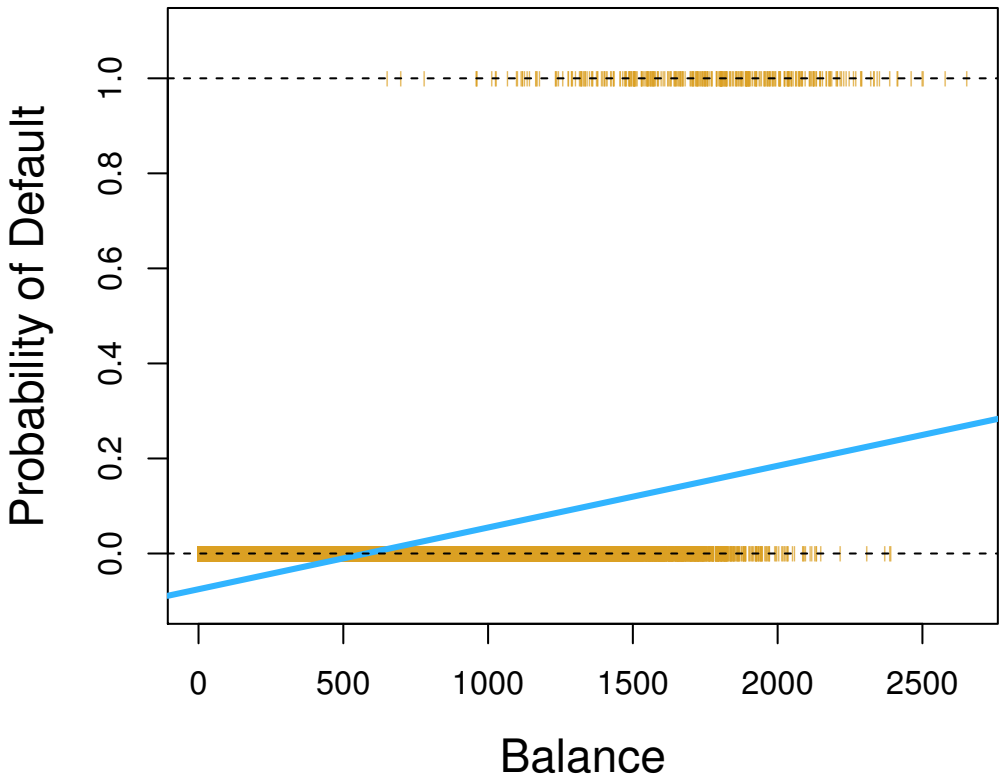
Logistic model

- Predicting `default = Yes` by modeling relationship between $p(X) = \mathbf{P}(Y = 1|X)$ and X by linear regression

$$p(X) = \beta_0 + \beta_1 X \qquad (4.1)$$

  gives fit on the right.

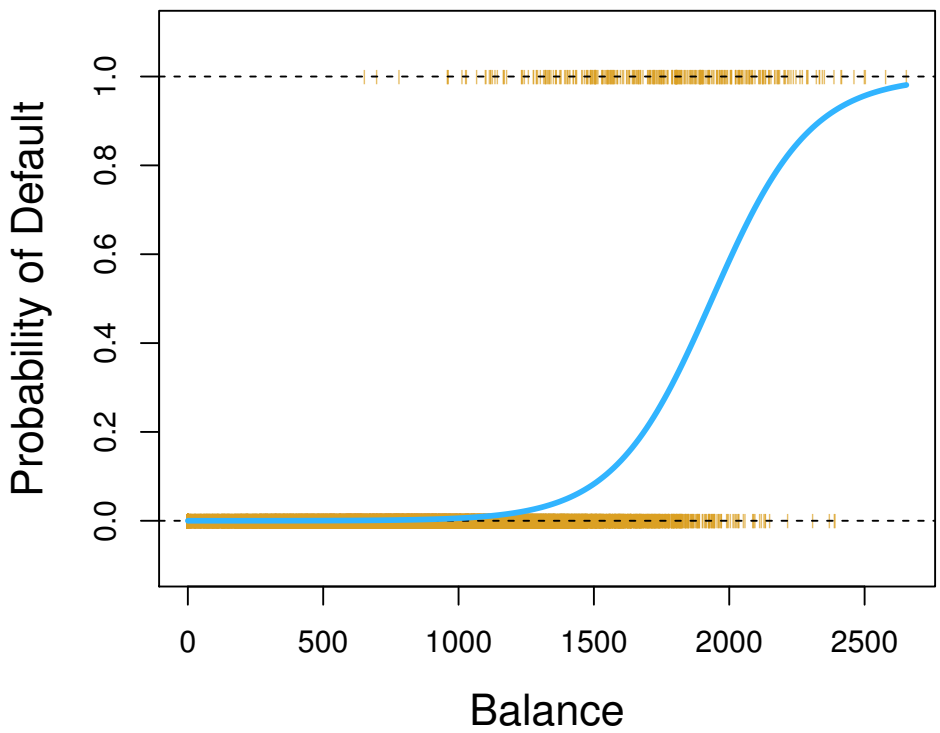
- Illustrates basic problem of fitting binary response coded with $\{0, 1\}$ with straight line: unless range of X limited, can always obtain probabilities outside $[0, 1]$.



Compose linear function with a **sigmoid** (monotone, S-shaped) function with values in $[0, 1]$, e.g., **logistic function**

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{4.2}$$

- Fit for `Default` data on the right.
- Average default rate in both cases (linear and logistic) 0.0333, close to overall proportion in data set.



- Rearranging (4.2) gives

$$\frac{p(X)}{1-p(X)} = \mathrm{e}^{\beta_0+\beta_1 X} \tag{4.3}$$

- Ratio on left: **odds**, $\in [0, \infty]$.
- Example: if 1 in 5 people default, odds are 1/4; for 9 in 10, odds are 9.
- Popular horse-racing terminology, as reflects appropriate betting strategy.
- Take logarithms on both sides of (4.3):

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X. \tag{4.4}$$

Lhs: **log-odds** or **logit**. Logistic regression model (4.2) has logit which is linear in X.

Logistic Regression

Logistic model: parameter β_1

- β_1: in linear regression, gives average change in Y per unit change in X; in logistic regression, reflects change in log-odds per unit change in X.
- Unit change in X changes odds by factor e^{β_1}.
  Due to nonlinearity, β_1 does not correspond to change in $p(X)$ due to unit change in X.
- Amount $p(X)$ changes depends on value of X.
- $\beta_1 > 0$ implies monotone increase of $p(X)$ with X, decrease for $\beta_1 < 0$.

$$p(X) = \frac{\mathrm{e}^{\beta_0+\beta_1 X}}{1+\mathrm{e}^{\beta_0+\beta_1 X}}$$

- **Maximum-likelihood estimation** (MLE) to determine estimates $\hat{\beta}_0, \hat{\beta}_1$ of coefficients β_0, β_1.
- **Likelihood function**

$$\ell(\beta_0, \beta_1) := \prod_{y_i=1} p(x_i) \cdot \prod_{y_i=0} (1 - p(x_i)), \tag{4.5}$$

  $p(x_i)$ determined from the observations (frequency).
- Estimates $\hat{\beta}_0, \hat{\beta}_1$ determined as $(\hat{\beta}_0, \hat{\beta}_1) := \arg\max \ell(\beta_0, \beta_1)$.
  This is a problem of numerical optimization methods, plenty of software available.
- Least squares can be viewed as a special case of MLE.

Logistic Regression

Estimating the regression coefficients

Coefficient estimates and statistics for logistic regression model on `Default` data set for predicting $\mathbf{P}(\texttt{default} = \texttt{Yes})$ with predictor `balance`:

	Coefficient	Standard error	z-statistic	p-value
β_0	-10.6513	0.3612	-29.5	< 0.0001
β_1	0.0055	0.0002	24.9	<0.0001

- Estimation accuracy measured by standard errors.
- z-statistic : analogous role here as t statistic in simple linear regression. For coefficient β_1:

$$z = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$$

- p-values strong evidence against $H_0 : \beta_1 = 0$, implying $p(X) = \mathrm{e}^{\beta_0}/(1+\mathrm{e}^{\beta_0})$.
- Intercept β_0 not of interest.

