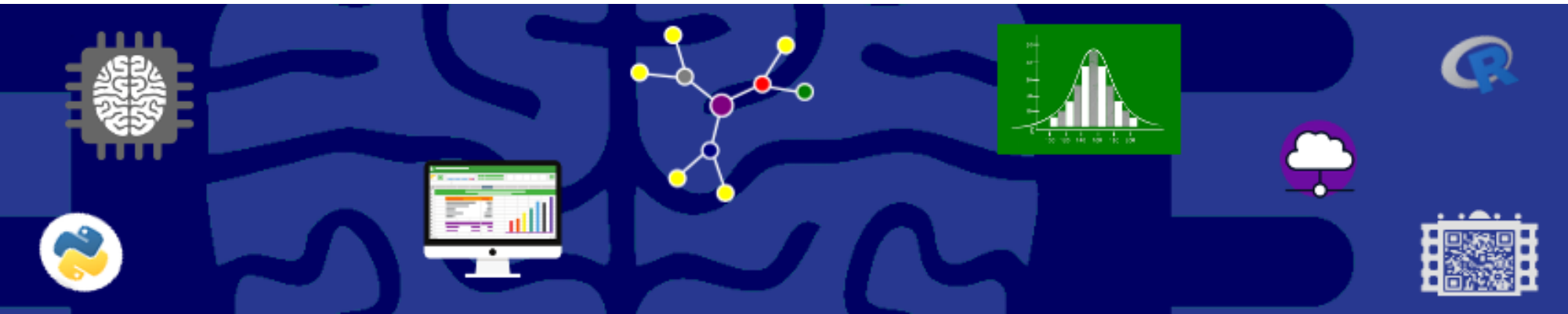
Introduction to Data Science

Winter Semester 2023/24

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: spam classification, fraud detection, tumor diagnostics, friend-foe distinction, medical triage ...
- 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.

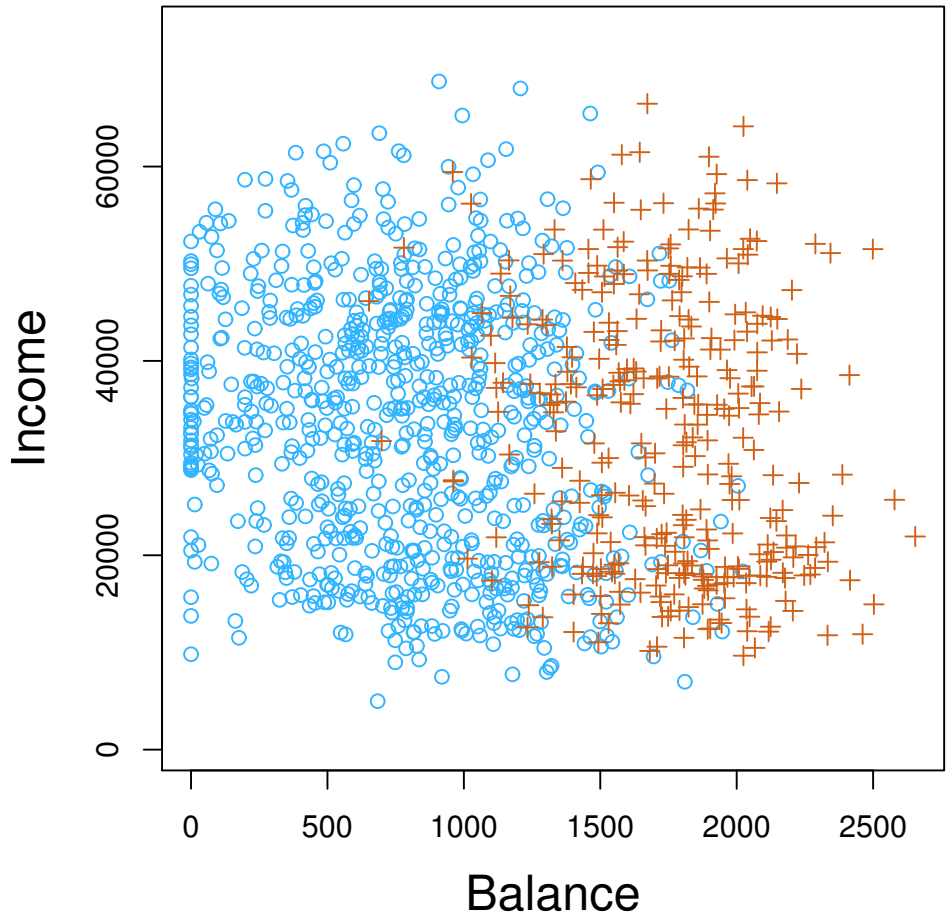
Overview of Classification

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'

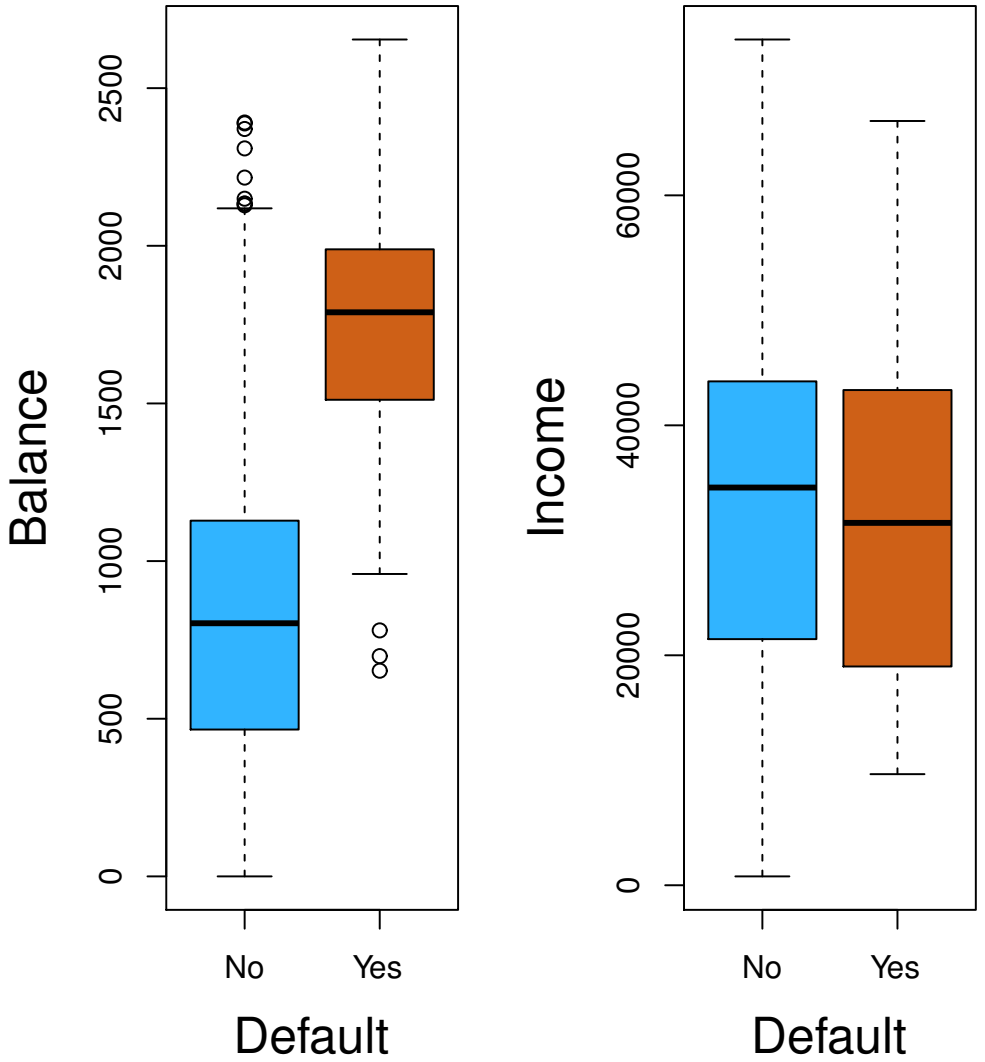
- annual `income` and
- monthly credit card `balance`.
- Response: binary variable `default` indicating whether or not a person defaulted on their credit card payment in a given month.
- Overall default rate: 3%.



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`.

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:

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

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 that 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*).
- But in general no way to map qualitative variable with > 2 values to quantitative response variable amenable to linear regression.

Why Not Linear Regression?

Ordering problem

- Sometimes underlying natural ordering exists (*mild, moderate, severe*).
- But 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 reverse coding gives same results.)

Why Not Linear Regression?

Ordering problem

- Sometimes underlying natural ordering exists (*mild, moderate, severe*).
- But 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 reverse coding gives same results.)
Interpret linear regression model response $X\hat{\beta}$ as estimate of probability

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

Why Not Linear Regression?

Ordering problem

- Sometimes underlying natural ordering exists (*mild, moderate, severe*).
- But 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 reverse coding gives same results.)
Interpret linear regression model response $X\hat{\beta}$ as 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

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

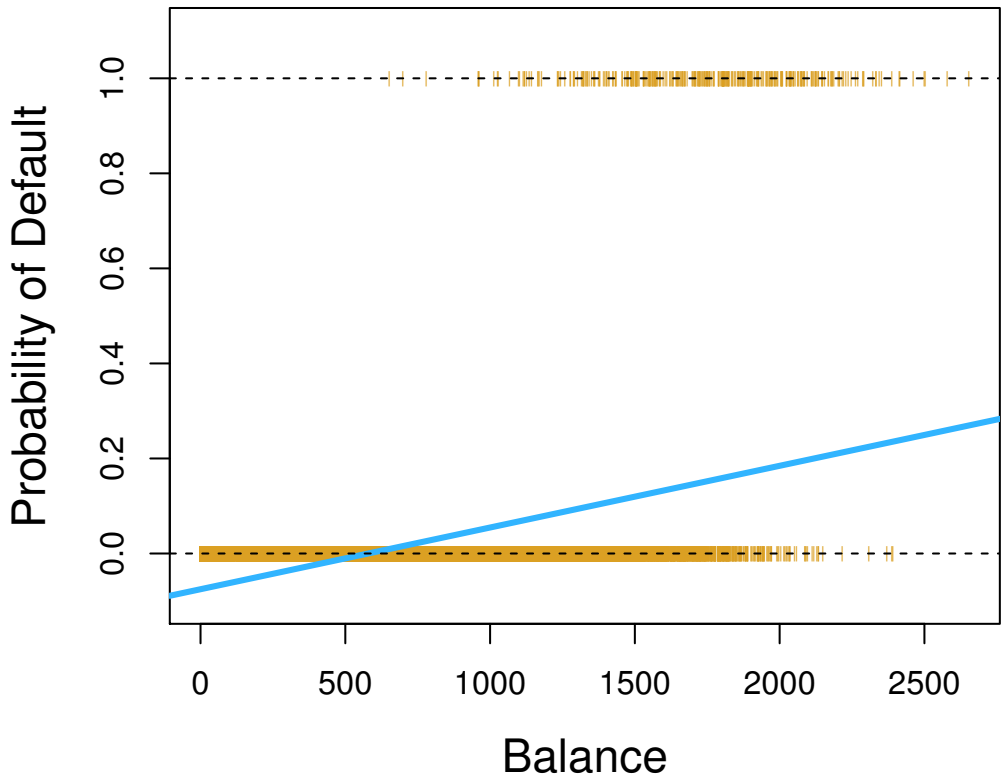
- Predict `default` $=$ `Yes` whenever, e.g., $p($`balance`$) > 0.5$.
- More conservative credit card company might prefer lower threshold, e.g., $p($`balance`$) > 0.1$.

- 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.

