Applied Statistics and Data Analysis

6. Predictive and classification methods

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine

paolo.vidoni@uniud.it

Partly based on Chapters 2, 3 and 4 of *An Introduction to Statistical Learning: with Applications in R* by G. James et al.

Table of contents

Summary

- Introduction to predictive modeling
- Predictive model accuracy
- Prediction using regression models
- The classification problem

Introduction to predictive modeling

Based also on Chapter 1 of *Applied Predictive Modeling* by M. Kuhn and K. Johnson

- Predictive modeling is a "process by which a model is created or chosen to try to best predict the probability of an outcome" (Geisser S., *Predictive Inference: An Introduction*. Chapman and Hall, 1993).
- The model or the mathematical tool which is developed is considered for giving accurate prediction.
- For example, insurance companies aim at predicting the risk of potential auto, health and life policy holders. This is crucial in order to determine if an individual will receive a policy and, if so, at what premium.
- Governments use predictive models for evaluating potential risks, with the aim of protecting their citizens. For example, biometric models for identifying terror suspects and models for fraud detection.
- Internet companies apply predictive models to guide consumers towards more satisfying products or more profitable investments.

- Although predictive models aim at indicating more satisfying products, better medical treatments and more profitable investments, they may generate inaccurate predictions and give wrong answers
- There are a number of reasons why predictive models fail. The main culprits are:
  - inadequate pre-processing of the data;
  - inadequate model selection and validation;
  - unjustified extrapolation (application of the model to data outside the range of the available observations);
  - over-fitting the model to the existing data.
- It is surely important to specify reliable and trustworthy predictive models, however the accuracy of our prediction will be affected by an irreducible error component.
- This unavoidable error term is related, for example, to the fact that relevant predictor variables may be missed, that there are unmeasurable, and then not exploitable, variables (such as those related to personal human behavior) and that prediction are usually constrained by our present and past knowledge.

Prediction versus inference

- Suppose that we observe a quantitative response Y and $p \geq 1$ different explanatory variables (predictors) $X = (X_1, \ldots, X_p)$ and that the following general model is defined

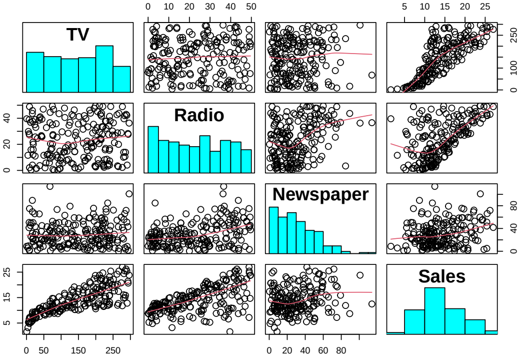
  $$Y = f(X) + \varepsilon$$

  The fixed, unknown function $f(\cdot)$ represents the systematic information on Y provided by X and ε is a zero-mean random error term. Regression models fall into this framework.
- **Inference**: understand the relation between Y and X and, in particular, how Y changes as a function of $X_1, \ldots, X_p$; the exact form of f is usually needed and it is essential to obtain an estimate $\widehat{f}$ based on the available data.
- **Prediction**: given a set of inputs for X, the aim is to predict the associated value for Y using a predictor $\widehat{Y} = \widehat{f}(X)$ or a prediction interval; $\widehat{f}$ may be treated as a *black box*.
- Applications may fall into the prediction setting, the inference setting, or a combination of the two.

Example: advertising data

Data from *An Introduction to Statistical Learning: with Applications in R* by G. James et al.

Data set on `sales` of a certain product (response variable Y) along with the advertising budgets for three different media, `TV`, `radio`, `newspaper` (predictor variables X_1, X_2, X_3); values in thousands related to $n = 200$ different markets.



A multiple linear regression model suggest that the effect of `newspaper` on `sales` is not statistically significant

Coefficients	Estimate	SE	p-value
Intercept	2.9389	0.3119	<0.0001
TV	0.0458	0.0014	<0.0001
radio	0.1885	0.0086	<0.0001
newspaper	-0.0010	0.0059	0.8599
	$s^2 = 2.841$	R^2_{adj}=0.896	

In the **inference framework** one may be interested, for example, in finding which media contributes significantly to sales, which media generates the biggest boost in sales, what is the increase in sales associated with a given increase in TV advertising.

In the **prediction framework** one may be interested, for example, in predicting the amount of sales given a fixed budget for the three media.

Table of contents

1. Summary and introduction
2. Predictive model accuracy
3. Prediction using regression models
4. The classification problem

Measuring the quality of fit

- By considering the available data (**training observations**) $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, a suitable model may be fitted, obtaining the estimate $\widehat{f}$.
- In order to evaluate the performance of the model, it can be useful to evaluate how well the predicted response values $\widehat{y}_i = \widehat{f}(\mathbf{x}_i)$ are close to the true response values y_i, $i = 1, \ldots, n$.

  A common measure, used in the regression setting, is the ***training* mean squared error (MSE)** given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(\mathbf{x}_i))^2$$

- The MSE is computed using the training observations, which have been previously employed for model fitting; therefore, the fitting accuracy is in fact measured instead of the predictive accuracy.
- Since the same data are used twice, the training MSE would give an overoptimistic predictive assessment of the model.