Logistic Regression

Making predictions: predictor `balance`

- Given this logistic regression model for `default` on `balance`, what probability for defaulting on payment can we predict for an individual with a balance of \$1000?

$$\hat{p}(X) = \frac{\mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+\mathrm{e}^{\hat{\beta}_0+\hat{\beta}_1 X}} \approx 0.00576 \approx 0.5\%.$$

- What about a balance of \$2000? Here

$$\hat{p}(X) \approx 0.5863 \approx 58\%.$$

Logistic Regression

Making predictions: predictor `student`

- For qualitative predictor variables, e.g., `student` in `Default` data set, use dummy variable taking value 1 for students, 0 for non-students.
- Resulting model: logistic regression of `default` on `student` status

	Coefficient	Standard error	z-statistic	p-value
β_0	-3.5041	0.0707	-49.55	< 0.0001
β_1	0.4049	0.1150	3.52	0.0004

- $\beta_1 > 0$, statistically significant.
- Model predicts higher default probability for students:

$$\hat{\mathbf{P}}(\texttt{default} = \texttt{Yes} | \texttt{student} = \texttt{Yes}) = \frac{\mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \cdot 1}}{1 + \mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \cdot 1}} \approx 0.0431,$$

$$\hat{\mathbf{P}}(\texttt{default} = \texttt{Yes} | \texttt{student} = \texttt{No}) = \frac{\mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \cdot 0}}{1 + \mathrm{e}^{\hat{\beta}_0 + \hat{\beta}_1 \cdot 0}} \approx 0.0292.$$

- For multiple predictor variables $X = (X_1, \ldots, X_p)$, generalize (4.4) to

$$\log \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{4.6}$$

or

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}, \tag{4.7}$$

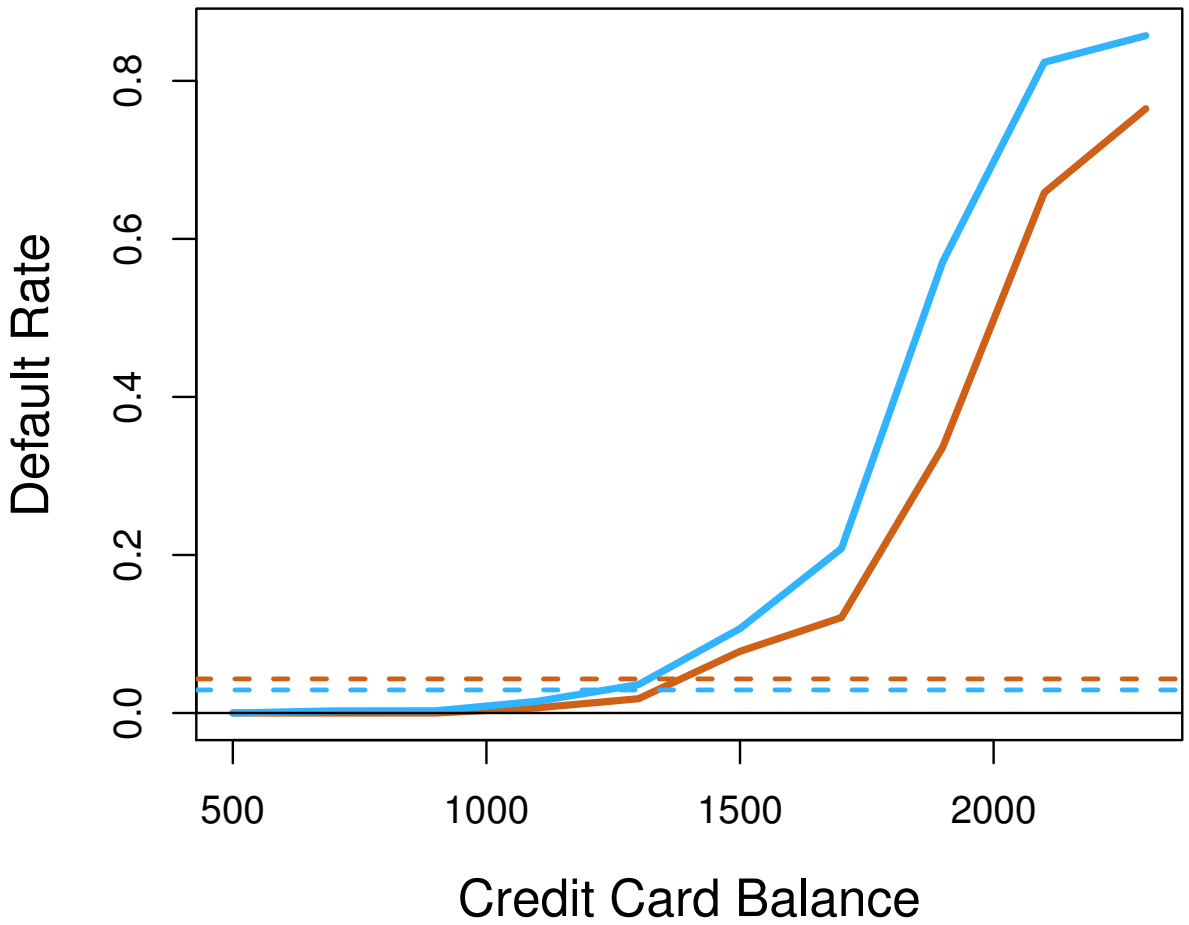
- Fit parameters again by MLE.
- Logistic regression predicting `default` based on `balance`, `income` and `student` status:

	Coefficient	Standard error	z-statistic	p-value
β_0	-10.8690	0.4923	-22.08	< 0.0001
β_1 (`balance`)	0.0057	0.0002	24.74	< 0.0001
β_2 (`income`)	0.0030	0.0082	0.37	0.7115
β_3 (`student`)	-0.6468	0.2362	-2.74	0.0062

Logistic Regression

Multiple logistic regression

- Coefficients of `balance` and `student` significant.
- Coefficient of `student` now negative! Explanation?

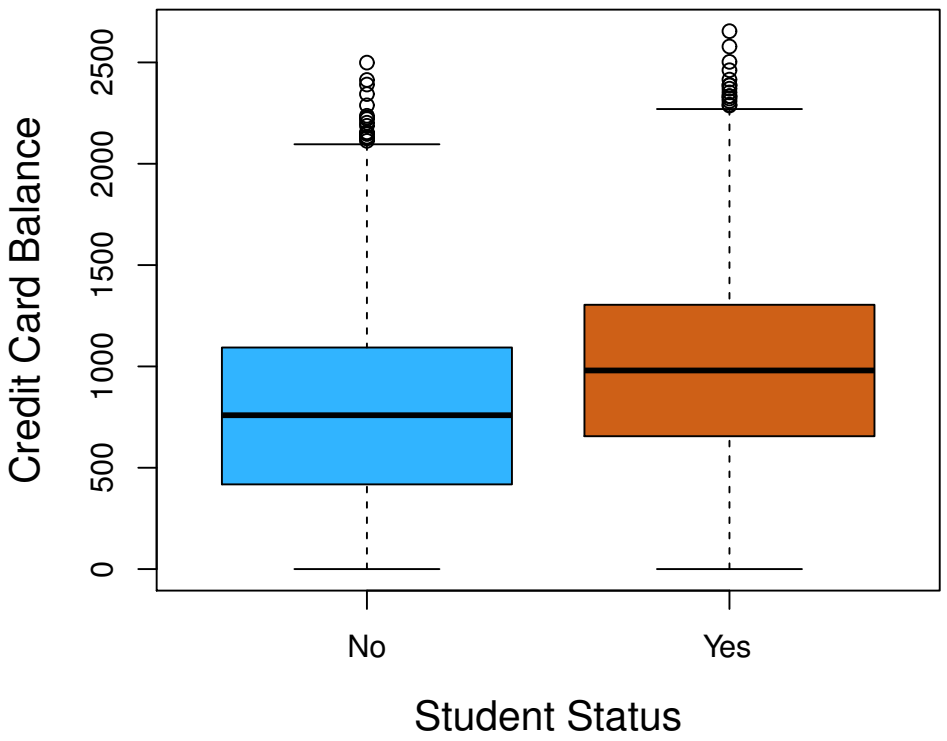


Orange: student
Blue: non-student

- Negative coefficient of `student`: for fixed value of `balance` and `income`, student *less* likely to default than non-student.
- Figure shows: student default rate at or below non-student rate for each value of `balance`.
- Horizontal broken lines: overall student default rate higher than non-student. Explains positive coefficient for `student` in single variable logistic regression.

Logistic Regression

Multiple logistic regression



- Variables `student` and `balance` correlated.
- Students tend to hold higher debt level, hence higher probability of default.
- Individual student with given balance will have lower default probability than non-student with same balance.