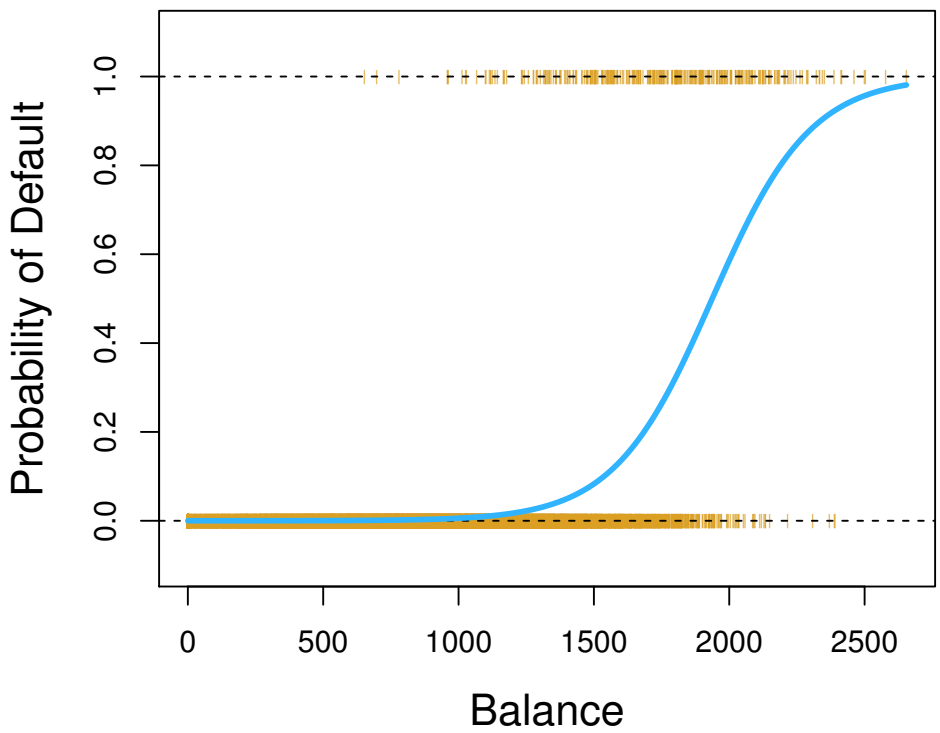
- Illustrates basic problem of fitting binary response coded with $\{0, 1\}$ with straight line: unless range of X limited, can always obtain probabilities outside $[0, 1]$.



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.[6]

[6] More specifically, these are known as *decimal* (European) odds, not to be confused with *fractional* (British) or money-line (American) odds.

- 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.[6]
- 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.

[6] More specifically, these are known as *decimal* (European) odds, not to be confused with *fractional* (British) or money-line (American) odds.

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 $\beta_1 < 0$.

Logistic Regression

Estimating the regression coefficients

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

- **Maximum-likelihood estimation** (MLE) to determine estimates $\hat{\beta}_0, \hat{\beta}_1$ of coefficients β_0, β_1.
- **Likelihood function**: probability of observing data $\{(x_i, y_i)_{i=1}^n\}$, $y_i \in \{0, 1\}$, if observations independent given values of β_0, β_1:

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

- Estimates determined as $(\hat{\beta}_0, \hat{\beta}_1) := \arg\max L(\beta_0, \beta_1)$.
  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

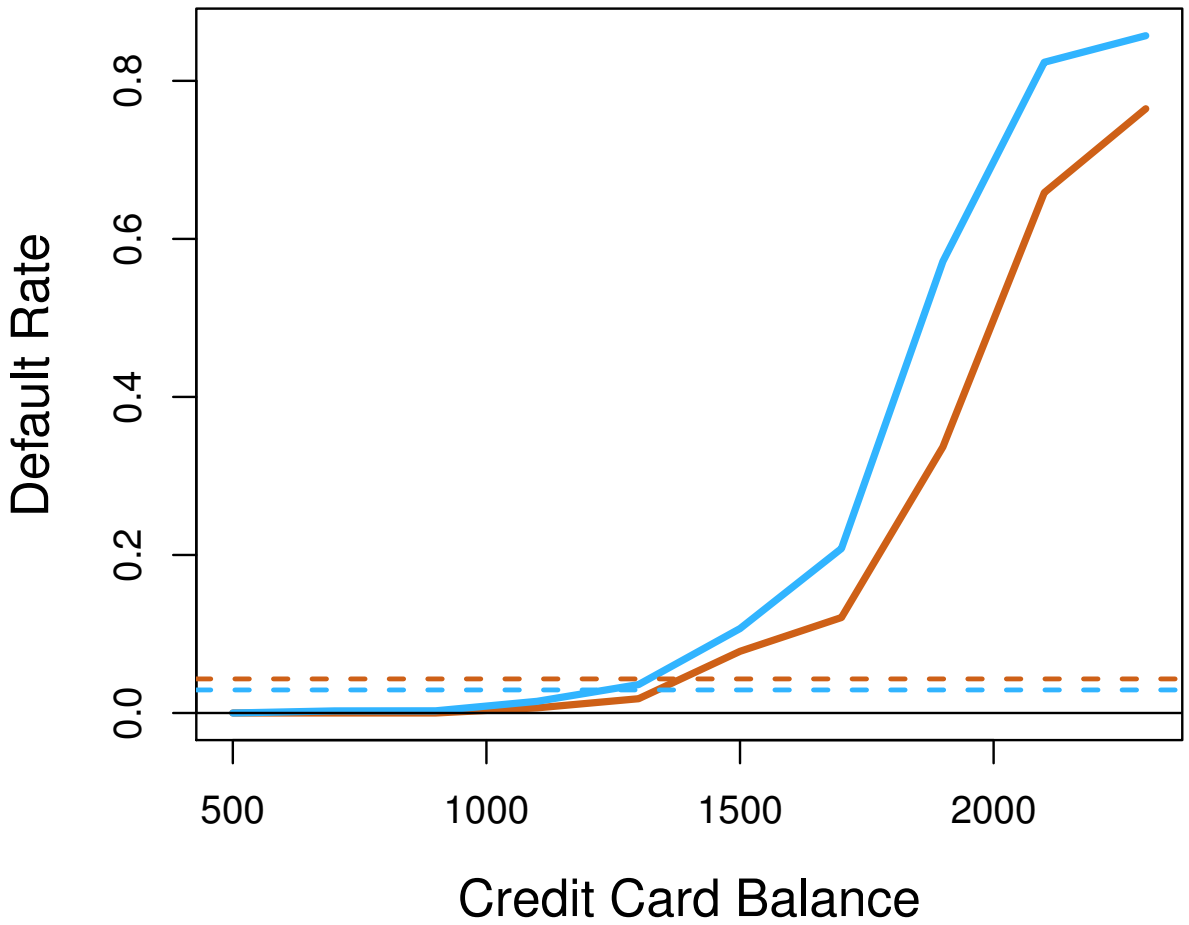
Multiple logistic regression

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

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.

Logistic Regression

Multiple logistic regression

- 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`.

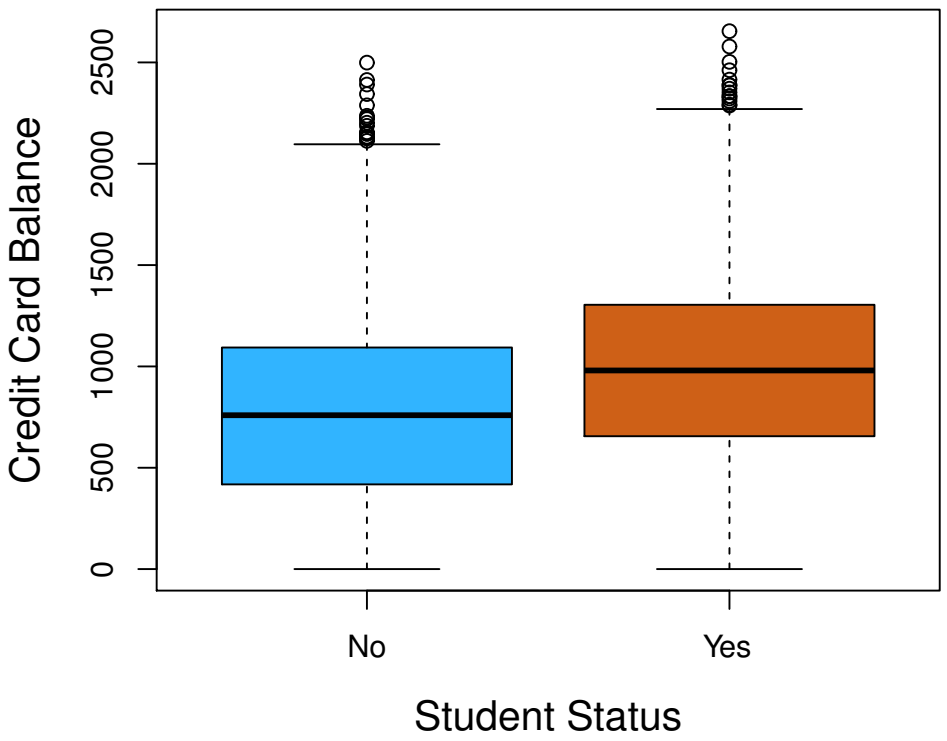
Logistic Regression

Multiple logistic regression

- 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

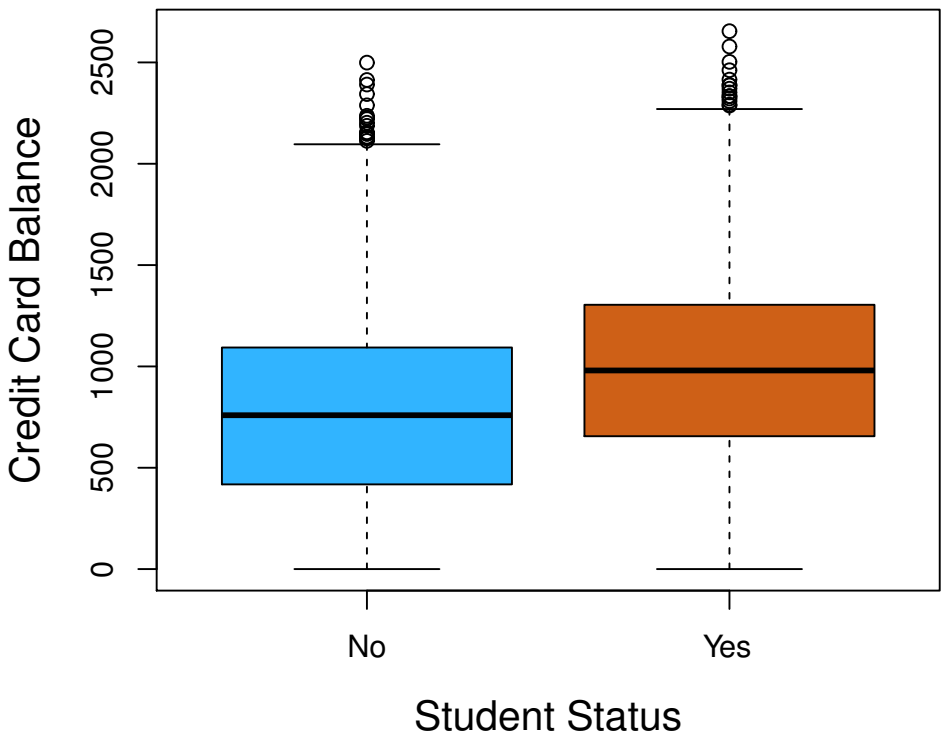
Multiple logistic regression



- 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.

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**.

Logistic Regression

Multiple logistic regression: example predictions

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

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

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

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

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: **multinomial logistic regression**.
- Software available, but LDA more popular for this case.

For $K > 2$ classes, select one class as **baseline**, say, class K. Replace (4.7) by

$$\mathbf{P}(Y = k|X = x) \approx p_k(x) := \frac{\mathrm{e}^{\beta_0^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_p^{(k)}x_p}}{1+\sum_{\ell=1}^{K-1}\mathrm{e}^{\beta_0^{(\ell)}+\beta_1^{(\ell)}x_1+\cdots+\beta_p^{(\ell)}x_p}}, \quad k = 1,\ldots,K-1,$$

and

$$\mathbf{P}(Y = K|X = x) \approx p_K(k) := \frac{1}{1+\sum_{\ell=1}^{K-1}\mathrm{e}^{\beta_0^{(\ell)}+\beta_1^{(\ell)}x_1+\cdots+\beta_p^{(\ell)}x_p}}.$$

- Analogously to (4.7), we have

$$\log\frac{p_k(x)}{p_K(x)} = \beta_0^{(k)} + \beta_1^{(k)}x_1 + \cdots + \beta_p^{(k)}x_p,$$

  i.e., log-odds between pairs of features linear in predictor variables.
- Log-odds between class pairs, predictions independent of choice of baseline, interpretation of coefficients changes.