- In order to evaluate the predictive performance of a model, it is more convenient to consider how well the predicted response values $\widehat{y}_{0j} = \widehat{f}(\mathbf{x}_{0j})$ are close to the true response values y_{0j}, $j = 1, \ldots, m$, with regard to previously unseen observations.
- While the model is fitted using the training observations, the prediction accuracy is now evaluated by considering the **test observations** $(\mathbf{x}_{01}, y_{01}), \ldots, (\mathbf{x}_{0m}, y_{0m})$, not used to train the statistical model.

  Thus, a measure of predictive fit is given by the ***test* MSE**

  $$testMSE = \frac{1}{m} \sum_{j=1}^{m} (y_{0j} - \widehat{f}(\mathbf{x}_{0j}))^2$$

  where $\widehat{f}$ is estimated using the training observations.
- In some settings, a test data set can be available; for example, when the original data set is sufficiently large.
- Whenever, as usual, no test data are available, methods for estimating test MSE using the training data can be considered. One important method in **cross-validation**.

- If one selects a statistical model by minimizing the training MSE there is no guarantee that the lowest test MSE is achieved.
- The training MSE is usually smaller than the test MSE. Furthermore, while the training MSE declines as model flexibility increases, the test MSE initially declines and then start to increase again.
- This phenomenon is known as **overfitting**: the model focuses exaggeratedly on patterns that are just caused by randomness and it misses the true properties of the unknown function f.
- The test MSE may be viewed as an estimate for the **expected test MSE** for the future random response Y_0, with a given value $\mathbf{x}_0$

  $$E[(Y_0 - \widehat{Y}_0)^2] = V(\widehat{f}(\mathbf{x}_0)) + [\mathrm{Bias}(\widehat{f}(\mathbf{x}_0))]^2 + V(\varepsilon)$$

  where $\widehat{Y}_0 = \widehat{f}(\mathbf{x}_0)$ is the point predictor based on the fitted model.
- This value can never lie below $V(\varepsilon)$, the **irreducible error term**. The minimum is achieved for models that simultaneously have low variance and low bias.
- The **bias-variance trade-off**: usually, as more flexible methods are considered, the variance will increase and the bias will decrease.

Table of contents

Regression versus classification problems

- Variables can be characterized as either quantitative or qualitative (also known as categorical).
- Predictive problems with a quantitative response are usually called **regression problems**, while those involving a qualitative response are often referred to as **classification problems**.
- The distinction is not always sharp, since logistic regression, which is often used as classification method, may be viewed as an extension of linear regression models with the aim of modeling probabilities on a transformed scale.
- The present section focuses on prediction using (multiple) linear regression models. Some notions, already discussed in the previous sections, are reviewed by studying two data sets. The next section is devoted to the analysis of classification problems.

Example: automobile bodily injury claims

Data from *Regression Modeling with Actuarial and Financial Applications* by E.W. Frees

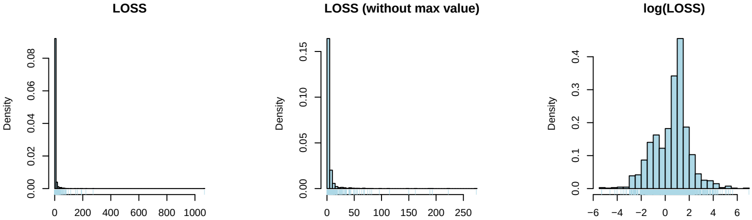
Data from the Insurance Research Council, US, collected in 2002 and regarding automobile bodily injury claims. Observations on the variables:

- `ATTORNEY`: whether the claimant is represented by an attorney (`yes`, `no`)
- `CLMSEX`: claimant gender (`male`, `female`)
- `MARITAL`: claimant marital status (married `M`, single `S`, widowed `W`, divorced `D`)
- `CLMINSUR`: whether or not the driver of the claimant's vehicle was uninsured (`yes`, `no`)
- `SEATBELT`: whether or not the claimant was wearing a seatbelt/child restraint (`yes`, `no`)
- `CLMAGE`: claimant's age
- `AGECLASS`: claimant's age split into five classes: `(-18]`, `(18,26]`, `(26,36]`, `(36,47]`, `(47+)`
- `LOSS`: claimant's total economic loss (in thousands)

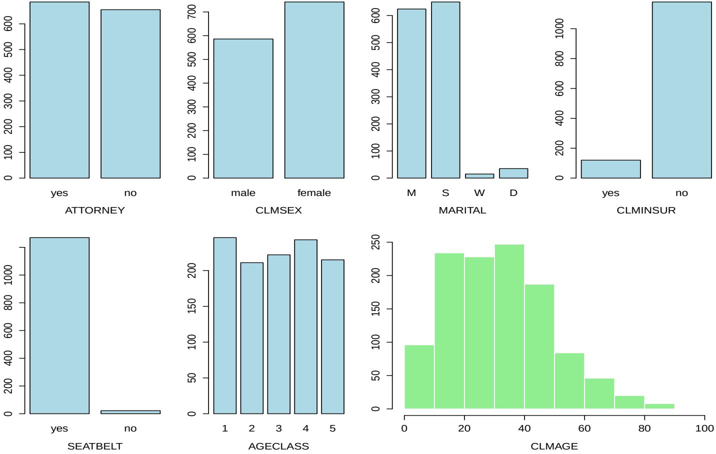
It is of interest to build a statistical model for predicting claim amounts in future policies, based on a sample of claim amounts of past policies.