- Overall: student riskier than non-student.
- But: student less risky than non-student with same balance.
- Illustrates subtleties of ignoring further relevant predictors.
- Phenomenon: **confounding**.

- Student with credit card balance of \$1 500, income of \$40 000 has estimated probability of default

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\cdot 1500+\hat{\beta}_2\cdot 40+\hat{\beta}_3\cdot 1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\cdot 1500+\hat{\beta}_2\cdot 40+\hat{\beta}_3\cdot 1}} \approx 0.0549.$$

- For non-student, same credit card balance and income, estimate is

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\cdot 1500+\hat{\beta}_2\cdot 40+\hat{\beta}_3\cdot 0}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\cdot 1500+\hat{\beta}_2\cdot 40+\hat{\beta}_3\cdot 0}} \approx 0.1054.$$

Note: model fit was performed with units of \$1 000 for variable `income`.

Logistic Regression

Logistic regression for several response classes

- Recall emergency room example with 3 response classes `stroke`, `drug overdose` and `epileptic seizure`.
- Would like to model

$$\begin{aligned}
&\mathbf{P}(Y = \texttt{stroke}|X),\\
&\mathbf{P}(Y = \texttt{drug overdose}|X),\\
&\mathbf{P}(Y = \texttt{epileptic seizure}|X)\\
&\qquad = 1 - \mathbf{P}(Y = \texttt{stroke}|X) - \mathbf{P}(Y = \texttt{drug overdose}|X).
\end{aligned}$$

- Can extend two-class logistic regression to more than two, software available, but LDA more popular for this case.

Contents

Linear Discriminant Analysis

Conditional distribution

- Recall (ideal) Bayes classifier: assign to x_0 class $k \in \{1, \ldots, K\}$ such that

$$\hat{y}_0 = \hat{f}(x_0) = \underset{1 \le k \le K}{\arg\max}\, \mathbf{P}(Y = k | X = x_0).$$

- Logistic regression: model $\mathbf{P}(Y = k | X = x_0)$ using logistic function (4.7) when $K = 2$.
- Alternative approach LDA: model distribution of *predictors* X, then use **Bayes' rule** to turn these into estimates for $\mathbf{P}(Y = k | X = x_0)$.
- Motivation:
  - Logistic regression often unstable even for well-separated classes.
  - For small n, predictors approximately Gaussian across classes, LDA more stable than logistic regression.
  - LDA popular for $K > 2$.

Linear Discriminant Analysis

Bayes' rule (events)

Given **probability space** $(\Omega, \mathfrak{A}, \mathbf{P})$, $A, B \in \mathfrak{A}$, $\mathbf{P}(B) > 0$, then the **conditional probability** of A given B is defined by

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Solving for $\mathbf{P}(A \cap B)$, exchanging roles of A and B, assuming $\mathbf{P}(A) > 0$, gives

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\,\mathbf{P}(A)}{\mathbf{P}(B)} \qquad \textbf{Bayes' rule} \quad [\text{Bayes, 1763}]$$

- A: unobservable state of nature, with **prior probability** $\mathbf{P}(A)$ of occurring;
- B: observable event, probability $\mathbf{P}(B)$ known as **evidence**;
- $\mathbf{P}(B|A)$: probability that A causes B to occur (**likelihood**);
- $\mathbf{P}(A|B)$: **posterior probability** of A knowing that B has occurred.
- Terms: **inverse probability**, **Bayesian inference**.

Linear Discriminant Analysis

Bayes' rule (partitions)

Given partition $\{A_j\}_{j\in\mathbb{N}}$ of Ω into exhaustive and exclusive disjoint events, de Morgan's rule and countable additivity give, assuming all $\mathbf{P}(A_j) > 0$,

$$\mathbf{P}(B) = \sum_{j\in\mathbb{N}} \mathbf{P}(B|A_j)\,\mathbf{P}(A_j) \qquad \text{(law of total probability)},$$

leading to another variant of Bayes' rule:

$$\mathbf{P}(A_k|B) = \frac{\mathbf{P}(B|A_k)\,\mathbf{P}(A_k)}{\sum_{j\in\mathbb{N}} \mathbf{P}(B|A_j)\,\mathbf{P}(A_j)},$$

giving posterior probability of each A_k after observing B.

Linear Discriminant Analysis

Bayes' rule (densities)

Given real-valued **random variables** X, Y with **probability density functions** (pdfs)

- $f_X(x)$, $f_Y(y)$: density of X, Y at value x, y,
- $f_{X|Y}(x|y)$: density of $(X|Y)$ at x having observed $Y = y$,
- $f_{Y|X}(y|x)$: analogously.

Then Bayes' theorem states that

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)\, f_X(x)}{\int f_{Y|X}(y|x)\, f_X(x)\, \mathrm{d}x}.$$

- $f_{Y|X}(y|x)$ is now called the **likelihood function**.
- $\int f_{Y|X}(y|x)\, f_X(x)\, \mathrm{d}x$ is calles the **normalizing factor** or **marginal**.
- Short form:

$$f_{X|Y} \propto f_{Y|X}\, f_X.$$

Linear Discriminant Analysis

Using Bayes' rule for classification

- Goal: classify observation into one of $K \geq 2$ classes.
- $\pi_k := \mathbf{P}(Y(X) = k)$, $1 \leq k \leq K$, for randomly chosen X: **prior** probability.
- $f_k(x) := \mathbf{P}(X = x | Y = k)$, $1 \leq k \leq K$, **density function**[5] of X in class k. In other words: $f_k(x)$ large if probability that $X = x$ is large in class k.
- Bayes' rule:

$$p_k(x) := \mathbf{P}(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}. \tag{4.8}$$

- Idea: instead of computing $p_k(X)$ directly, estimate $f_k(X)$ and π_k, $k = 1, \ldots, K$ and insert into (4.8).
- If all estimates accurate, should come close to Bayes classifier.

[5] Modify accordingly for non-discrete predictor variable.

Linear Discriminant Analysis

LDA, $p = 1$

- Assumption: assume single predictor X has Gaussian distribution in each class, i.e.,

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp \frac{-(x-\mu_k)^2}{2\sigma_k^2}, \qquad k = 1, \ldots, K.$$

- Assume further that $\sigma_1 = \cdots = \sigma_K = \sigma$.
- Insert into (4.8):

$$p_k(x) = \frac{\pi_k \exp \frac{-(x-\mu_k)^2}{2\sigma^2}}{\sum_{j=1}^K \pi_j \exp \frac{-(x-\mu_j)^2}{2\sigma^2}} \tag{4.9}$$

- Classification: assign x to class k for which (4.9) is largest.
- Equivalent: class k for which

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

is largest.

Linear Discriminant Analysis

LDA, $p = 1$

Example:

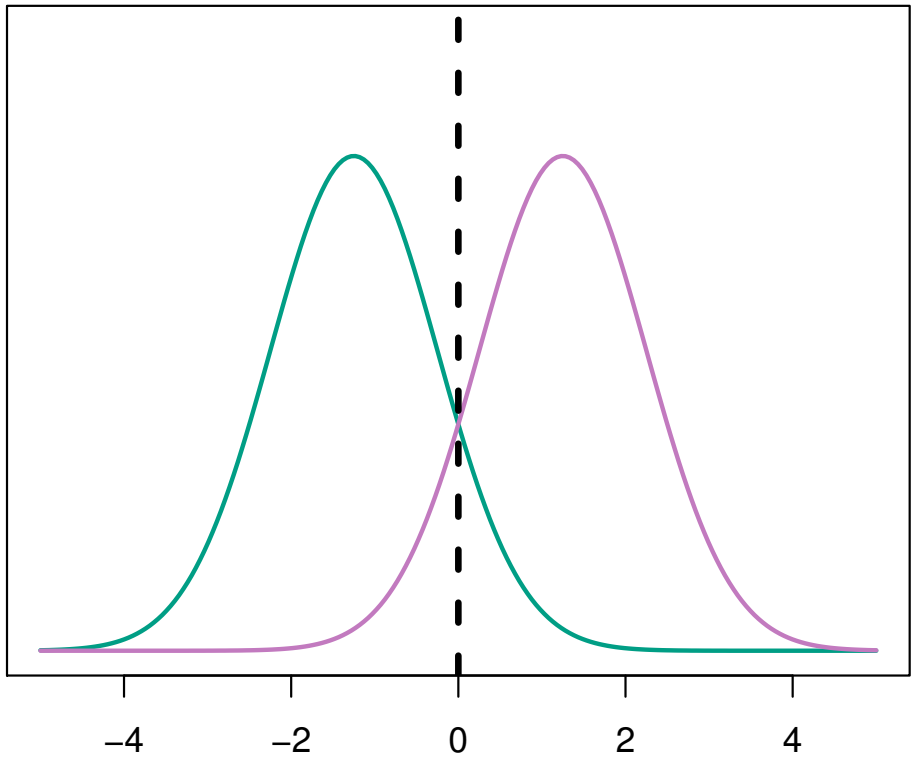
- $K = 2$, $\pi_1 = \pi_2$, assign x to class 1 if $\delta_1(x) > \delta_2(x)$ or

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2.$$

- Bayes decision boundary at

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

- In this case we can compute the Bayes classifier.