Logistic Regression

Softmax encoding

Alternative (but equivalent) encoding: no baseline, treat all classes symmetrically using

$$\mathbf{P}(Y = k|X = x) \approx \tilde{p}_k(x) := \frac{\mathrm{e}^{\beta_0^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_p^{(k)}x_p}}{\sum_{\ell=1}^{K} \mathrm{e}^{\beta_0^{(\ell)}+\beta_1^{(\ell)}x_1+\cdots+\beta_p^{(\ell)}x_p}}.$$

- Requires estimation of K sets of coefficients instead of $K-1$.
- Log-odds between class pairs given by

$$\log \frac{\tilde{p}_k(x)}{\tilde{p}_j(x)} = (\beta_0^{(k)} - \beta_0^{(j)}) + (\beta_1^{(k)} - \beta_1^{(j)})x_1 + \cdots + (\beta_p^{(k)} - \beta_p^{(j)})x_p.$$

Contents

Generative Models for Classification

Conditional distribution

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

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

Generative Models for Classification

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 $K = 2$.

Generative Models for Classification

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 $\mathbf{P}(Y = k | X = x_0)$.

Generative Models for Classification

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 and predictors approximately Gaussian across classes, LDA more stable than logistic regression.
  - LDA popular for $K > 2$.

Generative Models for Classification

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

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

Generative Models for Classification

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

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

Generative Models for Classification

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**.

Generative Models for Classification

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.

Generative Models for Classification

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:

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

Generative Models for Classification

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**[7] of X in class k. In other words: $f_k(x)$ large if probability that $X = x$ is large in class k.

[7] (probability mass function). Modify accordingly for non-discrete predictor variable.

Generative Models for Classification

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**[7] of X in class k. In other words: $f_k(x)$ large if probability that $X = x$ is large in class k.
- Bayes' rule: **posterior** probability given by

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

[7] (probability mass function). Modify accordingly for non-discrete predictor variable.

Generative Models for Classification

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**[7] of X in class k. In other words: $f_k(x)$ large if probability that $X = x$ is large in class k.
- Bayes' rule: **posterior** probability given by

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

- **Idea:** instead of computing $p_k(X) = \mathbf{P}(Y = k|X)$ 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 (maximize $p_k(x)$ over k).

[7] (probability mass function). Modify accordingly for non-discrete predictor variable.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$

- **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.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$

Example:

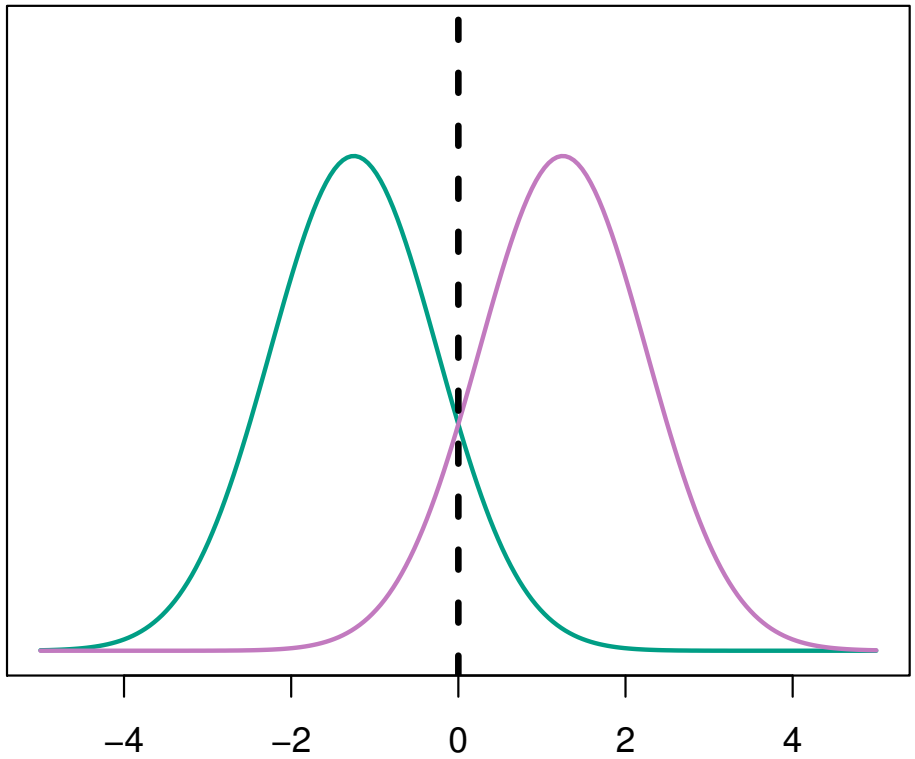
- $K = 2$, $\pi_1 = \pi_2$, assign x to class 1 if $\delta_1(x) > \delta_2(x)$ or

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

- 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).

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, estimating mean and variances

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

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

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

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, estimating mean and variances

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

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

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

- Prior probabilities estimated as

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

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, estimating mean and variances

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

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

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

- Prior probabilities estimated as

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

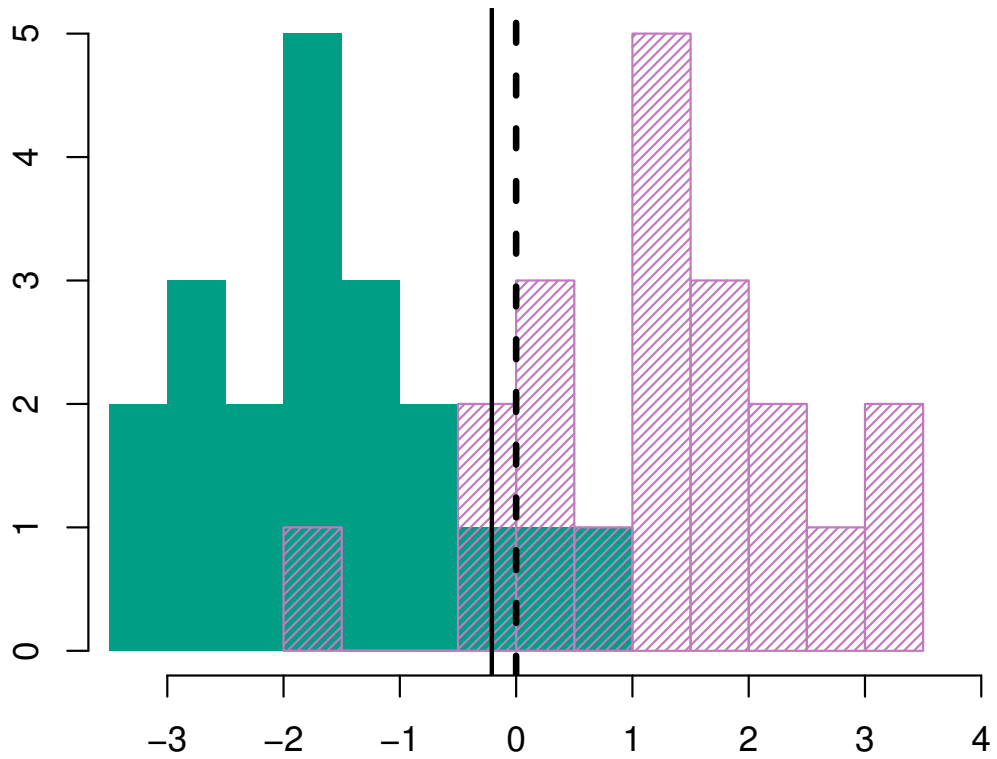
- Classifier now assigns new observation x to class k such that

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

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

Example: (right)

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



Simulated data from 2 classes (histograms), Decision boundaries: LDA solid, Bayes dashed.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$

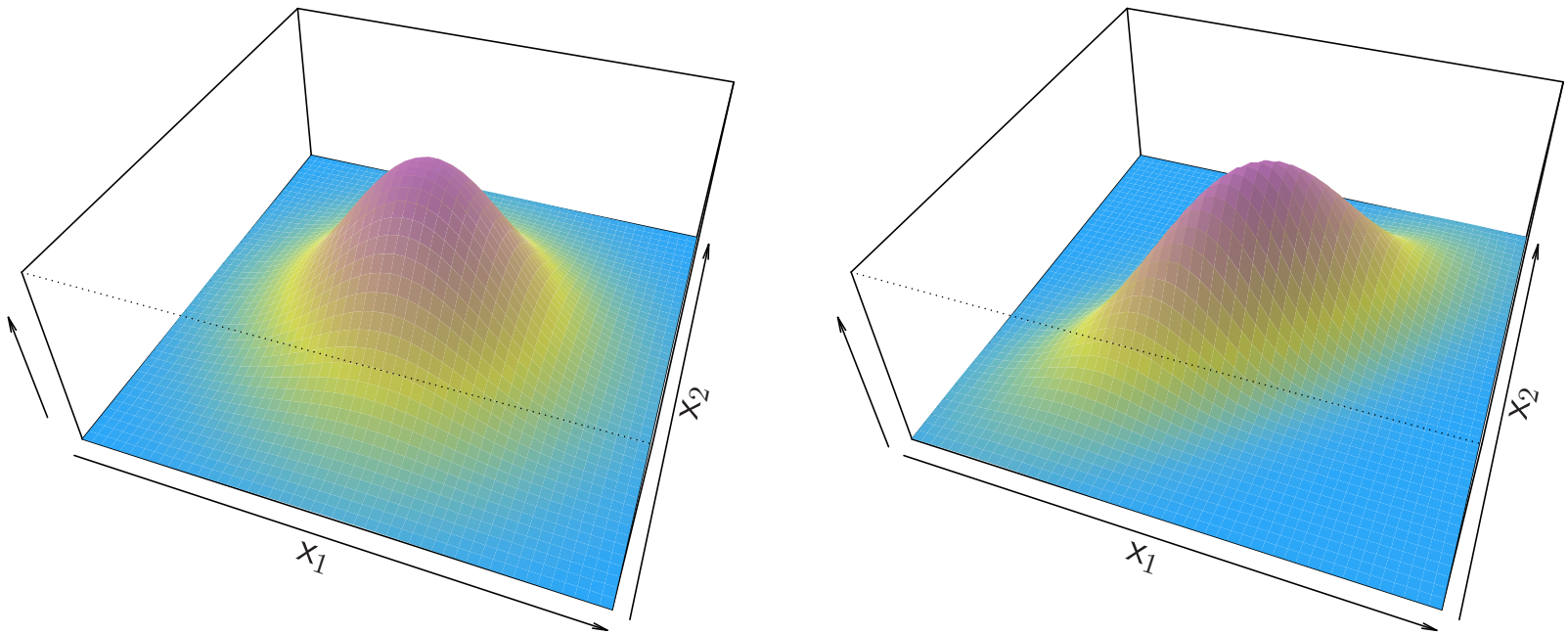
Recap: LDA classifier

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

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$

- 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.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$

- 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:

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

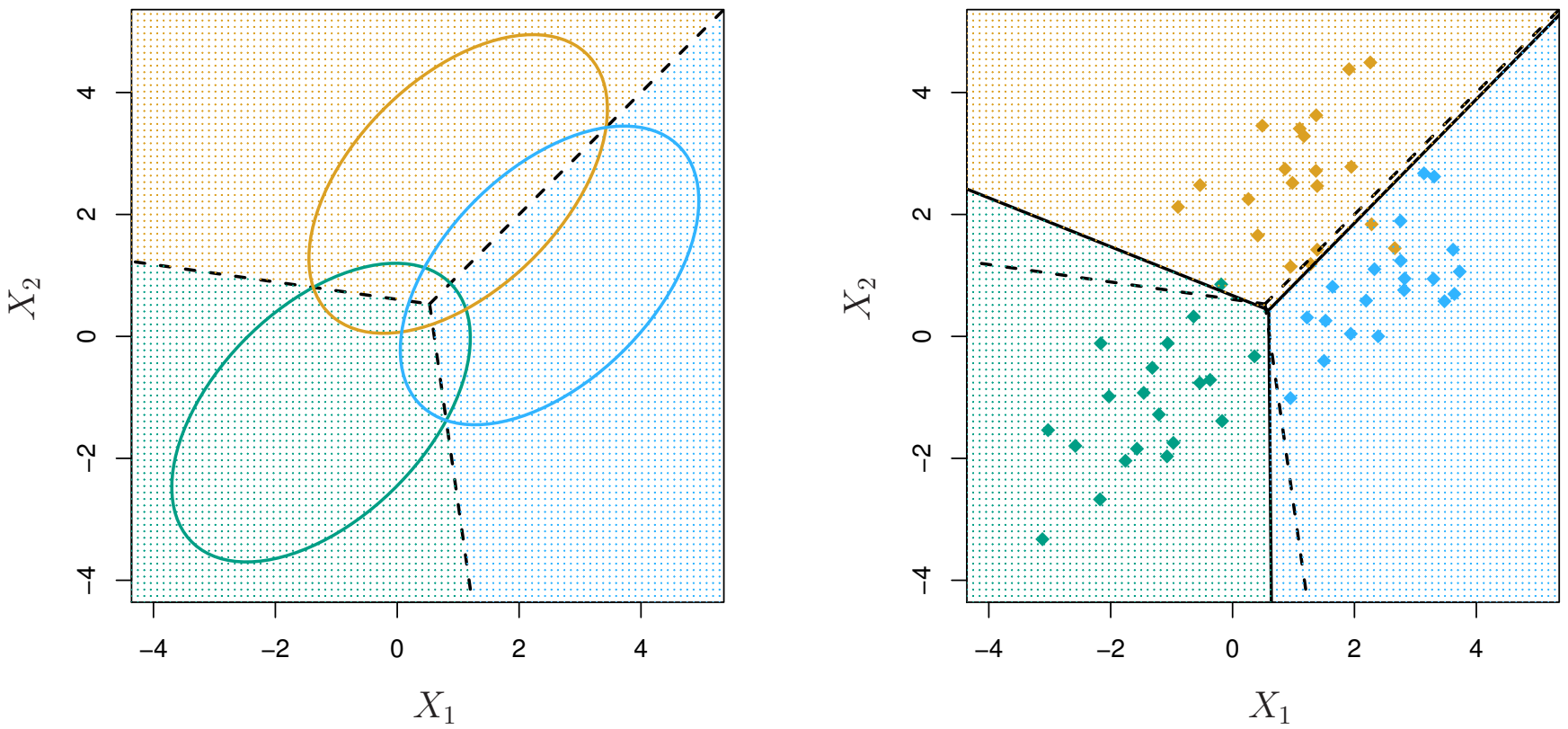
- LDA: for $p > 1$ assume within each class k: $X \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.
- Inserting pdf f_k into (4.8), we obtain Bayes classifier assigning observation $\boldsymbol{x}$ to class

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

This is the vector version of (4.12).

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, example



$p = 2$, $K = 3$, samples from 3 bivariate 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.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, example

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

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, 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.

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, 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

Generative Models for Classification

Linear Discriminant Analysis, $p = 1$, 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 $\boldsymbol{x}$.

Generative Models for Classification

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?

Generative Models for Classification

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 p to n poses danger of overfitting, but here $p = 2$, $n = 10\,000$.
  - Overall, true default rate in `Default` training data only 3.33%. Implies (useless) constant classifier $Y \equiv 0$ has this low error rate.
- Binary classification: Which errors are more likely, more consequential?

Generative Models for Classification

LDA applied to `Default` data set: classification error types

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 (0.24%) incorrectly labelled.

Generative Models for Classification

LDA applied to `Default` data set: classification error types

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 (0.24%) incorrectly labelled.
- However: of 333 who really defaulted, 252 (75.7%) missed by LDA.

Generative Models for Classification

LDA applied to `Default` data set: classification error types

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 (0.24%) incorrectly labelled.
- However: of 333 who really defaulted, 252 (75.7%) missed by LDA.
- For credit card company trying to identify high-risk individuals, such a *false negative error rate* probably unacceptable.

Generative Models for Classification

Binary misclassification errors

- **Class-specific** classification errors can be crucial.
- In screening procedures (medicine, airport security):
  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).

  probability of committing type-I error: α
  probability of committing type-II error: β
  power of test: $1 - \beta$

Generative Models for Classification

Example for binary classification error: mammography screening

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

Generative Models for Classification

Example for binary classification error: 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%

Generative Models for Classification

Example for binary classification error: 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?)

Generative Models for Classification

Example for binary classification error: 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?)

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

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

Generative Models for Classification

Example for binary classification error: 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]

Generative Models for Classification

LDA: modified threshold

- 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 216)

		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

Generative Models for Classification

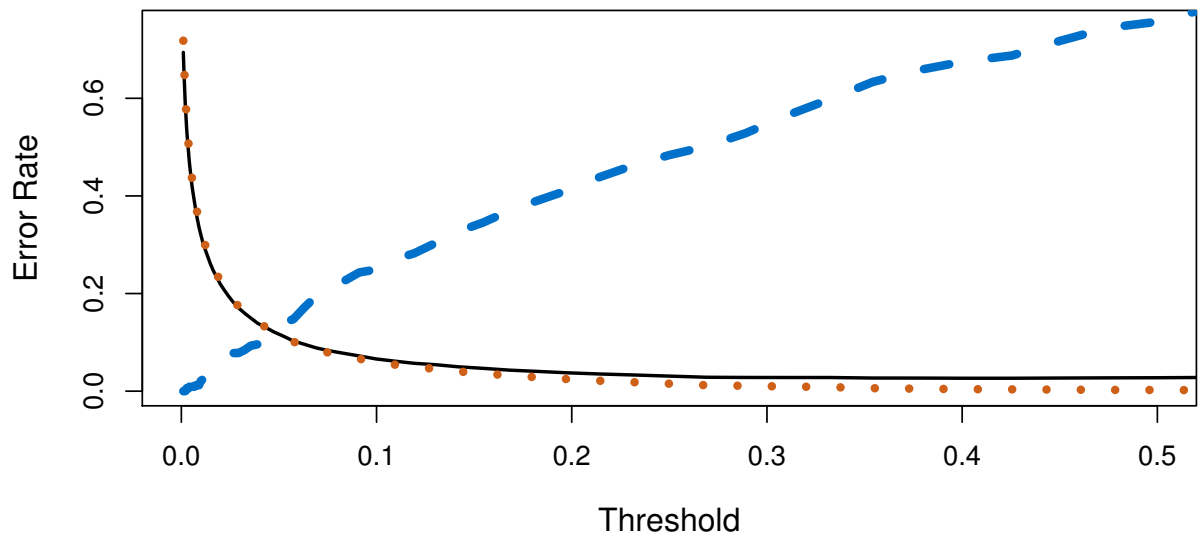
LDA: modified threshold

- 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 $(138 + 235)/10\,000 = 3.73\%$.

Generative Models for Classification

LDA: modified threshold

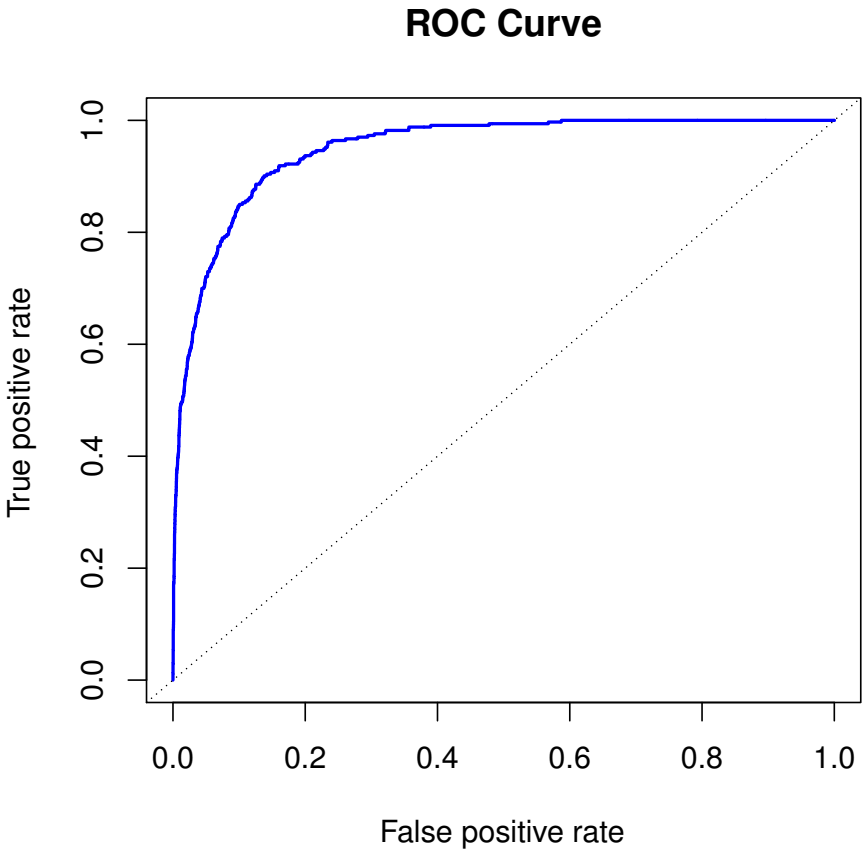
- LDA default prediction increases from 104 to 430. Default prediction error rate improves from $252/333 \approx 75.7\%$ to $138/333 \approx 41.4\%$.
- However, now 235 individuals who did not default are misclassified, raising the classification error in this class from $23/9\,667 \approx 0.24\%$ to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

Generative Models for Classification

Classification error: ROC curve



- Traces out false positive/true positive rate for all threshold values of LDA classifier in `Default` data set.
- True positive: sensitivity (ratio of defaulters correctly classified)
- False positive: 1−specificity (ratio of non-defaulters incorrectly classified).
- Optimal ROC curve: follows left/top boundaries. (top?)
- 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.

Generative Models for Classification

Classification error: 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.

Generative Models for Classification

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 $\boldsymbol{x}$.