The severity refers to the amount of claim. The response variable is quantitative and it is given by the claimant's total economic loss.

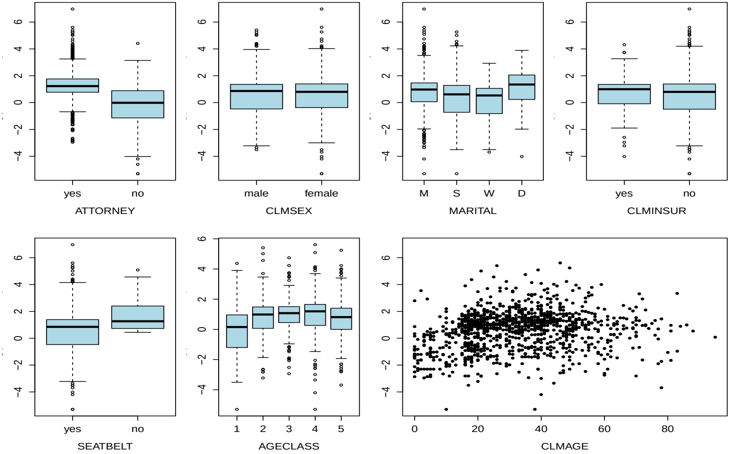
The logarithmic scale is chosen, as motivated by the following histograms, so that the response variable is `logLOSS`.



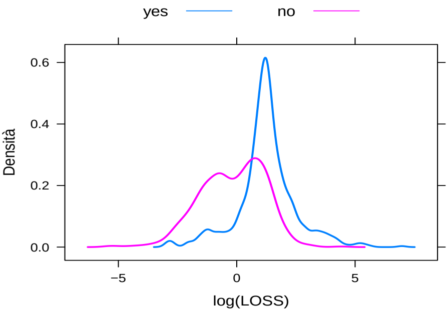
Information about policies enables a marginal description of the explanatory variables



A further analysis suggests that a significant relationship exists between `logLOSS` and some explanatory variables



Being represented by an attorney seems to be important, as shown by considering the density estimate of `logLOSS` in case of `ATTORNEY=yes` and `ATTORNEY=no`



Also other variables may matter, such as `SEATBELT`, `MARITAL` and `AGECLASS`.

On the contrary `CLMSEX` and `CLMINSUR` seem not to have an effect. The effect of `CLMAGE` is not clear.

The joint effects of ATTORNEY (qualitative variable coded as a dummy variable) and CLMAGE on the response logLOSS can be described using the following multiple linear regression model

$$\texttt{logLOSS} = \beta_0 + \beta_1\,\texttt{ATTORNEY} + \beta_2\,\texttt{CLMAGE} + \varepsilon$$

	Estimate	SE	p-value
Intercept	0.7376	0.0851	< 0.0001
ATTORNEYno	-1.3699	0.0729	< 0.0001
CLMAGE	0.0160	0.0021	< 0.0001
	$n - p = 1148$	$s = 1.23$	R^2_{adj}=0.259

Both the coefficients are strongly significant, pointing to a relevant effect of both ATTORNEY and CLMAGE on the mean response.

The dummy variable flags those observations that have the level of ATTORNEY equal to no. Since its estimated coefficient is negative, subjects without an attorney have a smaller mean response.

Note that 189 observations were deleted due to missingness.

It is easy to estimate the mean of the `logLOSS` and then to transform back the results on the scale of the original `LOSS` variable.

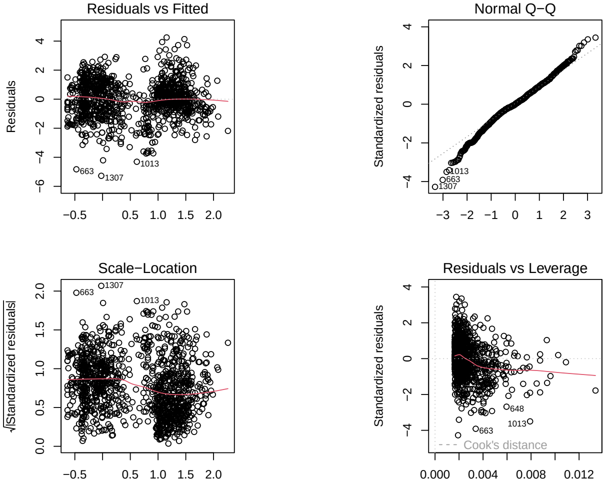
The estimated mean response for a subject of age 30, in case of `ATTORNEY=yes`, is $\widehat{y} = 0.7376 + 30 \cdot 0.016 = 1.22$, with

- Standard error: 0.05
- Confidence interval: $[1.12, 1.32]$
- Estimate on LOSS scale: $e^{1.22} = 3.38$
- Interval on LOSS scale: $[e^{1.12}, e^{1.32}] = [3.06, 3.73]$

The estimated mean response for a subject of age 30, in case of `ATTORNEY=no`, is $\widehat{y} = 0.7376 - 1.3699 + 30 \cdot 0.016 = -0.15$, with

- Standard error: 0.05
- Confidence interval: $[-0.26, -0.05]$
- Estimate on LOSS scale: $e^{-0.15} = 0.86$
- Interval on LOSS scale: $[e^{-0.26}, e^{-0.05}] = [0.77, 0.95]$

The diagnostic plots are moderately good, then the model is acceptable for practical purposes.