Two univariate normal densities with $\sigma_1 = \sigma_2 = 1$ and $\mu_1 = -\mu_2 = 1.25$, Bayes decision boundary (dashed black line).

Linear Discriminant Analysis

LDA, $p = 1$, estimating mean and variances

- LDA uses estimates for (ususally unknown) mean and variance:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{y_i = k} x_i, \qquad \hat{\sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{y_i = k} (x_i - \hat{\mu}_k)^2, \tag{4.10}$$

  N: total # observations, n_k: # observations in class k.

- Prior probabilities estimated as

$$\hat{\pi}_k = \frac{n_k}{n}. \tag{4.11}$$

- Classifier now assigns new observation to class k such that

$$k = \underset{1 \le k \le K}{\arg\max}\, \hat{\delta}_k(x), \qquad \hat{\delta}_k(x) := x \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k. \tag{4.12}$$

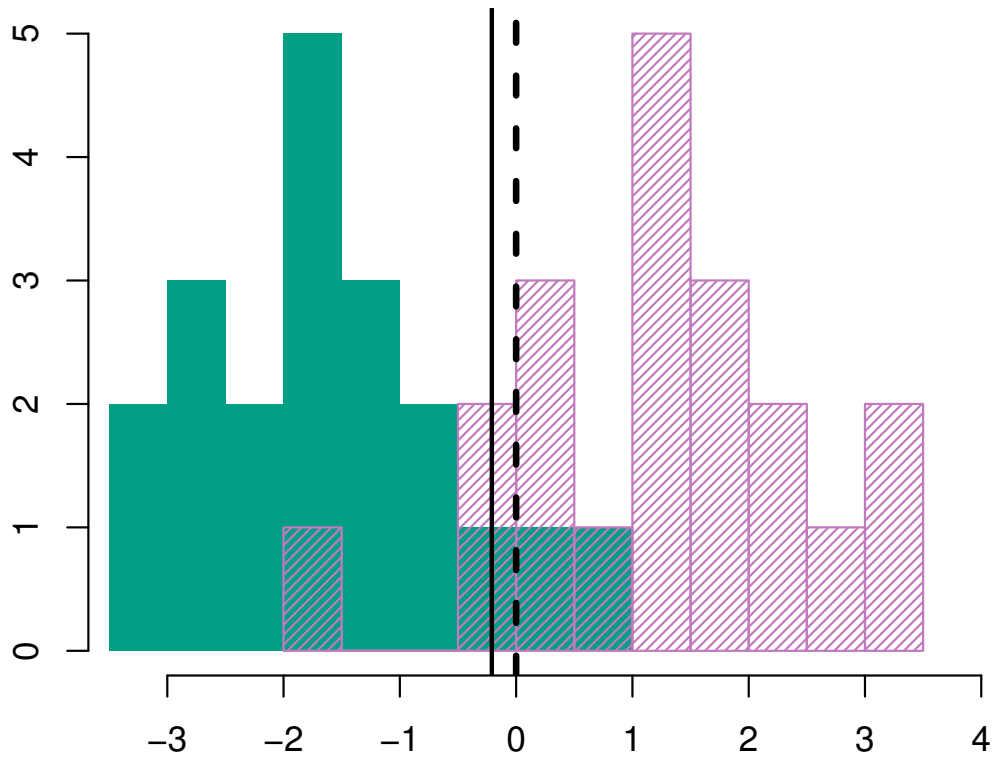
- LDA: **discriminant functions** $\hat{\delta}_k(x)$ are *linear* in x.

Linear Discriminant Analysis

LDA, $p = 1$, example

Example: (right)

- $K = 2$, $n = 20$ random observations from each class, estimate σ, μ_k, π_k.
- Bayes decision boundary given by solid black line; observations to the left assigned to green, otherwise purple.
- $n_1 = n_2 = 20 \Rightarrow \hat{\pi}_1 = \hat{\pi}_2$ and decision boundary at $(\hat{\mu}_1 + \hat{\mu}_2)/2$, slightly to left of Bayes decision boundary at $(\mu_1 + \mu_2)/2 = 0$.
- Bayes error rate: 10.6%, LDA test error rate: 11.1 %, i.e., only 0.5% over optimal!



Simulated data from 2 classes, LDA / Bayes decision boundaries.

Linear Discriminant Analysis

LDA, $p = 1$

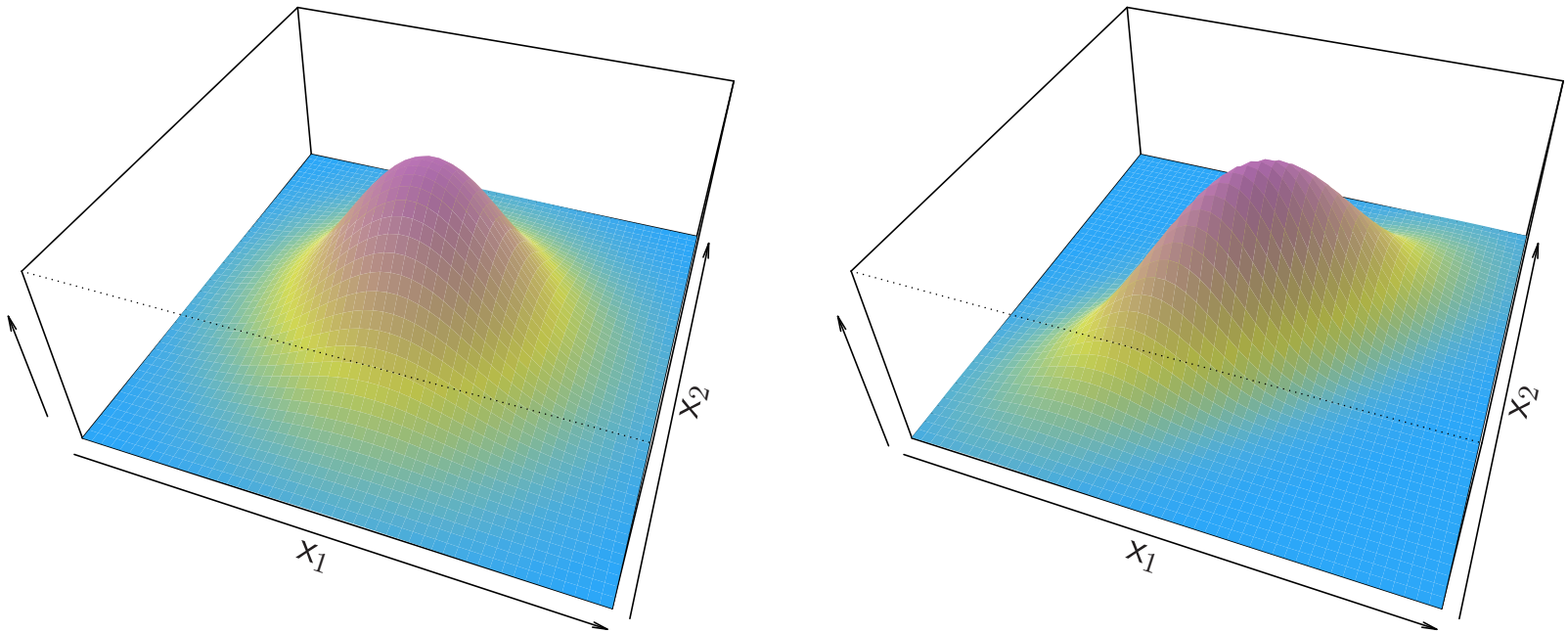
Recap: LDA classifier

- assumes observations within each class follow normal distribution,
- class-specific mean, common variance σ^2,
- estimates lead to classifier (4.12).