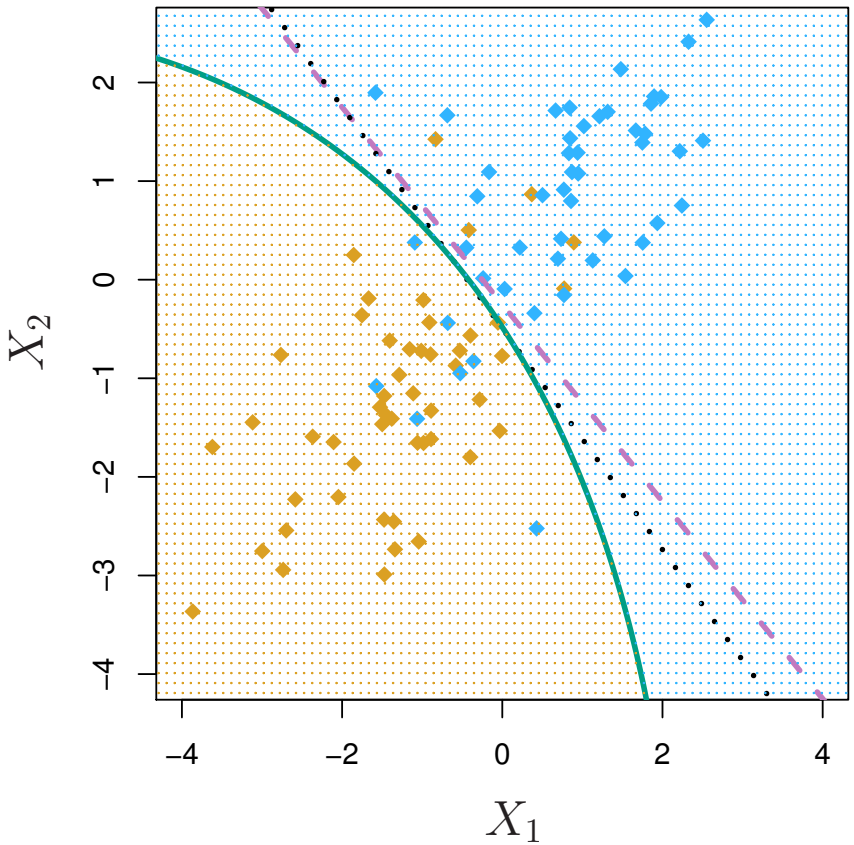
Generative Models for Classification

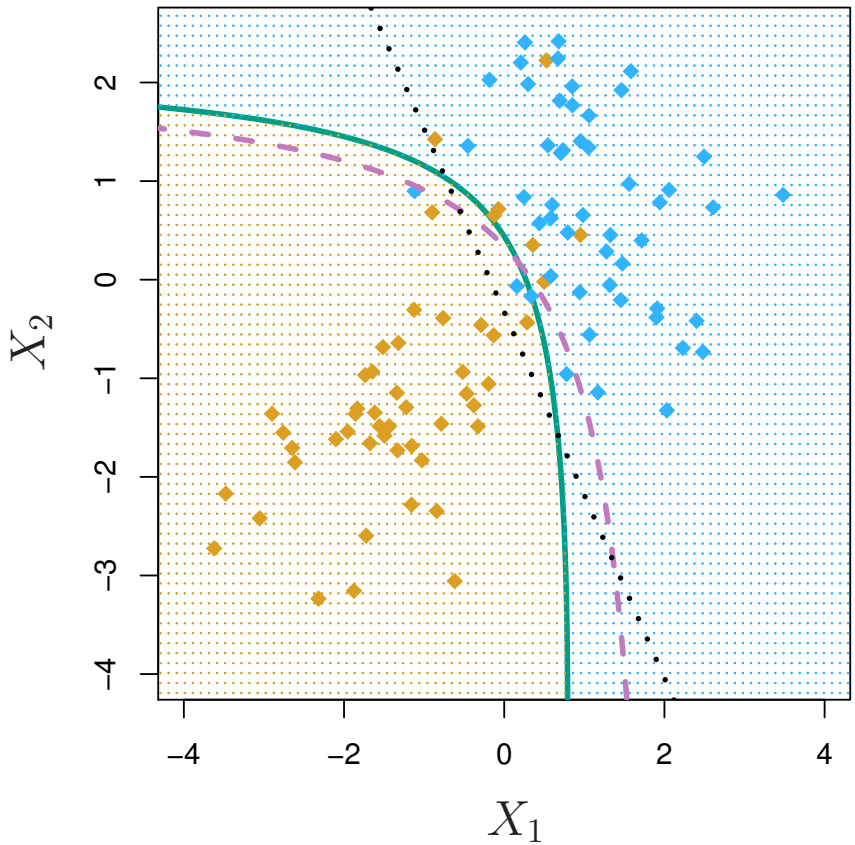
Quadratic discriminant analysis

- 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.

Generative Models for Classification

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:$\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

Generative Models for Classification

Naive Bayes

- Bayes' theorem in the form (4.8) gave a basis for approximating the conditional probabilities $p_k(x) = \mathbf{P}(Y = k|X = x)$ (posterior probabilities).
- Priors π_k easy to estimate from data (relative frequencies); within-class densities $f_k(x)$ more challenging.
- For LDA/QDA: strong assumption that within-class distributions jointly Gaussian with class-specific means and common/class-specific covariance matrix. Reduces problem of estimating K densities $f_k(x)$ to that of estimating K mean vectors $\mu_k \in \mathbb{R}^p$ and one or K covariance matrices in $\mathbb{R}^{p\times p}$, respectively.
- Assumption for **naive Bayes classifier**: within each class, the p predictors are independent, i.e.,

$$f_k(x) = f_{k,1}(x_1) \cdot f_{k,2}(x_2) \cdot \ \ldots \ \cdot f_{k,p}(x_p), \qquad k = 1, \ldots, K,$$

$f_{k,j}$: density function of j-th predictor variable among observations belonging to class k.

Generative Models for Classification

Naive Bayes

- No need to estimate correlations/dependencies between predictor variables.
- Convenient when n not large relative to p.
- Introduces some bias, reduces variance.
- Plug into (4.8)

$$\mathbf{P}(Y = k|X = x) \approx \frac{\pi_k f_{k,1}(x_1) \cdot f_{k,2}(x_2) \cdot \ \ldots \ \cdot f_{k,p}(x_p)}{\sum_{\ell=1}^{K} \pi_\ell f_{\ell,1}(x_1) \cdot f_{\ell,2}(x_2) \cdot \ \ldots \ \cdot f_{\ell,p}(x_p)}, \qquad (4.16)$$

$$k = 1, \ldots, K.$$

Generative Models for Classification

Naive Bayes

Options for estimating univariate densities $f_{k,j}$ from training data $x_{j,1}, \ldots x_{j,n}$:

- For quantitative X_j assume $X_j | Y = k \sim \mathrm{N}(\mu_{k,j}, \sigma^2_{k,j})$.
  Corresponds to QDA with diagonal within-class covariance matrices.
- For quantitative X_j, estimate $f_{k,j}$ in non-parametric way: histogram of j-th predictor in class k; $f_{k,j}(x_j) \approx$ fraction of class k observations in same histogram bin as x_j.
  (Or **kernel density estimator**, smoothed version of histogram).
- For qualitative X_j: count relative frequency of values of X_j in each class.

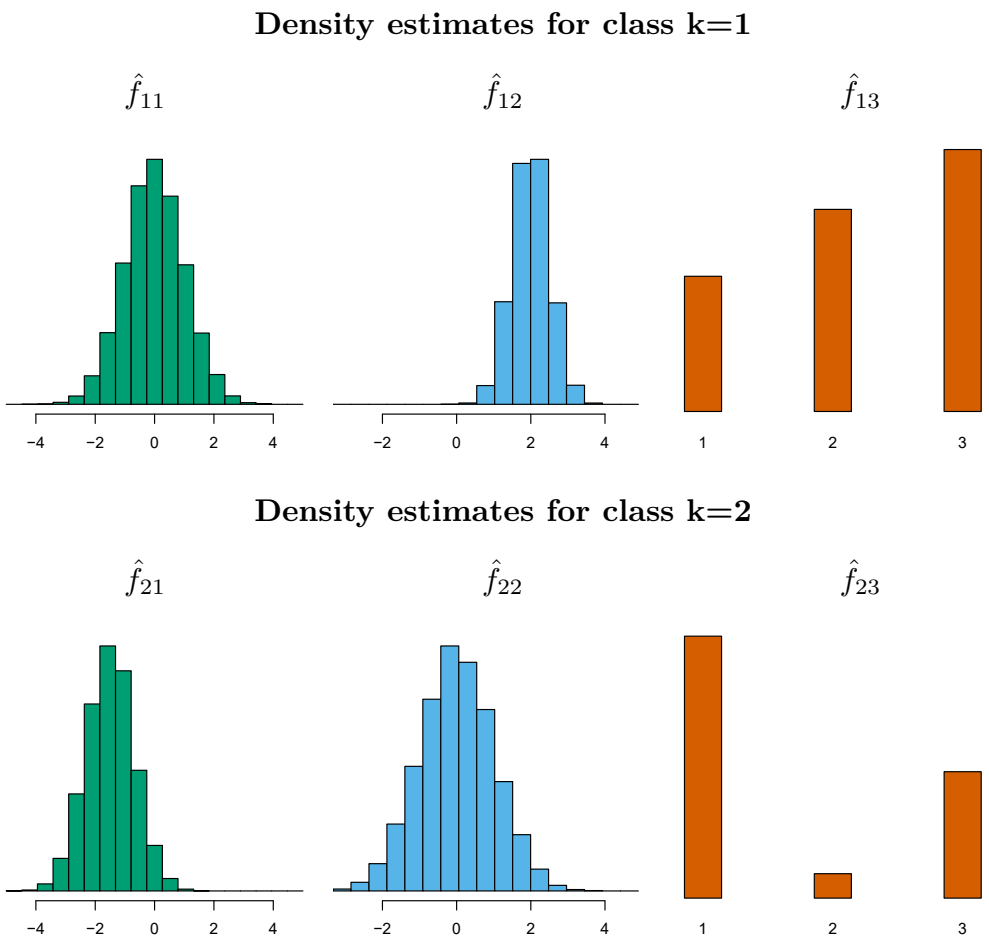
  Example: $X_j \in \{1, 2, 3\}$, $n = 100$ observations in class k; if number of times X_j takes on values 1,2,3 are 32, 55, 13, respectively, then estimate

$$f_{k,j}(x_j) = \begin{cases} 0.32 & x_j = 1, \\ 0.55 & x_j = 2, \\ 0.13 & x_j = 3. \end{cases}$$

Generative Models for Classification

Naive Bayes: toy example

- $p = 3$ predictors for $K = 2$ classes.
- X_1, X_2 quantitative, X_3 qualitative with 3 levels.
- $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$.



Generative Models for Classification

Naive Bayes: toy example

Now classify new observation $x^* = (0.4, 1.5, 1)^\top$.
For the data shown, we obtain

$$\hat{f}_{1,1}(0.4) = 0.368, \qquad \hat{f}_{1,2}(1.5) = 0.484, \qquad \hat{f}_{1,3}(1) = 0.226,$$
$$\hat{f}_{2,1}(0.4) = 0.030, \qquad \hat{f}_{2,2}(1.5) = 0.130, \qquad \hat{f}_{2,3}(1) = 0.616.$$

Plug into (4.16) to obtain posterior probability estimates

$$\mathbf{P}(Y = 1|X = x^*) = 0.944, \qquad \mathbf{P}(Y = 2|X = x^*) = 0.056.$$

Generative Models for Classification

Naive Bayes: `Default` data set

Naive Bayes applied to `Default` data set: predict `default` from `balance` and `student status` (cf. Slide 216).

- threshold 0.5;
- quantitative predictor Gaussian, predictors independent within classes.

		True default status		
		No	Yes	Total
Predicted default status	No	9 621	244	9 865
	Yes	46	89	135
	Total	9 667	333	10 000

Observations:

- LDA has slightly lower overall error rate (275 vs. 290 per 10 000);
- naive Bayes correctly predicts higher fraction of true defaulters (89 vs. 81 of 333).

Generative Models for Classification

Naive Bayes: `Default` data set

Lower threshold to 0.2:

		True default status		
		No	Yes	Total
Predicted default status	No	9 339	130	9 865
	Yes	328	203	135
	Total	9 667	333	10 000

Observations:

- Naive Bayes again has higher overall error rate than LDA with same threshold (458 vs. 373 per 10 000), but correctly predicts almost 2/3 (203 of 333) of true defaults.
- Since $n = 10\,000$ and $p = 2$ here, variance reduction afforded by naive Bayes not necessarily worthwhile.
- Expect greater payoff relative to LDA/QDA for smaller n / larger p.

Contents

A Comparison of Classification Methods

Analytical comparison: recall KNN

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.

A Comparison of Classification Methods

Analytical comparison: KNN vs. QDA

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 (cf. table on Slide 192).

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

Analytical comparison

We compare the structural differences between logistic regression, LDA, QDA, naive Bayes and KNN by designating class K as a baseline and comparing the ratio

$$\log \frac{\mathbf{P}(Y = k|X = x)}{\mathbf{P}(Y = K|X = x)}, \qquad k = 1, \ldots, K,$$

for which the maximizing k is the classification result.