With the aim of looking for a better model specification, the following regression model is specified (according to R^2_{adj}, AIC and the test MSE estimated by simple Cross Validation)

$$\texttt{logLOSS} = \beta_0 + \beta_1\,\texttt{ATTORNEY} + \beta_2\,\texttt{CLMAGE} + \beta_3\,\texttt{CLMAGE}^2 + \beta_4\,\texttt{SEATBELT} + \varepsilon$$

The qualitative variables `ATTORNEY` and `SEATBELT` are coded as dummy variables and the quadratic effect of `CLMAGE` is also considered.

	Estimate	SE	p-value
Intercept	-0.2249	0.1376	0.1024
ATTORNEYno	-1.3522	0.0725	< 0.0001
CLMAGE	0.0828	0.0075	< 0.0001
CLMAGE^2	-0.0009	0.0001	< 0.0001
SEATBELTno	0.9241	0.2681	0.0006
	$s^2 = 1.404$		$R^2_{\mathrm{adj}} = 0.321$

Diagnostic plots do not highlight any serious flaw in the model.

One of the main usage of the model is for **out-of-sample predictions**.

The **point predictor** for the response variable Y_0, which corresponds to a certain value $\mathbf{x}_0$ for the covariates, is $\widehat{Y}_0 = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}$.

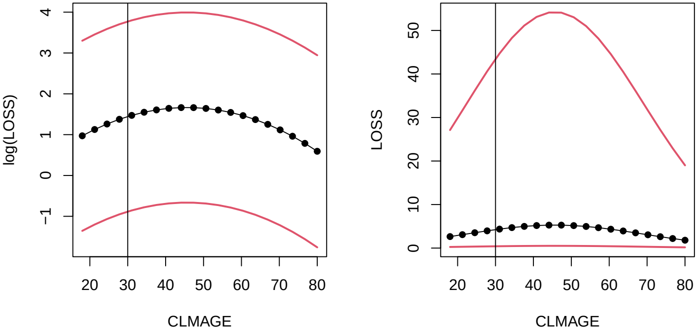
The **prediction error** (and then the SE of prediction) is made of two parts: the randomness of Y_0 and estimation error associated to the linear predictor $\widehat{\mu}_0 = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}$.

The objective is to predict the `LOSS` for an insured of age 30, that was represented by an attorney and that was wearing a seatbelt.

- The estimated variance for the fitted model (which correspond to the training MSE) is 1.40, which is close to 1.41, that is the test MSE assessed by cross-validation: here the overoptimism given by the training data is small.
- The 95% **prediction interval** for `log(LOSS)` is $[-0.89, 3.77]$, much wider than the confidence interval for $\widehat{\mu}_0$ given by $[1.33, 1.55]$.
- Adopting the original scale, the intervals are $[0.41, 43.43]$ and $[3.80, 4.72]$, respectively. Using the cross-validation-based SE of prediction, the prediction interval becomes $[0.41, 43.62]$.

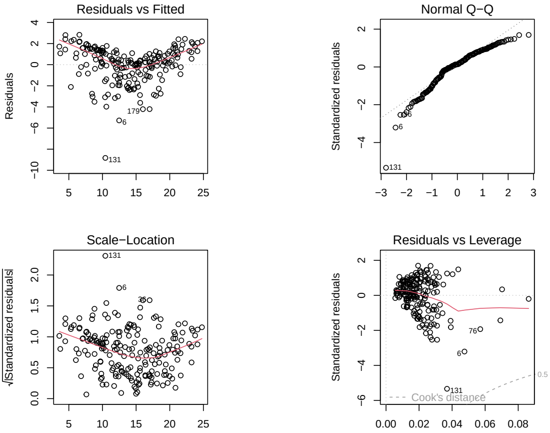
It is possible to compute the 95% prediction intervals for `logLOSS` and `LOSS`, for insureds represented by an attorney, wearing a seatbelt and with age ranging from 18 to 80 years.

The vertical lines identify the prediction intervals given before, for an insured of age 30.



Example: advertising data

The linear additive regression model for the `sales`, considered before, does not give an adequate description of the data set. This is confirmed by the diagnostic plots, which suggest some possible nonlinear effects of the covariates.

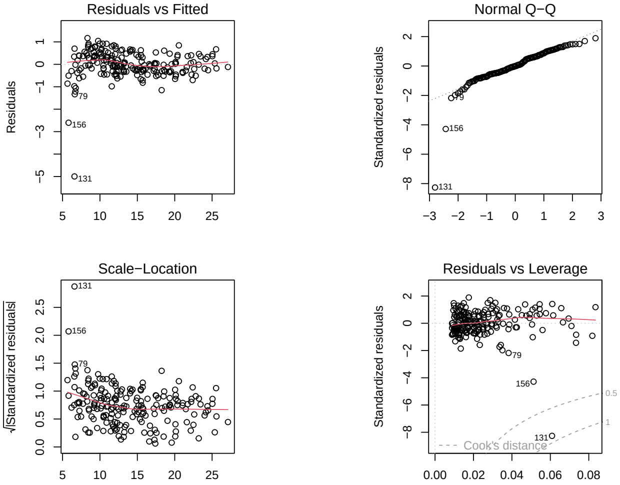


A model with additive effects of TV and radio plus their interaction and the quadratic effect of TV is considered

$$\texttt{sales} = \beta_0 + \beta_1\,\texttt{TV} + \beta_2\,\texttt{radio} + \beta_3\,\texttt{TV}^2 + \beta_4\,\texttt{TV:radio} + \varepsilon$$

It has a much smaller AIC than the additive model. Indeed, the estimated variance s^2 is smaller too, while R^2_{adj} is higher.