Linear Discriminant Analysis

LDA, $p > 1$

- For multiple predictor variables $X = (X_1, \ldots, X_p)$, assume observations follow multivariate normal distributions with class-specific mean, common covariance matrix.



Probability density functions (pdf) of two bivariate ($p = 2$) Gaussian distributions. Left: uncorrelated, right: correlation 0.7.

Linear Discriminant Analysis

LDA, $p > 1$

- Multivariate Gaussian:

$$X \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \mathbf{E}[X] \in \mathbb{R}^p, \quad \boldsymbol{\Sigma} = \mathbf{Cov}(X) \in \mathbb{R}^{p \times p}.$$

- Pdf:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{4.13}$$

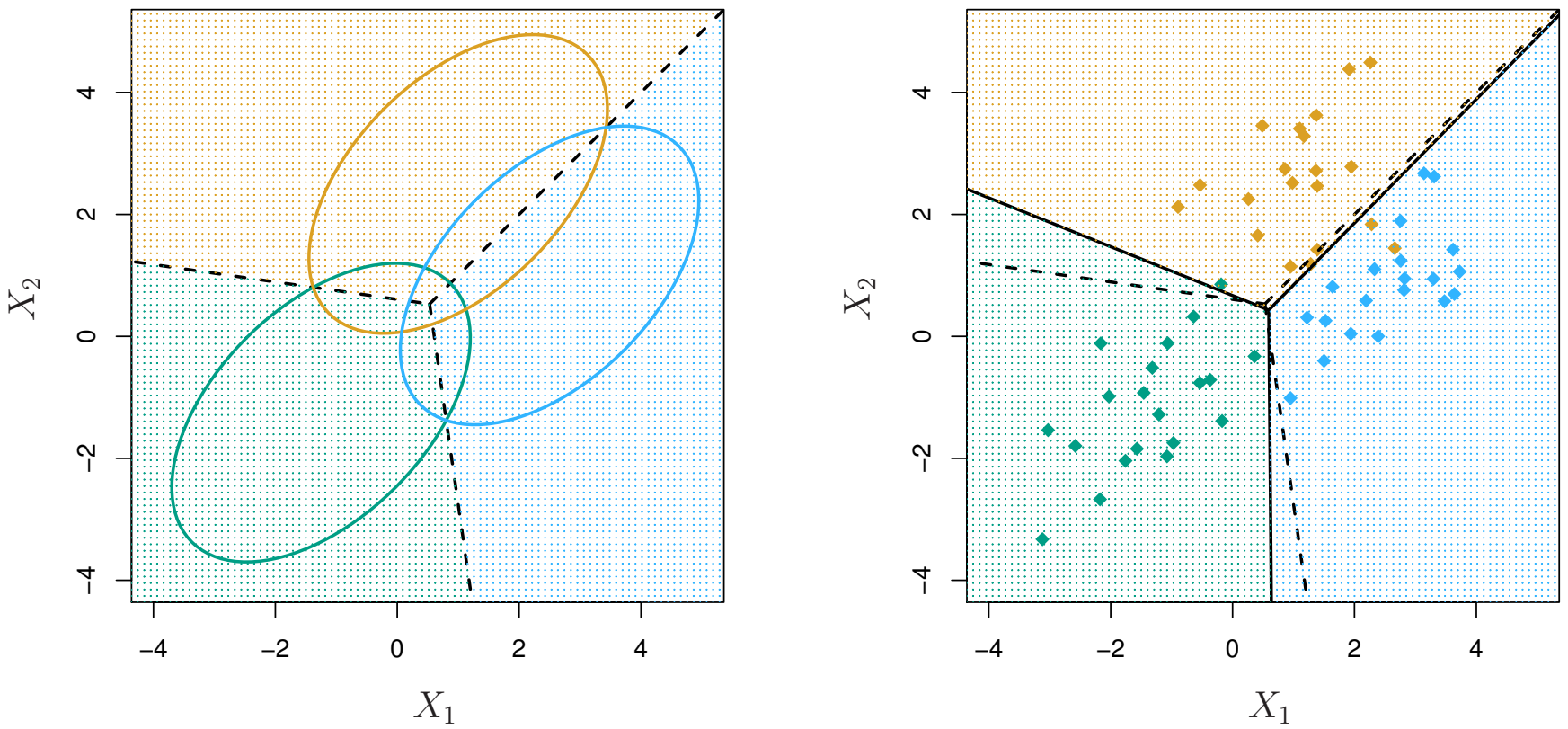
- For $p > 1$ assume within each class k: $X \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.
- Inserting pdf f_k into (4.8), we obtain Bayes classifier assigning observation $\boldsymbol{x}$ to class

$$k = \underset{1 \le k \le K}{\arg\max}\ \delta_k(\boldsymbol{x}), \qquad \delta_k(\boldsymbol{x}) := \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k. \tag{4.14}$$

This is the vector version of (4.12).

Linear Discriminant Analysis

LDA, $p > 1$, example



$p = 2$, $K = 3$, samples from three Gaussian distributions with means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$, common covariance matrix.

Left: 95%-ellipses, Bayes decision boundaries dashed.

Right: $n = 20$ random samples drawn from each class, their LDA classifications, Bayes decision boundary dashed, LDA decision boundary solid.

Linear Discriminant Analysis

LDA, $p > 1$, example

- Bayes decision boundaries: $\delta_j(\boldsymbol{x}) = \delta_k(\boldsymbol{x}), \quad j, k = 1, 2, 3,\ j < k.$
- LDA decision boundaries: $\hat{\delta}_j(\boldsymbol{x}) = \hat{\delta}_k(\boldsymbol{x}), \quad j, k = 1, 2, 3,\ j < k.$
- Unknown parameters

$$\{\pi_k\}_{k=1}^K, \quad \{\boldsymbol{\mu}_k\}_{k=1}^K, \quad \boldsymbol{\Sigma}$$

  estimated using formulas analogous to $p = 1$ case.
- Test error rates:
  Bayes: 0.0746
  LDA: 0.0770
- Again, conditional probability $\delta_k(\boldsymbol{x})$ in (4.14) *linear* in $\boldsymbol{x}$.

Linear Discriminant Analysis

LDA applied to `Default` data set

- Predict probability of defaulting on credit card payments given `balance` and `student` status.
- LDA model fit to $n = 10,000$ training samples results in *training* error rate of 2.75%. Low?
- Caveats:
  - Training error rates generally lower than test error rates.
  - High ratio of free parameters p to n poses overfitting danger, but here $p = 2$, $n = 10,1000$.
  - Overall, true default rate in `Default` training data only 3.33%.
    Implies (useless) constant classifier $Y \equiv 0$ has this low error rate.

Linear Discriminant Analysis

LDA applied to `Default` data set

Confusion matrix for LDA applied to `Default`:

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- Two types of misclassification errors.
- LDA: predicts 104 of 10,000 will default; of those, only 81 really defaulted. Hence, only 23 of 9,667 incorrectly labelled.
- However: of 333 who really defaulted, 252 (75.7%) missed by LDA.
- For credit card company trying to identify high-risk individuals: this false negative error rate probably unacceptable.

Linear Discriminant Analysis

Binary misclassification errors

- **Class-specific** classification errors can be crucial.
- In screening procedures (medical, airport):
  sensitivity: ratio of true positives identified;
  specificity: ratio of true negatives identified.
- In `Default` example:
  sensitivity $= 81/333 \approx 24.3\%$;
  specificity $= 9\,644/9\,667 \approx 99.8\%$.
- In hypothesis testing:
  type-I error: rejection of true null hypothesis (**false positive** finding);
  type-II error: failing to reject false null hypothesis (**false negative** finding).

Linear Discriminant Analysis

Example: mammography screening

What is the probability that a woman has breast cancer given (only) a positive result after undergoing a mammography screening?