For LDA, we obtain an expression of the form

$$\log \frac{\mathbf{P}(Y = k|X = x)}{\mathbf{P}(Y = K|X = x)} = a_k + \sum_{j=1}^{p} b_{k,j}\, x_j \ .$$

For QDA, this becomes

$$\log \frac{\mathbf{P}(Y = k|X = x)}{\mathbf{P}(Y = K|X = x)} = a_k + \sum_{j=1}^{p} b_{k,j}\, x_j + \sum_{k,\ell=1}^{p} c_{k,j,\ell}\, x_j x_\ell \ .$$

For naive Bayes, the corresponding expression is

$$\log \frac{\mathbf{P}(Y = k|X = x)}{\mathbf{P}(Y = K|X = x)} = a_k + \sum_{j=1}^{p} g_{k,j}(x_j)$$

with nonlinear univariate functions $g_{k,j}$. Such a structure is known as a **generalized linear model**, to be discussed later.

A Comparison of Classification Methods

Analytical comparison

- LDA is seen to be a special case of QDA for which $c_{k,j\ell} \equiv 0$ for all k, j, ℓ.
- Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{k,j}(x_j) = b_{k,j}\, x_j$. Not obvious from derivation of, e.g., LDA and naive Bayes: LDA allows dependent predictors within class while naive Bayes assumes independence!
- If naive Bayes uses as $f_{k,j}$ the density of $\mathrm{N}(\mu_{k,j}, \sigma_j^2)$, then $g_{k,j}(x_j) = b_{k,j}\, x_j$ with $b_{k,j} = (\mu_{k,j} - \mu_{K,j})/\sigma_j^2$. This correspnds to a special case of LDA with $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^1, \ldots, \sigma_p^2)$.
- Neither QDA nor naive Bayes is a special case of the other. Naive Bayes flexible as allows arbitrary $g_{k,j}$, but structure always an additive model. (Function of x_j (only) added to function of x_ℓ (only) for $j \neq \ell$. By contrast, QDA includes multiplicative terms $c_{k,j\ell}\, x_j x_\ell$, hence potentially more accurate when interactions among predictors relevant for classification.

None of these methods uniformly dominates the others: choice of method will depend on true within-class predictor distribution as well as n vs. p (bias-variance trade-off).

A Comparison of Classification Methods

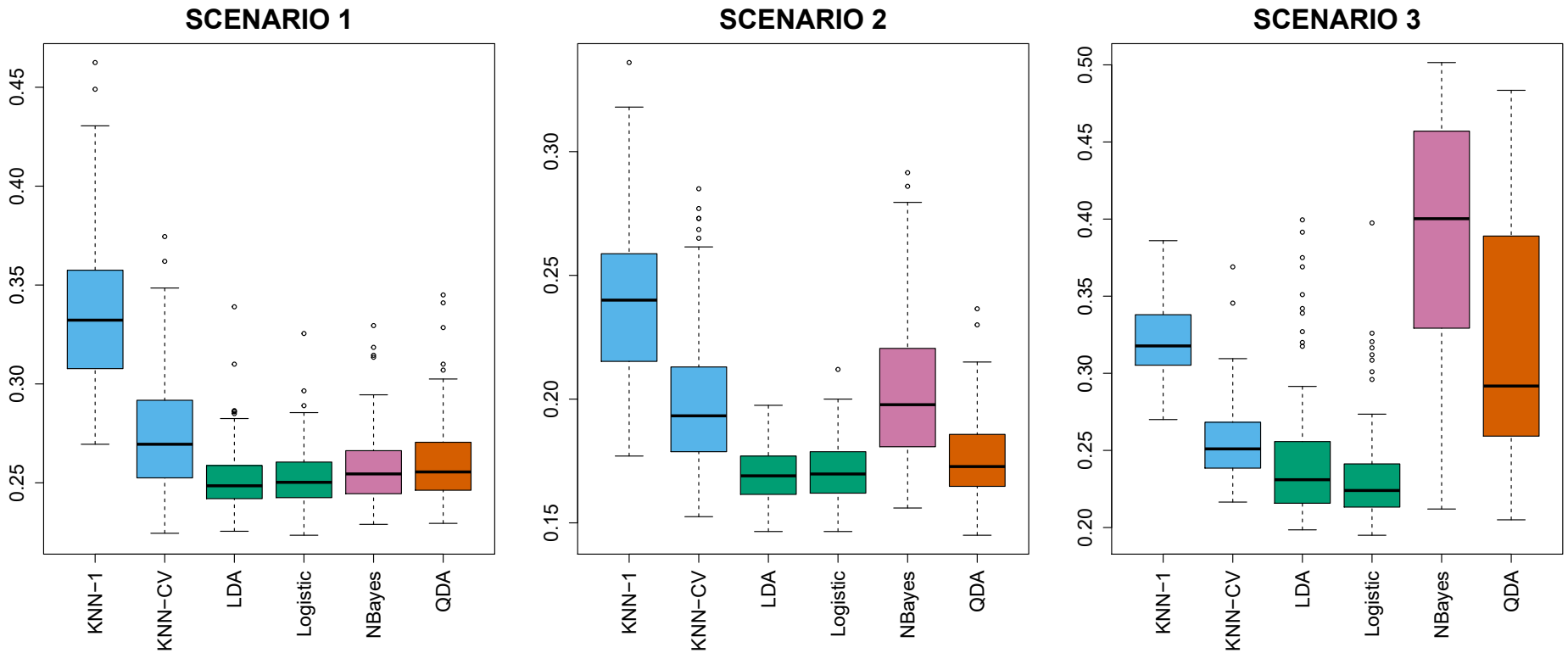
Empirical comparison

Six binary classification scenarios for comparison:
$p = 2$ (both quantitative); 3 linear, 3 nonlinear decision boundaries.
For each 100 random training data sets, larger simulated test data set.
For KNN used $K = 1$ and K determined 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's 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). Naive Bayes slightly better than QDA since independent predictors assumption correct.

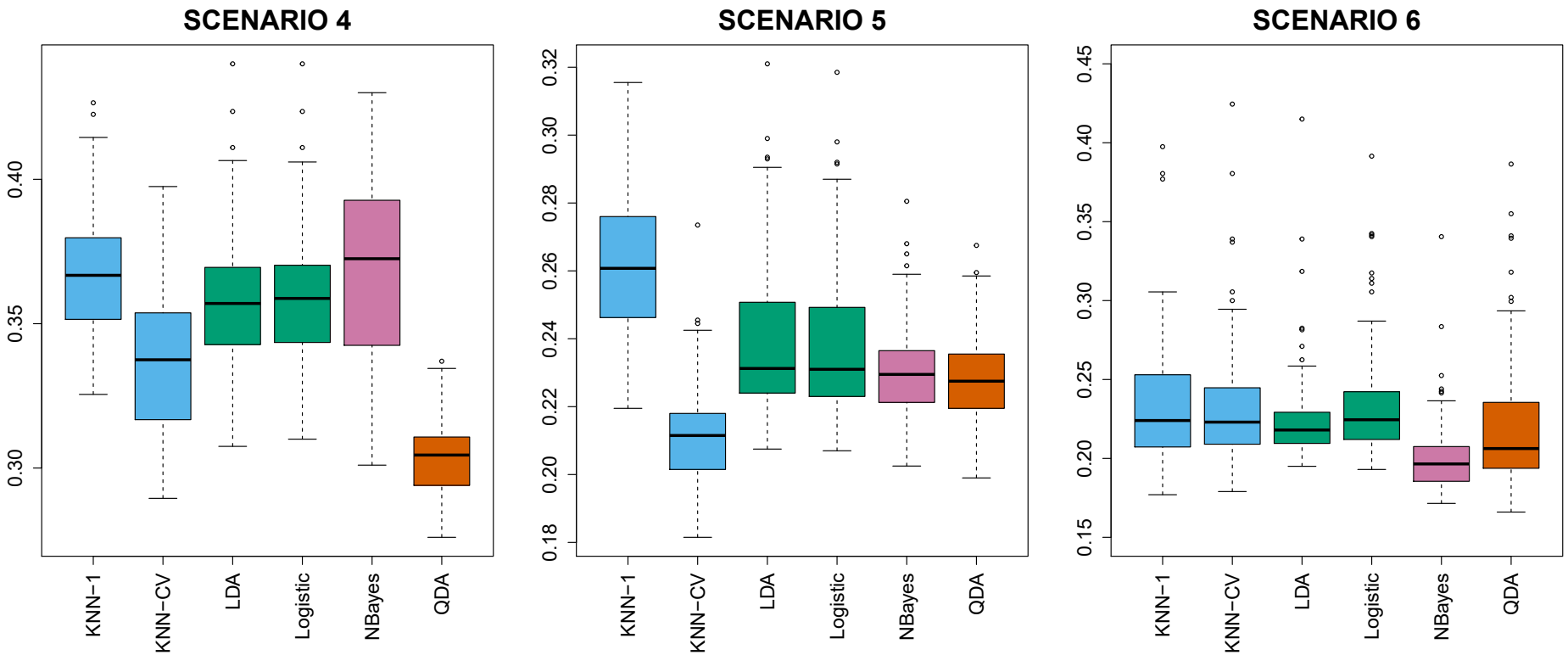
Scenario 2: Same as Scenario 1 except that within each class the 2 predictors had correlation -0.5. Little change, expect naive Bayes, which performs poorly since independent predictors assumption violated.

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/QDA violated. Logistic regression outperforms LDA. QDA deteriorates considerably due to non-normality. Naive Bayes again performs poorly due to lack of independence.



A Comparison of Classification Methods

Empirical comparison



Test error rates, nonlinear decision boundary.

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. Naive Bayes again performs poorly due to lack of independence.

Scenario 5: In each class observations generated by normals with uncorrelated predictors, responses sampled from logistic function applied to a complicated non-linear function of predictors. QDA, naive Bayes slightly better than linear methods, while much more flexible KNN-CV method gave best results. KNN with K = 1 gave worst results of all methods, highlights fact that even when the data exhibits a complex non-linear relationship, a non-parametric method such as KNN can still give poor results if level of smoothness not chosen correctly.

Scenario 6: Observations generated from normal distribution with different diagonal covariance matrix for each class, but sample size just n = 6 in each class. Naive Bayes performed very well, because its assumptions met. LDA, logistic regression perform poorly since true decision boundary is non-linear due to the unequal covariance matrices. QDA performed a bit worse than naive Bayes, because given small sample size, former incurred too much variance in estimating the correlation between the predictors within each class. KNN's performance also suffers due to the very small sample size.

A Comparison of Classification Methods

Empirical comparison

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.

Contents

Generalized Linear Models

The `bikeshare` data set

If the response variable Y is neither quantitative nor qualitative, then neither the regression nor the classification approaches discussed so far apply.

In `Bikeshare` data set

- response is `bikers`: number $\in \mathbb{N}_0$ of hourly users of a bike sharing program in Washington, DC (USA); neither qualitative nor quantitative.