	Estimate	SE	p-value
Intercept	5.1371	0.1927	<0.0001
TV	0.0509	0.0022	<0.0001
radio	0.0352	0.0059	<0.0001
TV^2	-0.0001	0.0000	<0.0001
TV:radio	0.0011	0.0000	<0.0001
	$s^2 = 0.389$		$R^2_{\text{adj}} = 0.986$



The model is apparently better, though there are some worrisome local deviations.

Once the multiple regression model is fitted, it is easy to predict the response Y_0 on the basis of a set of values for the predictors $\mathbf{x}_0$.

While confidence intervals quantify the uncertainty surrounding the average sales over a large number of markets, prediction intervals can be used to quantify the uncertainty surrounding sales for a particular market.

For example, given that 100 is spent on TV advertising and 20 is spent on radio advertising in each market (in thousands $), the 95% confidence interval for $\mu_0 = \mathbf{x}_0^T \boldsymbol{\beta}$ is $[11.864, 12.112]$.

On the other hand, assuming the same spent amount for TV and radio advertising, the 95% prediction interval for Y_0 is $[10.752, 13.225]$.

Note that both intervals are centered at 11.988, but the prediction interval is wider, reflecting the increased uncertainty about sales for a given market in comparison to the average sales over many locations.

Table of contents

1. Summary and introduction
2. Predictive model accuracy
3. Prediction using regression models
4. The classification problem

The classification setting

- In the prediction framework, the interest response variable may be qualitative (categorical) rather than quantitative. In such instances, the process of **predicting** the category of a new observation is referred to as **classification**.
- There are some connections with the regression setting, as often the methods used for classification predict the probability of each of the categories of the response variable, and such probabilities are numerical scores (like the fitted values of a regression model).
- The classification problem is ubiquitous in all the applications of statistics.
- Some examples: an email filter must classify a new email as `spam` or `not spam`; a person who arrives at the hospital emergency room with some symptoms has to be attributed to one of three medical conditions (such as `stroke`, `drug overdose`, `epileptic seizure`).

Example: credit scoring

Data from *Regression: Models, Methods and Applications* by L. Fahrmeir et al.

A bank loan service has to evaluate the possible loan insolvency for a customer, and decide whether *customer will default* or *customer will pay back*: this is a problem which typically arises in the credit scoring setting.

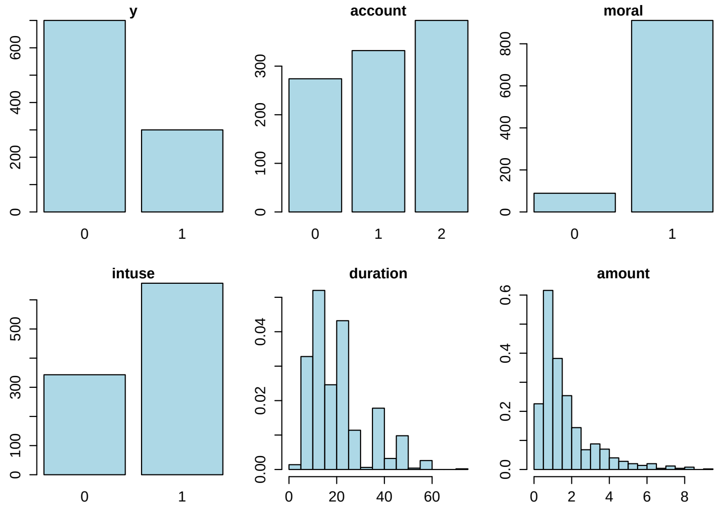
Data set on $n = 1000$ private credits issued by a German bank. Training data where every client is associated with a binary response

- `Y`: the client did not pay back his loan 1, the client paid back 0

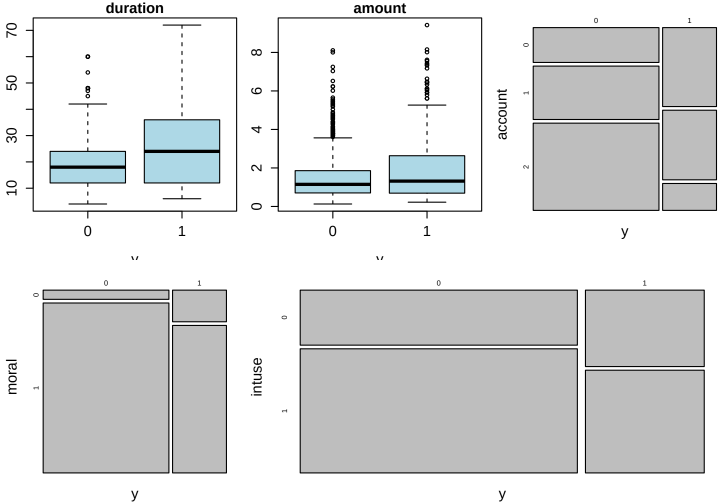
and five explanatory variables (predictors)