Data on breast cancer screening test: [Kerlikowske & al., 1996]

Prevalence	1% (proportion of women who have breast cancer)
Sensitivity	90%
Specificity	91%

Bayes' rule: $Y \in \{0, 1\}$ (cancer?), $X \in \{0, 1\}$ (test positive?)

$$\mathbf{P}(Y = 1|X = 1) = \frac{\mathbf{P}(X = 1|Y = 1) \cdot \mathbf{P}(Y = 1)}{\mathbf{P}(X = 1|Y = 1) \cdot \mathbf{P}(Y = 1) + \mathbf{P}(X = 1|Y = 0) \cdot \mathbf{P}(Y = 0)}$$

$$= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + (1 - 0.91)(1 - 0.01)} \approx 9.2\%.$$

Do medical professionals know this? Don't count on it!
[Hoffrage & Gigerenzer, 1998]

Linear Discriminant Analysis

Modified LDA

- LDA approximates Bayes classifier, which has lowest *overall* error rate.
- However, sometimes important to achieve low error within a particular class of interest (credit card company, interested in defaulting customers).
- Bayes classifier: assign observation x to class k for which $p_k(x)$ largest.
  In two-class case of `Default` data set: assign to `default` class if

$$\mathbf{P}(\texttt{default} = \texttt{Yes} | X = x) > 0.5.$$

- To increase sensitivity to default, instead use lower threshold of

$$\mathbf{P}(\texttt{default} = \texttt{Yes} | X = x) > 0.2.$$

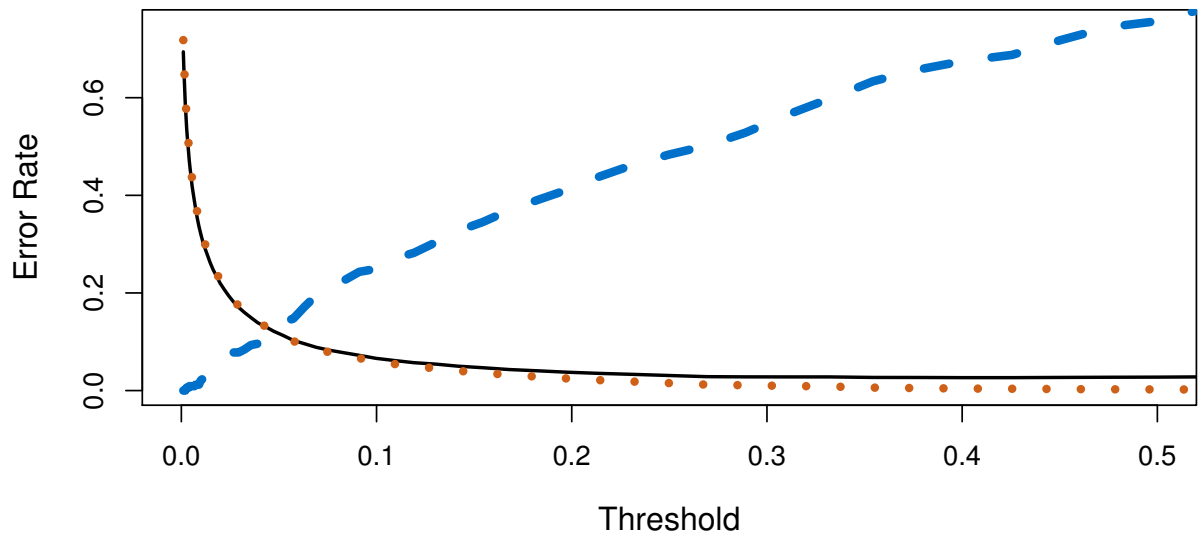
Modifies confusion table as follows: (cf. Slide 201)

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Linear Discriminant Analysis

Modified LDA

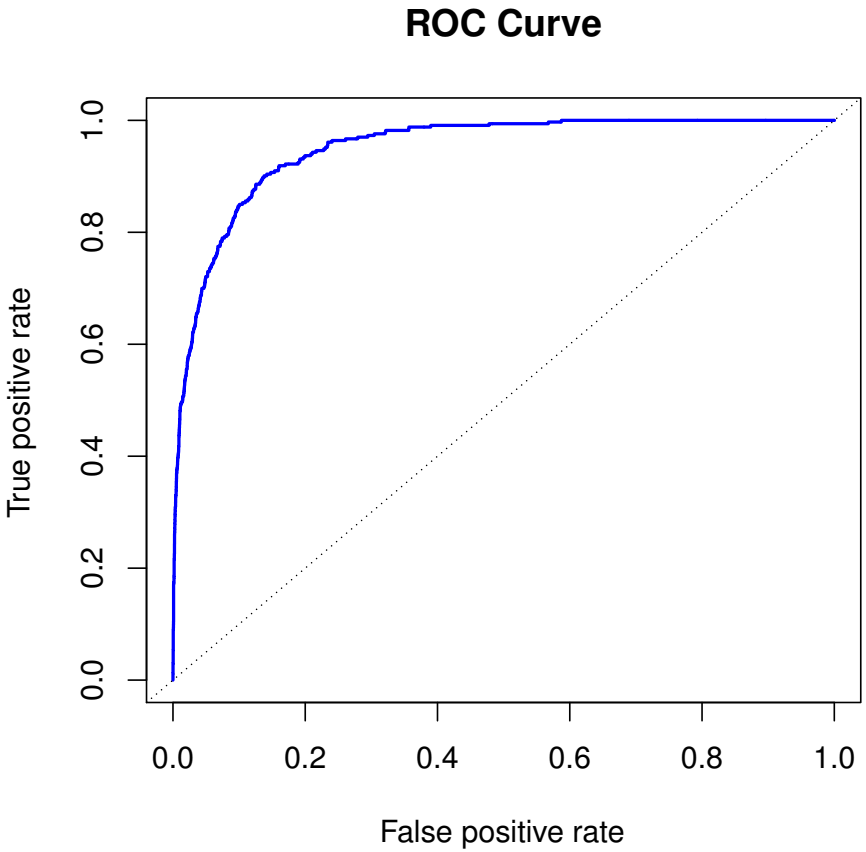
- LDA default prediction increases from 104 to 430. Default prediction error rate improves from $252/333 \approx 75.7\%$ to $138/333 \approx 41.4\%$.
- However, now 235 individuals who did not default are misclassified, raising the classification error in this class from $23/9,667 \approx 0.24\%$ to $235/9,667 \approx 2.4\%$, with an overall classification error of $(138 + 235)/10,000 = 3.73\%$.



`Default` data set: error rates versus threshold for LDA-assignment into defaulting class: black: overall, blue: fraction of defaulting customers misclassified; red: misclassified non-defaulting customers.

Linear Discriminant Analysis

ROC curve



- Traces out false positive/true positive rate for all threshold values of LDA classifier in `Default` data set.
- True positive: sensitivity (ratio defaulters correctly classified)
- False positive: 1−specificity (ratio of non-defaulters incorrectly classified).
- Optimal ROC curve: follows left/top boundaries.
- Dotted line: "no-information classifier", i.e., if student status and credit card balance unrelated to default.

- **Receiver Operating Characteristics** (ROC): simultaneous plot of both error types for all possible thresholds.
- **Area under the ROC curve** (AUC): overall performance of classifier summarized over all threshold values. Here AUC= 0.95 close to optimum 1.

Linear Discriminant Analysis

Summary of terminology

Possible results when applying a classifier (diagnostic test) to a population:

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
Total		N*	P*	

- Epidemiology context:
  +: disease, −: non-disease.
- Hypothesis testing context:
  −: null hypothesis, +: alternative (non-null) hypothesis.
- `Default` data set context:
  +: defaulting customer, −: non-defaulting customer.

Linear Discriminant Analysis

Performance measures for binary classification

Name	Definition	Synonyms
False Pos. rate	FP/N	Type-I error, 1− specificity
True. Pos. rate	TP/P	1− Type-II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1− false discovery proportion
Neg. Pred. value	TN/N*	

N: population negative
P: population positive
N*: predicted negative
P*: predicted positive

- **Quadratic discriminant analysis** (QDA): assume observations within each class follow Gaussian distribution, but each class has distinct covariance matrix, i.e., observation in k-th class given by random variable

$$X \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Assign observation $X = \boldsymbol{x}$ to class k which maximizes discriminant

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\log\det\boldsymbol{\Sigma}_k + \log\pi_k \\ &= -\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log\det\boldsymbol{\Sigma}_k + \log\pi_k. \end{aligned} \tag{4.15}$$

- Now discriminants depend *quadratically* on observation $\boldsymbol{x}$.