Predictor variables:

- `mnth`: month of the year
- `hr`: hour of the day $\in \{0, 1, \ldots, 23\}$
- `workingday`: indicator variable for non-holiday and non-weekend
- `temp`: normalized temperature in $^\circ$C
- `weathersit`: $\in \{\text{clear}, \text{misty/cloudy}, \text{light rain/snow}, \text{heavy rain/snow}\}$

Generalized Linear Models

Linear regression on `Bikeshare` data set

Linear regression of `bikers` onto the predictors (baseline for `weathersit` was clear skies; `mnth` and `hr` on separate plot):

	Coefficient	Std. error	t-statistic	p-value
intercept	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	- 12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	- 66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15

Generalized Linear Models

Linear regression on `Bikeshare` data set

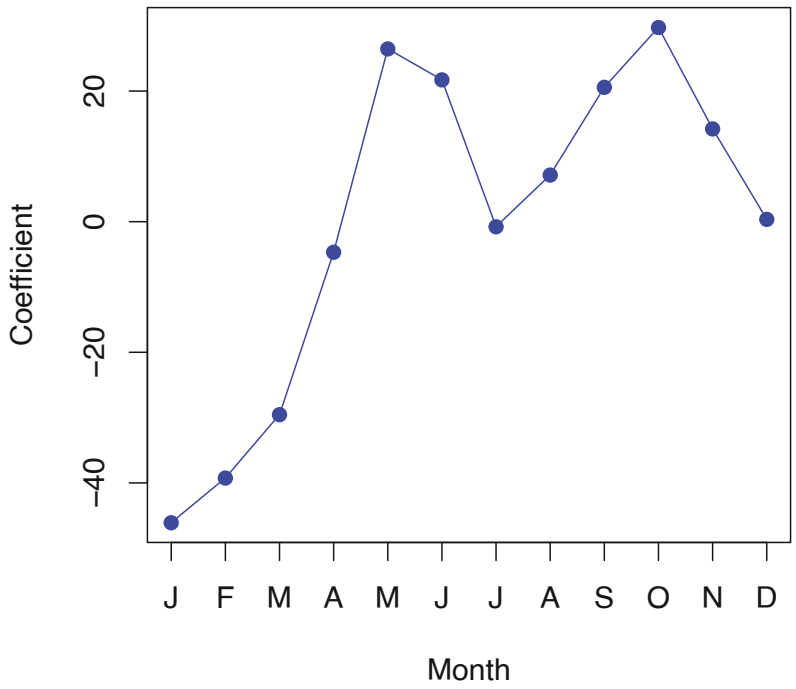
Linear regression of `bikers` onto the predictors (baseline for `weathersit` was clear skies; `mnth` and `hr` on separate plot):

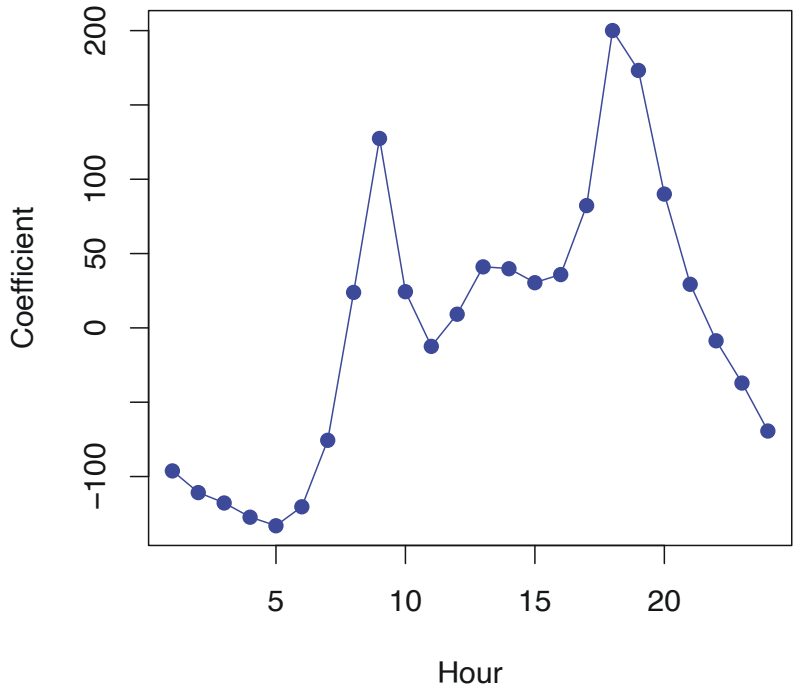
	Coefficient	Std. error	t-statistic	p-value
intercept	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	- 12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	- 66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15

- Change in `weathersit` from clear to cloudy results in, on average, 12.89 fewer bikers/hour.
- Change from cloudy to light rain/snow an additional 53.6 fewer bikers/hour.

Generalized Linear Models

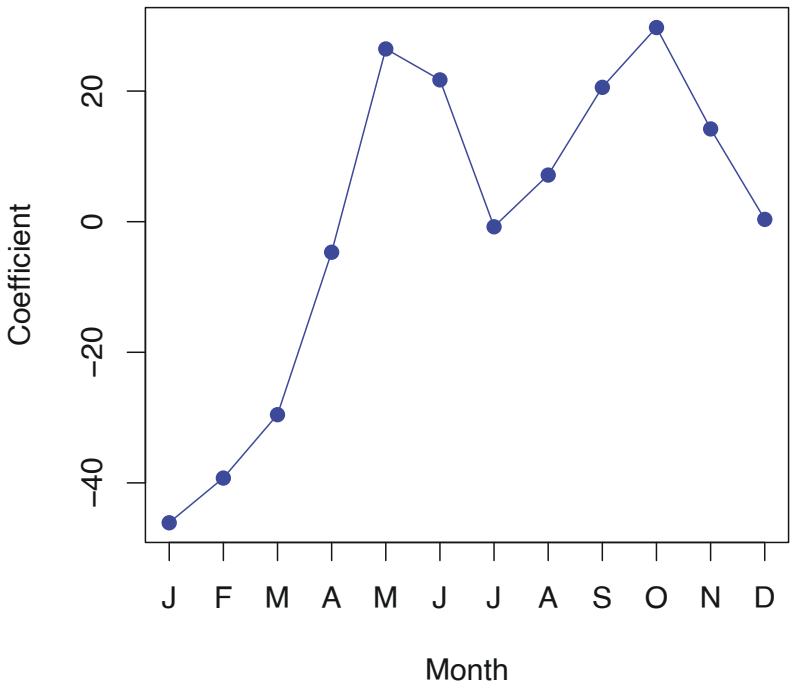
Linear regression on `Bikeshare` data set: `mnth` and `hr` predictors

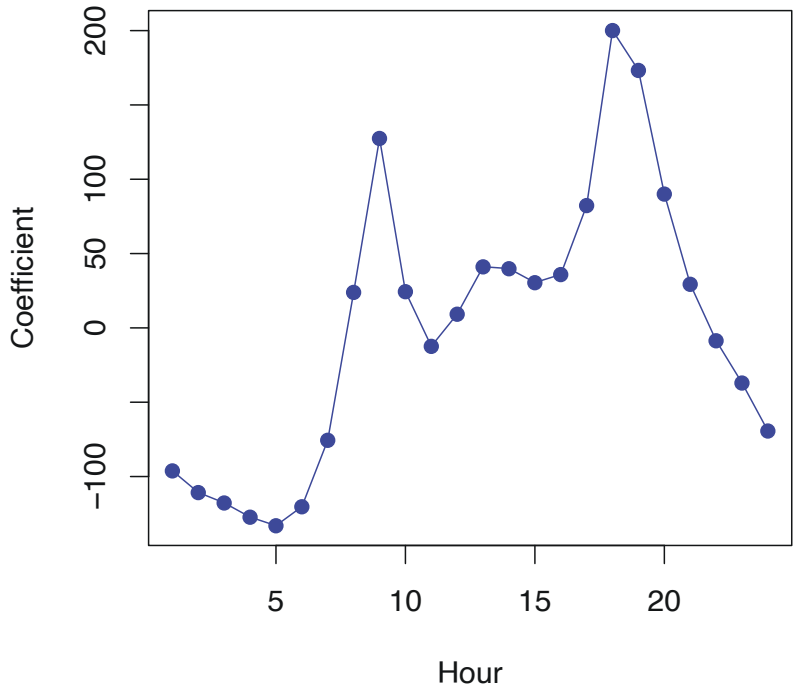




Generalized Linear Models

Linear regression on `Bikeshare` data set: `mnth` and `hr` predictors





- Bike usage high in spring/fall, lowest in winter months.
- Bike usage high during rush hour 9:00 – 18:00, lowest overnight.
- Plausible results up to here.

Generalized Linear Models

Linear regression on `Bikeshare` data set: criticism

Issue: 9.6% of fitted values in `Bikeshare` data set negative, i.e., linear regression model predicts negative number of users during 9.6% of hours in data set.

- Ability to perform meaningful predictions from the data?
- Accuracy of coefficient estimates, confidence intervals, and other outputs of regression model?

Generalized Linear Models

Linear regression on `Bikeshare` data set: criticism

Issue: 9.6% of fitted values in `Bikeshare` data set negative, i.e., linear regression model predicts negative number of users during 9.6% of hours in data set.

- Ability to perform meaningful predictions from the data?
- Accuracy of coefficient estimates, confidence intervals, and other outputs of regression model?

Issue: expect small variance in `bikers` when expected value of `bikers` small.

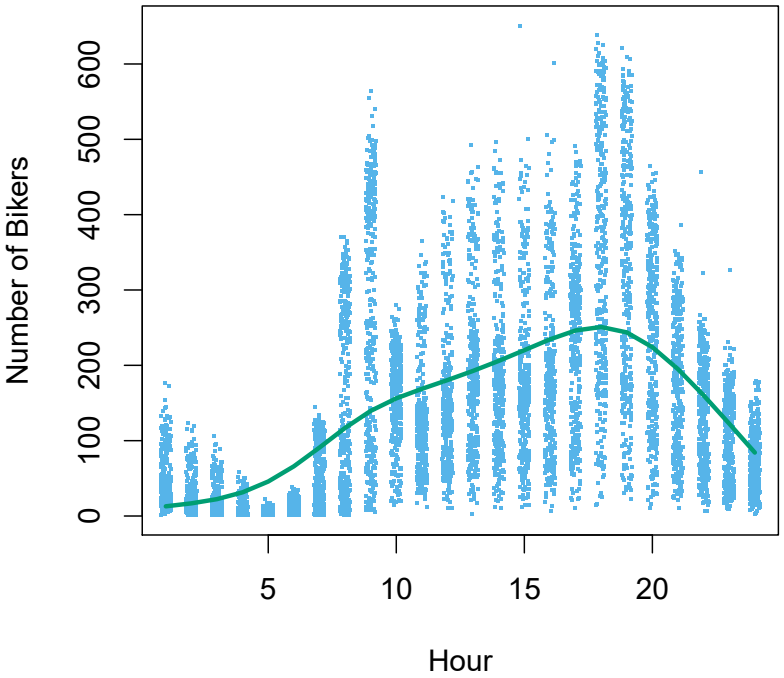
- At 2:00 during heavy December storm, expect few bikers and low variance.
- Borne out in data: Dec–Feb between 1:00 and 4:00 when raining: 5.05 bikers on average with standard deviation 3.73.
- Apr–Jun between 7:00 and 10:00 under clear skies; 243.59 bikers on average with standard deviation 131.7.
  Compare with linear regression assumptions:

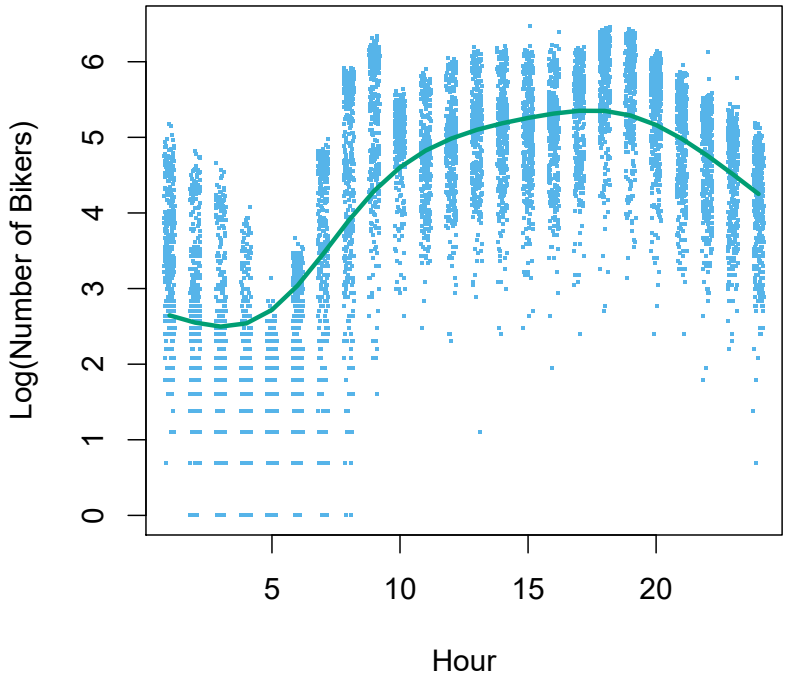
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad \textbf{Var}\,\varepsilon \text{ constant.}$$

Generalized Linear Models

Linear regression on `Bikeshare` data set: criticism

Fitting $\log Y$ rather than Y sometimes helps, but not here (0 counts, interpretation of coefficients).





Issue: response `bikers` integer-valued, regression model yields continuous values. Linear regression does not properly reflect discrete nature of response.