- `account`: no running account 0, bad running account 1, good running account 2
- `duration`: duration of the credit (in months)
- `amount`: credit amount (in thousands euros)
- `moral`: previous payment behavior, bad 0, good 1
- `intuse`: intended use, business, 0, private 1

The data set enables a marginal description of the response variable and of the explanatory variables



A further analysis points to the potential relationships between the binary response and the explanatory variables



Classification of a categorical response

- Many of the concepts related to predictive model accuracy transfer over to the classification setting, with some modifications.
- The case of a binary response is initially considered. The response for the i-th unit is coded as $y_i \in \{0, 1\}$, $i = 1, \ldots, n$.
- A classification method makes use of the training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ to build a **classifier** that, for a given set of predictors $\mathbf{x}_0$, returns a binary classification $\widehat{y}_0 \in \{0, 1\}$, which is very similar to what obtained for linear regression.
- The most common approach for quantifying the accuracy of the classifier is the ***training* error rate**, the proportion of mistakes that are made if the classifier is applied to the training observations

$$trainingER = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \widehat{y}_i)$$

where $\widehat{y}_i$ the predicted i-th observation and $I(y_i \neq \widehat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \widehat{y}_i$ and 0 otherwise.

- Since the same data are used twice, the training error rate would give an overoptimistic predictive assessment of the classifier.
- As in the regression setting, it is more convenient to evaluate the error rate on observations that were not used for training.

  Thus, given the **test observations** $(\mathbf{x}_{01}, y_{01}), \dots, (\mathbf{x}_{0m}, y_{0m})$, the ***test* error rate** is

$$testER = \frac{1}{m}\sum_{j=1}^{m} I(y_{0j} \neq \widehat{y}_{0j})$$

  which may be also computed from the training observations using **cross-validation**.

- The test error rates is an estimate of the **prediction** (**classification**) **error** for the future random response Y_0 corresponding to $\mathbf{x}_0$

$$E[I(Y_0 \neq \widehat{Y}_0)] = P(Y_0 \neq \widehat{Y}_0)$$

  with $\widehat{Y}_0$ the associated classifier (predictor) .

- The aim is to define a classifier which minimizes the prediction error.

The Bayes classifier

- It is possible to show that the **best** classifier (with the smallest classification error) assigns an observation with predictor $\mathbf{x}_0$ to the class 1 if

  $$P(Y_0 = 1|\mathbf{X}_0 = \mathbf{x}_0) > P(Y_0 = 0|\mathbf{X}_0 = \mathbf{x}_0)$$

  and to class 0 otherwise. (Note that the costs of all errors are assumed to be the same)
- This very simple classifier is known as the **Bayes classifier**, which is a gold standard achieving the best classification.

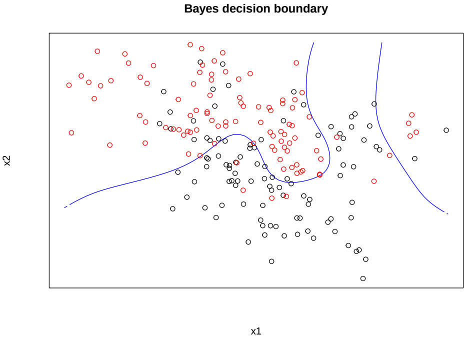
  Unfortunately, for real data, the conditional probability distribution of Y_0 given $\mathbf{X}_0 = \mathbf{x}_0$ is **not known**, as it depends on the true distribution of the response variable. Therefore, computing the Bayes classifier is impossible.
- Classification methods try to approximate such conditional probability following different assumptions and methodologies. Their performances would depend on how good this approximation is.

Example: two-predictors simulated data

Data from *An Introduction to Statistical Learning: with Applications in R* by G. James et al.

Two continuous predictors, X_1 and X_2, and simulated binary responses (100 observations in each class, red and **black** circles).

Since the true model is known, the conditional probability is available and the Bayes classifier can be obtained. The blue line represents the points with probability 0.5: the **Bayes decision boundary**.



Classification based on logistic regression

- Logistic regression models specify directly the conditional probability of the response, as approximated by the model and given by

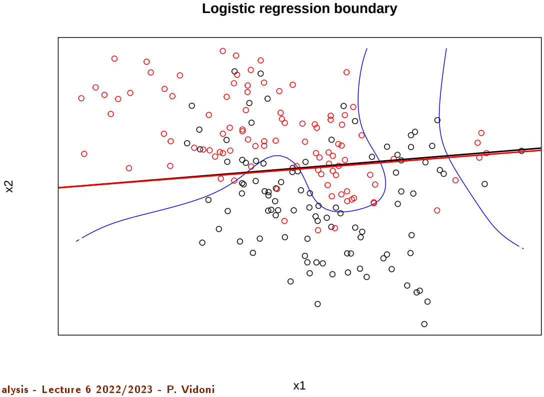
$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}$$

- The decision boundary, which corresponds to the the values $\mathbf{x}$ such that $\widehat{P}(Y = 1|\mathbf{X} = \mathbf{x}) = 0.5$, has the following linear form $\mathbf{x}^T \widehat{\boldsymbol{\beta}} = 0$.
- Classification based on logistic regression may be satisfactory whenever the Bayes decision boundary is roughly linear.
- All the theory developed for regression models readily applies, even though, when the goal is classification, less attention is paid to inference on the coefficients, focusing instead on the classification performances.
- Extension to more than two categories for the response are possible, but not used very often.