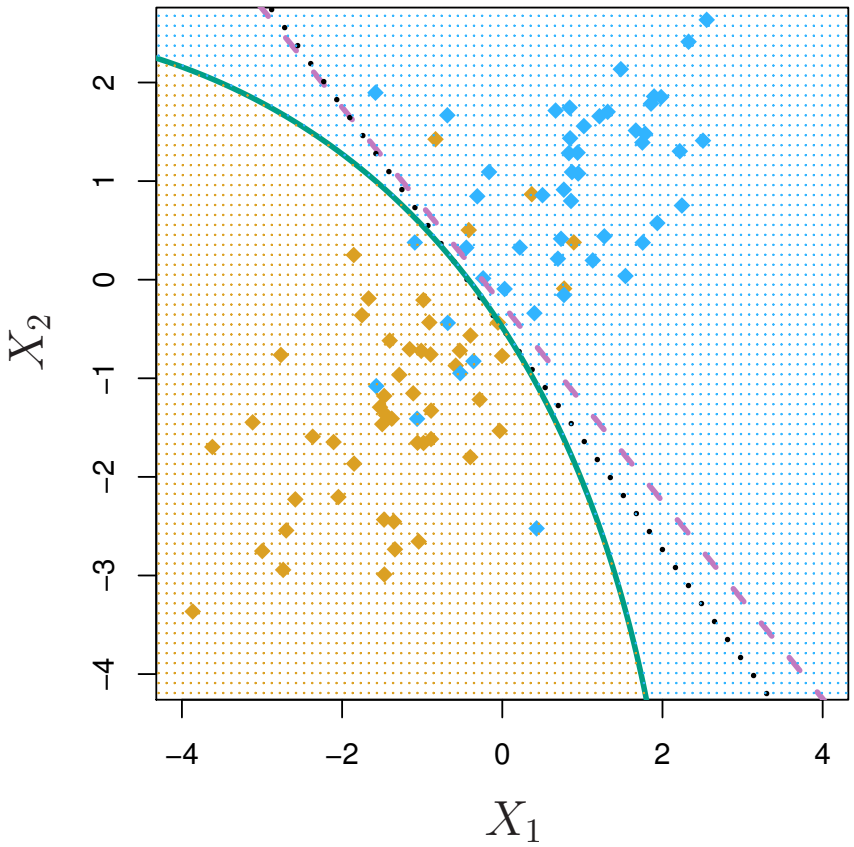
Linear Discriminant Analysis

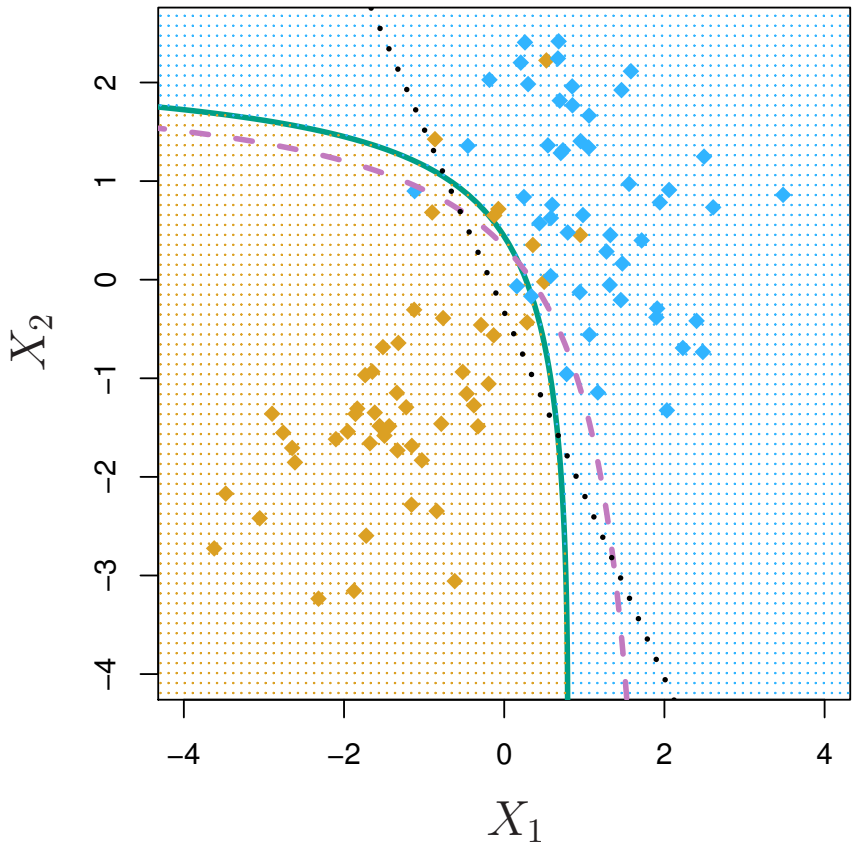
Quadratic discriminant analysis

- Requires estimation of π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$.
- Possible advantage of QDA over LDA: bias-variance trade-off.
- LDA estimates single covariance matrix: $p(p+1)/2$ parameters.
  QDA estimates K covariance matrices: $Kp(p+1)/2$ parameters.
- For 50 predictors this amounts to $K \cdot 1275$ parameters.
- LDA: larger bias, use for few training observations;
  QDA: larger variance, use for many training observations or when common covariance matrix known to be false.

Linear Discriminant Analysis

Example: LDA vs. QDA





Two-class problem, decision boundaries: Bayes (purple dashed), LDA (black dotted) and QDA (green solid). Shading: QDA classification. Left:$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Right:$\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

Contents

LDA vs. logistic regression: consider $p = 1$, $K = 2$.

- $p_1(x)$, $p_2(x) = 1 - p_1(x)$: probability x belongs to class 1, 2, respectively.
- log-odds for LDA:

$$\log \frac{p_1(x)}{1 - p_1(x)} = \log \frac{p_1(x)}{p_2(x)} = c_0 + c_1 x,$$

  c_0, c_1 functions of μ_1, μ_2, σ^2.
- log-odds for logistic regression:

$$\log \frac{p_1(x)}{1 - p_1(x)} = \beta_0 + \beta_1 x.$$

- Both linear in x, hence produce linear decision boundaries.
- β_0, β_1 via MLE, c_0, c_1 from estimation of mean, variance of Gaussians.
- Same relation between LDA and logistic regression holds for $p > 1$.
- LDA and logistic regression can give differing results if assumptions on Gaussian distribution not met, in this case logistic regression superior.

KNN

- Prediction for observation $X = x$ based on K training observations closest to x. Class selected based on majority of neighbors.
- Non-parametric, no assumptions on shape of decision boundary, hence expected to be superior to LDA and logistic regression when decision boundary highly nonlinear.
- KNN, however, gives no information on relative importance of predictor variables.

QDA

- Compromise between non-parametric KNN and linear LDA/logistic regression.
- Less flexible than KNN.
- Makes some assumptions on decision boundary shape, can perform better for limited observation numbers.

A Comparison of Classification Methods

Comparison scenarios

Six scenarios for comparison: 3 linear, 3 nonlinear decision boundaries.
100 random training data sets each.
For KNN used $K = 1$ and value chosen by *cross validation* (later).

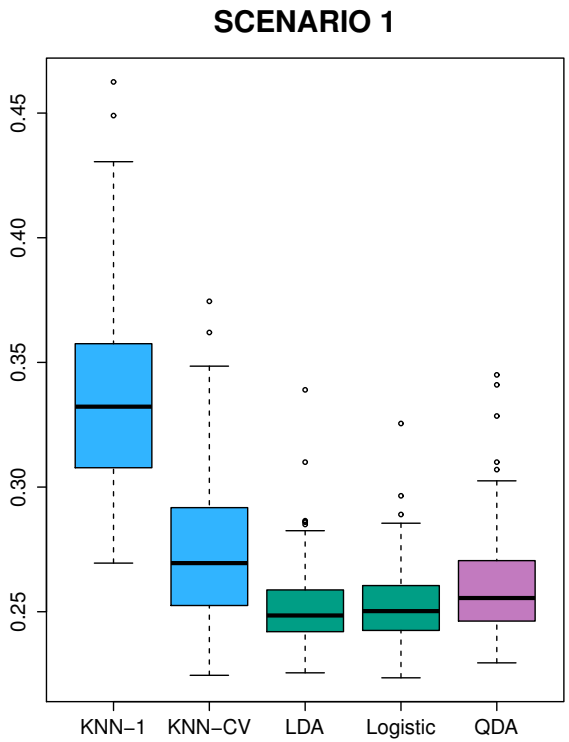
Scenario 1 20 training observations in each of 2 classes; in each class: uncorrelated Gaussian with separate means. LDA performs well, KNN high variance not offset by reduction in bias. QDA worse than LDA since classifier more flexible than necessary. Logistic regression: only slightly worse than LDA (linear decision boundary).