Poisson distribution: probability distribution on $\mathbb{N}_0$ (discrete), models number of times an event occurs in fixed interval of time or space.

- Models (rare) events occurring in a fixed interval with average frequency λ: e.g. # incoming phone calls per day, # web accesses per hour, etc.
- **intensity parameter** $\lambda > 0$ (or „rate“).
- **Probability mass function** of RV $X \sim \mathrm{Pois}(\lambda)$

$$\mathbf{P}(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}, \qquad k \in \mathbb{N}_0.$$

- $\mathbf{E}[X] = \mathbf{Var}\, X = \lambda$ (the larger the mean, the larger the variance)
- For **binomial distribution** $\mathrm{B}(n, p)$ with n large and p small, $\mathrm{Pois}(\lambda) \approx \mathrm{B}(n, p)$ with $\lambda = np$.

Generalized Linear Models

Poisson regression on `Bikeshare` data set: modeling

Model # hourly bike sharing users under during particular hour of day, under particular set of weather conditions, and during particular month of the year as Poisson random variable Y.

- Choosing $\lambda = \mathbf{E}[Y] = 5$ means that, under given conditions,

$$\mathbf{P}(Y=0) = \frac{\mathrm{e}^{-5}5^0}{0!} = 0.0067, \qquad \mathbf{P}(Y=1) = \frac{\mathrm{e}^{-5}5^1}{1!} = 0.0337,$$

$$\mathbf{P}(Y=2) = \frac{\mathrm{e}^{-5}5^2}{2!} = 0.0842, \ldots$$

- Crucial model feature: allow λ to vary with covariates: $\lambda = \lambda(X_1, \ldots, X_p)$.
- Poisson regression model:

$$\log \lambda(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

ensures positive values of λ for all values of covariates.

Generalized Linear Models

Poisson regression on `Bikeshare` data set: estimating parameters, fitted model

Maximum likelihood to determine coefficients: make likelihood of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda(\boldsymbol{x}_i)}\lambda(\boldsymbol{x}_i)^{y_i}}{y_i!}, \qquad \lambda(\boldsymbol{x}_i) = \mathrm{e}^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_p x_{p,i}}$$

as large as possible (cf. logistic regression).

Generalized Linear Models

Poisson regression on Bikeshare data set: estimating parameters, fitted model

Maximum likelihood to determine coefficients: make likelihood of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda(\boldsymbol{x}_i)} \lambda(\boldsymbol{x}_i)^{y_i}}{y_i!}, \qquad \lambda(\boldsymbol{x}_i) = \mathrm{e}^{\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}}$$

as large as possible (cf. logistic regression).

Resulting Poisson regression fit for `Bikeshare` data:

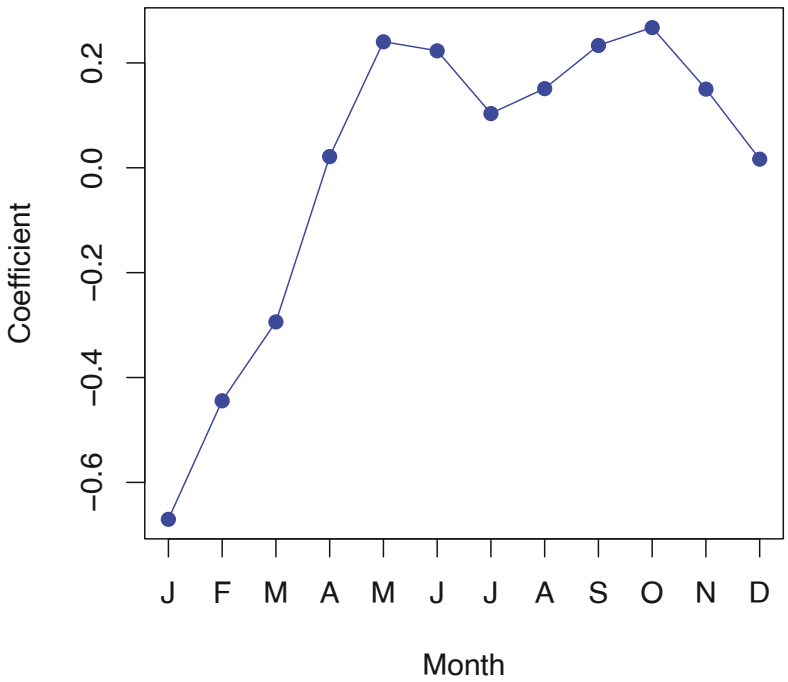
	Coefficient	Std. error	z-statistic	p-value
intercept	4.12	0.01	683.96	0.00
`workingday`	0.01	0.00	7.50	0.00
`temp`	0.79	0.01	68.43	0.00
`weathersit[cloudy/misty]`	-0.08	0.00	-34.53	0.00
`weathersit[light rain/snow]`	-0.58	0.00	-141.91	0.00
`weathersit[heavy rain/snow]`	-0.93	0.17	-5.55	0.00

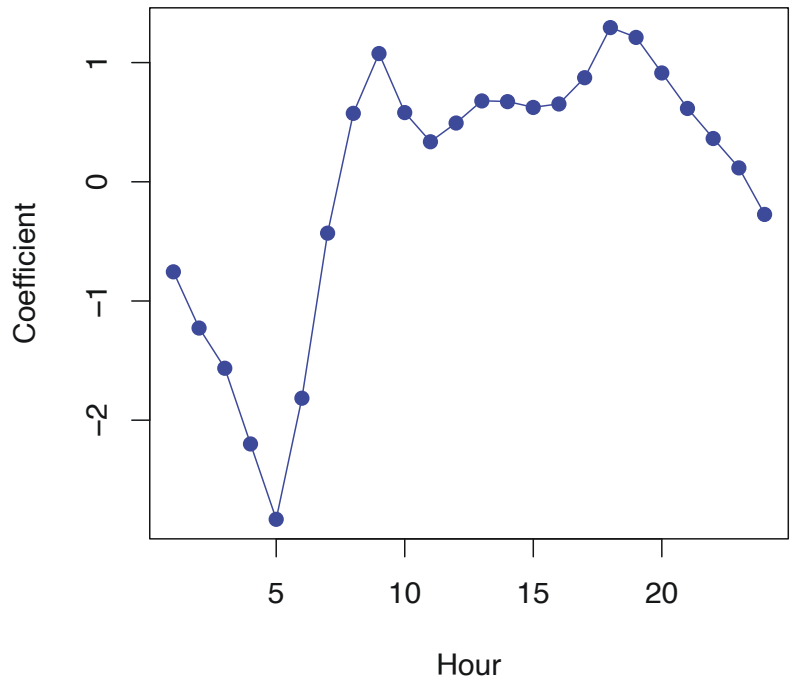
- Coefficient associated with `workingday` statistically significant under Poisson regression model, but not under linear regression model.
- Bike usage increases with temperature, decreases as weather worsens.

Generalized Linear Models

Poisson regression on `Bikeshare` data set: fitted model

Covariates `mnth` and `hr`:





- Bike usage again highest in spring and fall and during rush hour, lowest during winter and in early morning hours.

- **Interpretation:** increase in X_j by one unit associated with change in $\mathbf{E}[Y] = \lambda$ by a factor of $\exp(\beta_j)$.
  Example: change in weather from clear to cloudy associated with change in mean bike use by factor of $\exp(-0.08) = 0.923$, i.e., on average only 92.3% of clear sky usage.
  Further worsening to rain: additional factor $\exp(-0.58) = 0.56$ reduction in mean bike use relative to merely cloudy conditions.

Generalized Linear Models

Poisson regression on `Bikeshare` data set: distinction with linear regression

- **Interpretation:** increase in X_j by one unit associated with change in $\mathbf{E}[Y] = \lambda$ by a factor of $\exp(\beta_j)$.
  Example: change in weather from clear to cloudy associated with change in mean bike use by factor of $\exp(-0.08) = 0.923$, i.e., on average only 92.3% of clear sky usage.
  Further worsening to rain: additional factor $\exp(-0.58) = 0.56$ reduction in mean bike use relative to merely cloudy conditions.
- **Mean-variance relationship:** In view of $\lambda = \mathbf{E}[Y] = \mathbf{Var}\, Y$ for $Y \sim \text{Pois}(\lambda)$, by modeling bike usage with Poisson regression we implicitly assume that mean bike usage in a given hour equals variance of bike usage during that hour.
  Poisson regression model able to handle mean-variance relationship seen in `Bikeshare` data in a way the linear regression model is not.

Generalized Linear Models

Poisson regression on `Bikeshare` data set: distinction with linear regression

- **Interpretation:** increase in X_j by one unit associated with change in $\mathbf{E}[Y] = \lambda$ by a factor of $\exp(\beta_j)$.
  Example: change in weather from clear to cloudy associated with change in mean bike use by factor of $\exp(-0.08) = 0.923$, i.e., on average only 92.3% of clear sky usage.
  Further worsening to rain: additional factor $\exp(-0.58) = 0.56$ reduction in mean bike use relative to merely cloudy conditions.
- **Mean-variance relationship:** In view of $\lambda = \mathbf{E}[Y] = \mathbf{Var}\, Y$ for $Y \sim \text{Pois}(\lambda)$, by modeling bike usage with Poisson regression we implicitly assume that mean bike usage in a given hour equals variance of bike usage during that hour.
  Poisson regression model able to handle mean-variance relationship seen in `Bikeshare` data in a way the linear regression model is not.
- **Nonnegative fitted values:** No negative predictions using Poisson regression model.

Generalized Linear Models

Generalized linear models in general

Have discussed 3 types of regression models: logistic, linear and Poisson.

1. In each case, $Y|X_1, \ldots, X_p$ belongs to family of distributions:

   logistic: Bernoulli, linear: Gaussian, logistic: Poisson.

Have discussed 3 types of regression models: logistic, linear and Poisson.

1. In each case, $Y|X_1, \ldots, X_p$ belongs to family of distributions:

   logistic: Bernoulli, linear: Gaussian, logistic: Poisson.

2. Each approach models expectation of Y as a function of the predictors:

$$\mathbf{E}[Y|X_1, \ldots, X_p] = \begin{cases} \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p & \text{linear,} \\ \mathbf{P}(Y = 1|X_1, \ldots, X_p) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}} & \text{logistic,} \\ \lambda(X_1, \ldots, X_p) = \mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p} & \text{Poisson.} \end{cases}$$

Have discussed 3 types of regression models: logistic, linear and Poisson.

1. In each case, $Y|X_1,\ldots,X_p$ belongs to family of distributions:

   logistic: Bernoulli, linear: Gaussian, logistic: Poisson.

2. Each approach models expectation of Y as a function of the predictors:

$$\mathbf{E}\left[Y|X_1,\ldots,X_p\right] = \begin{cases} \beta_0+\beta_1X_1+\cdots+\beta_pX_p & \text{linear,} \\ \mathbf{P}(Y=1|X_1,\ldots,X_p) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}} & \text{logistic,} \\ \lambda(X_1,\ldots,X_p) = \mathrm{e}^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p} & \text{Poisson.} \end{cases}$$

3. The general common structure of these models is that a function of $\mathbf{E}\left[Y|X_1,\ldots\right.$ is a linear function of the predictors:

$$\eta(\mathbf{E}\left[Y|X_1,\ldots,X_p\right]) = \beta_0+\beta_1X_1+\cdots+\beta_pX_p,$$

$\eta(\mu)=\mu$ (linear), $\eta(\mu)=\log(\mu/(1-\mu))$ (logistic), $\eta(\mu)=\log\mu$ (Poisson).

Generalized Linear Models

Generalized linear models in general

Gaussian, Bernoulli and Poisson belong to **exponential family** of probability distributions. (Others include *exponential*, *Gamma*, *negative binomaial*, etc.)

General approach:

1. Model response Y as coming from particular member of exponential family;
2. transform expectation of response so that transformed expectation is linear function of predictors.

Any regression approach following this general recipe is known as a **generalized linear model(GLM)**.

Hence linear, logistic, and Poisson regression are three examples of GLMs. Other examples not covered here: *Gamma regression*, *negative binomial regression*.