Example: two-predictors simulated data

The classification rule based on a logistic regression model has a linear boundary such that $\widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \widehat{\beta}_2 \, x_2 = 0$.

In this case, this is quite different from the Bayes decision boundary, and very similar to that one obtained using a **linear regression model** for the binary response, given by $\widehat{\beta}_0 + \widehat{\beta}_1 \, x_1 + \widehat{\beta}_2 \, x_2 = 0.5$.



Linear discriminant analysis (LDA)

- Linear regression for binary (or categorical) response often gives results very similar to logistic regression, as shown in the two-predictor example.

  Yet, it is fundamentally unsatisfactory, as it may give fitted probabilities outside $[0, 1]$.
- Linear models for classification can be defined in an indirect way, starting from a model for the predictors and obtaining the conditional probability of interest $P(Y_i = y_i | \mathbf{X}_i = \mathbf{x}_i)$ by the Bayes' theorem.
- This is the essence of **Linear Discrimination Analysis** (**LDA**), that ends up in a linear classification boundary in the space of (continuous) predictors.

- Suppose that the response variable Y can take on $S \geq 2$ unordered values y_s, $s = 1 \ldots, S$, and that $\pi_s = P(Y = y_s)$ is the associated (prior) marginal probability.

  Let $f_s(\mathbf{x}) = f(\mathbf{x}|Y = y_s)$ denote the density function of the p-dimensional vector $\mathbf{X}$ of continuous predictors for a response observation in the s-th category.
- According to the Bayes' theorem,

  $$P(Y = y_s|\mathbf{X} = \mathbf{x}) = \frac{\pi_s f_s(\mathbf{x})}{\sum_{r=1}^{S} \pi_r f_r(\mathbf{x})}$$

  so that the Bayes classifier assigns an observation with predictor $\mathbf{x}$ to the category for which $P(Y = y_s|\mathbf{X} = \mathbf{x})$ is largest.
- An approximate Bayes classifier is then defined by considering suitable estimates for $f_s(\mathbf{x})$ (and, if needed, for the membership probability π_s), $s = 1, \ldots, S$, using the training observations.

- The LDA requires that, for the case $p = 1$, $f_s(x)$ is the density of a Gaussian distribution $N(\mu_s, \sigma^2)$, with a class-specific mean value μ_s and a constant variance σ^2.
- The approximation for the Bayes classifier is obtained by plugging into $P(Y = y_s|\mathbf{X} = \mathbf{x})$ the estimates

$$\widehat{\mu}_s = \frac{1}{n_s} \sum_{i:y_i=y_s} x_i, \quad \widehat{\sigma}^2 = \frac{1}{n-S} \sum_{s=1}^{S} \sum_{i:y_i=y_s} (x_i - \widehat{\mu}_s)^2, \quad \widehat{\pi}_s = \frac{n_s}{n},$$

  with n_s the number of training observations belonging to the s-class.
- The LDA assigns an observation with $X = x$ to the class for which (using a simple transformation)

$$\delta_s(x) = x\,\frac{\widehat{\mu}_s}{\widehat{\sigma}^2} - \frac{\widehat{\mu}_s^2}{2\widehat{\sigma}^2} + \log(\widehat{\pi}_s)$$

  is largest. This function is *linear* in x. The procedure can be easily extended to the case with $p > 1$.

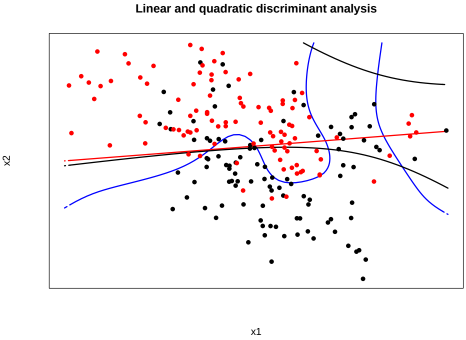
- If there are two categories, $\delta_1(x) = \delta_2(x)$ defines a linear decision boundary.
- LDA and logistic regression often perform very similarly, though actually LDA is numerically more stable.
- LDA assumes the **normality of the continuous predictors**, with class-specific means and **common variance**, so that it is not suitable for categorical predictors.

  If the assumptions hold, and if the sample is very large (so that estimation error is very small), then the LDA classification rule approximates well the Bayes rule.
- Relaxing the assumption of common variance leads to **Quadratic Discriminant Analysis** (**QDA**), resulting in quadratic classification boundaries. It usually requires larger samples, and it is not so much used in practice.

Example: two-predictors simulated data

The classification rules based on the linear discriminant analysis and on the **quadratic discriminant analysis** have, respectively, a linear boundary and a quadratic boundary.