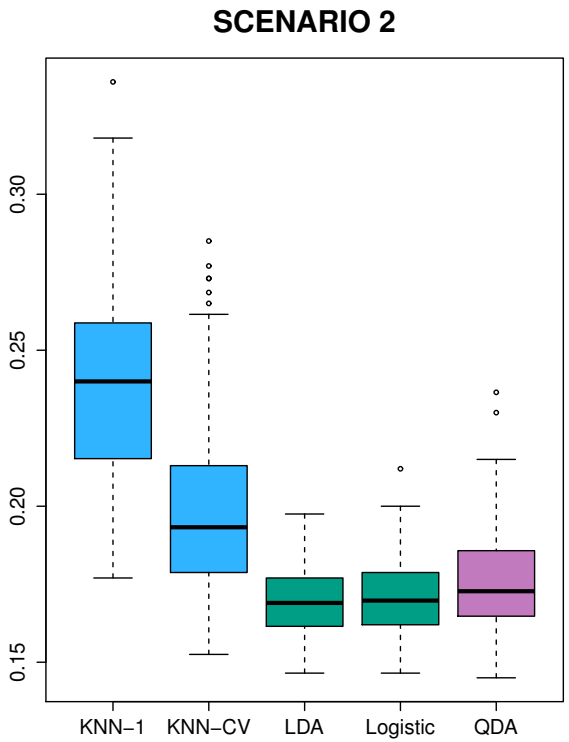
Scenario 2 Same as Scenario 1 except that within each class the 2 predictors had correlation -0.5. Little change.

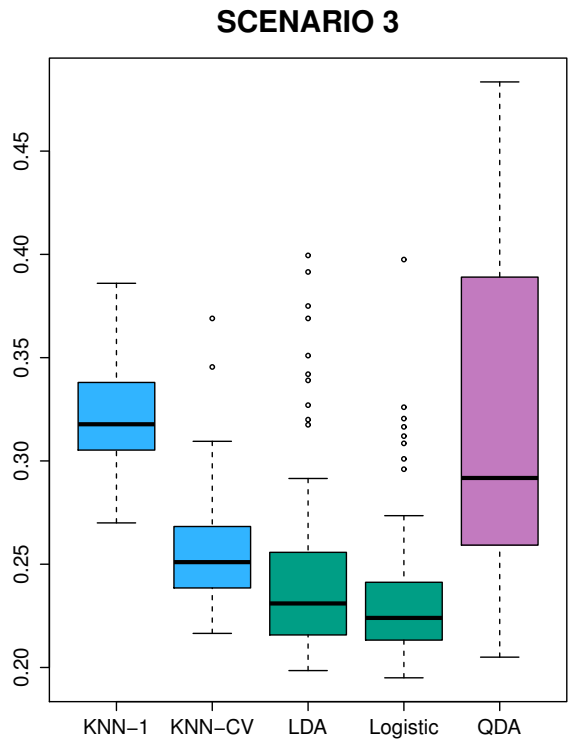
Scenario 3 X_1, X_2 from t-distribution (heavier tails than Gaussian), 50 observations per class. Decision boundary still linear, so assumptions of logistic regression satisfied, but those of LDA violated. Logistic regression outperforms LDA. QDA deteriorates considerably due to non-normality.

A Comparison of Classification Methods

Comparison scenarios







A Comparison of Classification Methods

Comparison scenarios

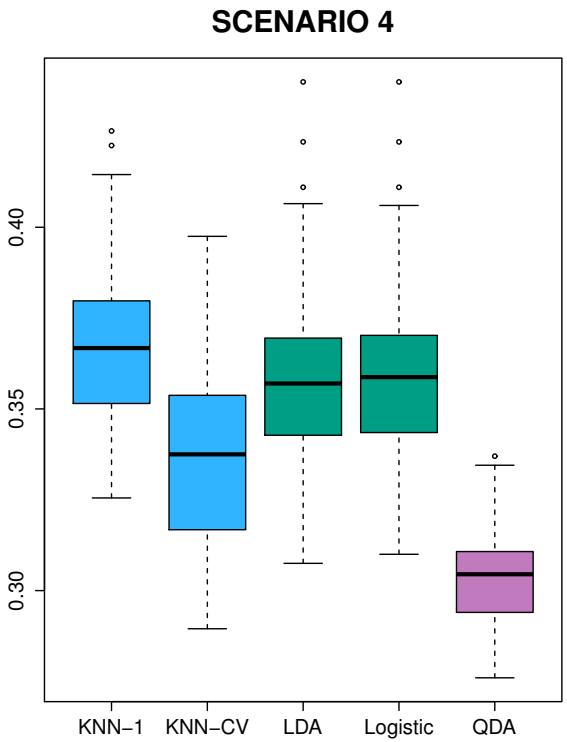
Scenario 4 Normal distribution, correlation 0.5 in the first, -0.5 in second class. Corresponds to QDA assumptions, quadratic decision boundaries. QDA outperforms all other methods.

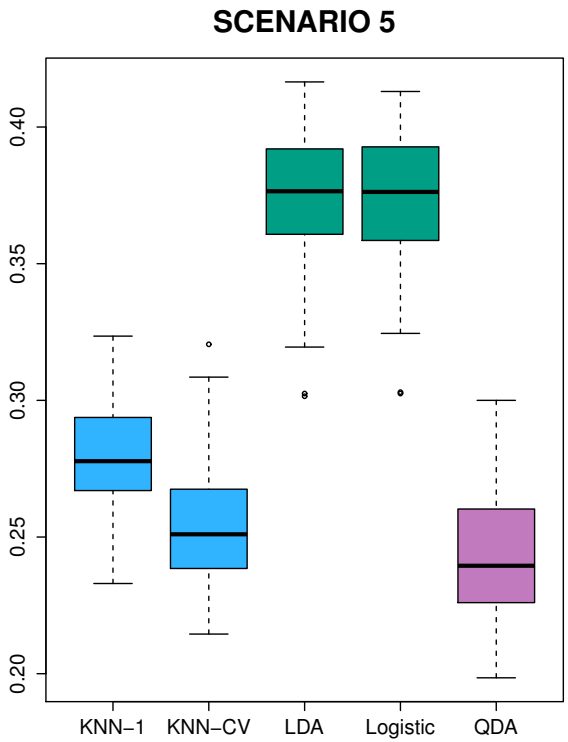
Scenario 5 In each class observations generated by normals with uncorrelated predictors, responses sampled from logistic function using X_1^2, X_2^2 and $X_1 X_2$ as predictors. Quadratic decision boundary. QDA best followed closely by KNN-CV. Linear methods perform poorly.

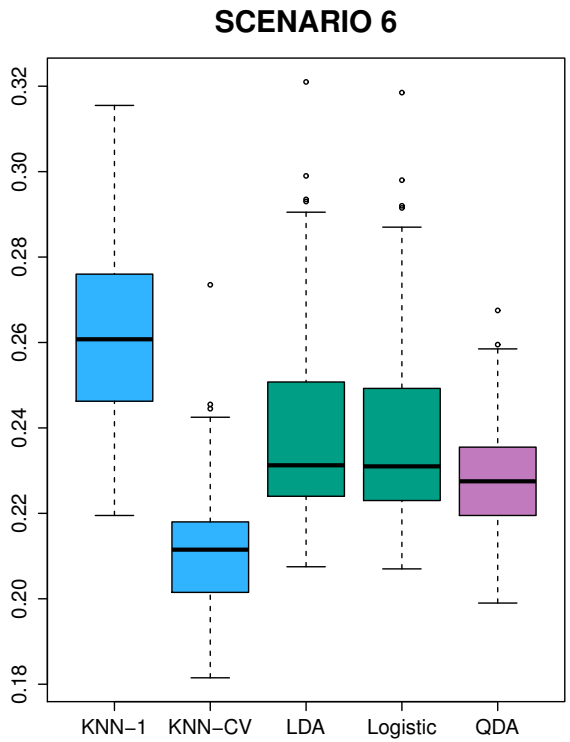
Scenario 6 Same as in 5, except now responses sampled from a more non-linear expression. Now even QDA can no longer correctly model complex decision boundary. QDA better than the linear methods, but more flexible KNN-CV gave best results. But: KNN with $K = 1$ gives worst results.

A Comparison of Classification Methods

Comparison scenarios







A Comparison of Classification Methods

Comparison scenarios

Summary:

- No method superior in all situations.
- Linear decision boundaries: LDA/logistic regression will perform well.
- Moderately nonlinear decision boundaries: QDA can be better.
- More highly nonlinear decision boundaries: high-variance method such as KNN may have advantages, but correct choice of smoothness (flexibility) parameter can be crucial.
- Next chapter: methods for finding the right amount of smoothing.