In this case, they are quite different from the Bayes decision boundary. The LDA produces similar classifications to those obtained using a logistic regression model.



k-Nearest Neighbors (kNN)

- The method of **k-Nearest Neighbors (kNN)** is a simple procedure, pertaining to the class of *instance-based classification methods*, that classify a new unit by using the observations in the training set with similar predictor values.
- To classify a new observation with predictor $\mathbf{x}_0$, the kNN classifier identifies the $k > 0$ points in the training data that are closest to $\mathbf{x}_0$, forming the set $\mathcal{N}_0$. Then, the conditional probability is estimated by

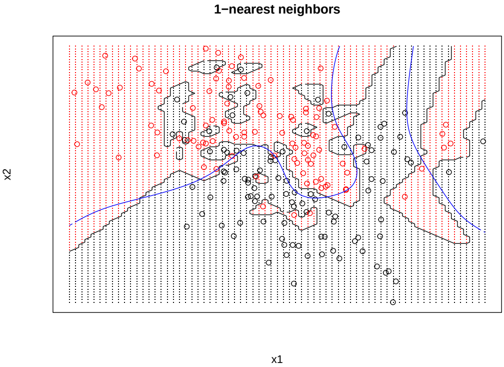
$$\widehat{P}(Y_0 = 1|\mathbf{X}_0 = \mathbf{x}_0) = \frac{1}{k} \sum_{i \in \mathcal{N}_0} y_i$$

- The value k is a positive integer that has a strong impact on the performance of the method.
- The best choices for k may lead to rather good performances (close to the Bayes classifier), despite the simplicity of the method.

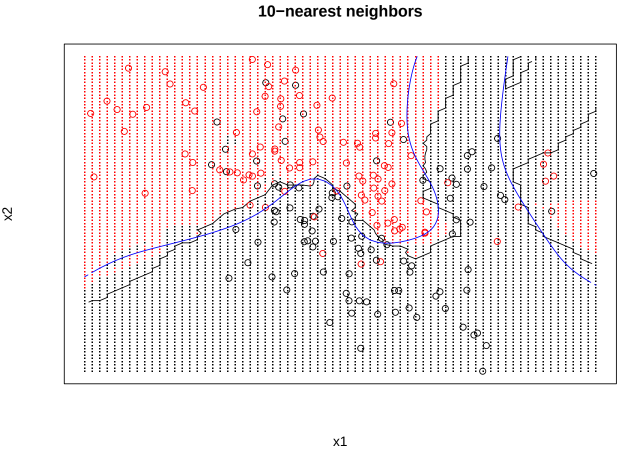
Example: two-predictors simulated data

The choice $k = 1$ uses only the closest point for classification. The plot below reports the classification done for a dense grid of values, giving the **classification boundary**.

The classification is even too detailed, with respect to that given by the Bayes decision boundary.



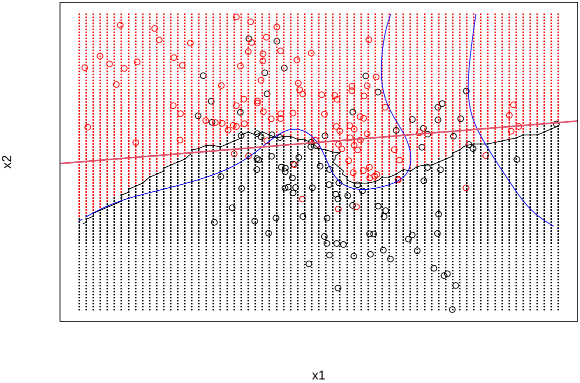
The choice $k = 10$ gives a less flexible classification, with a **classification boundary** much closer to the Bayes decision boundary.



The choice $k = 50$ gives a rough classification. The set of neighbors of each point is enlarged too much.

However, the **classifier** is closer to the Bayes classifier than the linear boundary obtained via logistic regression.

50−nearest neighbors



Comments

- The kNN classifier can be very effective with numeric predictors, especially when the Bayes decision boundary is highly nonlinear.

  The method can also be used with categorical or mixed predictors, by suitable definition of which are the neighbors of each observation.
- In order to choose the best value for k, the optimality criterion is that of *minimizing the rate of wrong classifications*. This is easy to do for the training data, but once again it is important to use, for example, cross-validation to avoid over-optimistic assessments.
- k plays the role of the amount of smoothing/flexibility in the decision procedure, with the usual bias-variance trade off:
  - small k leads to small bias and large variance (classification susceptible to noise);
  - large k leads to large bias and small variance.
- kNN can be used also for regression problems, when the goal is to approximate a continuous response.

Confusion matrix

- The predictive performance of a binary classifier can be summarized using a **confusion matrix**, which cross-classifies the observed frequencies and the predicted ones

	$Y = 0$ (obs)	$Y = 1$ (obs)
$Y = 0$ (pred)	True Negative (TN)	False Negative (FN)
$Y = 1$ (pred)	False Positive (FP)	True Positive (TP)

- The percentages of correct classification corresponds to the
  - true positive rate (sensitivity): TP/(TP+FN);
  - true negative rate (specificity): TN/(FP+TN);
  - total accuracy rate: (TP+TN)/(FP+TN+TP+FN).
- Further useful measures are the
  - positive predictive value: TP/(FP+TP);
  - negative predictive value: TN/(TN+FN);
  - log-odds ratio: $\log($TP*TN/(FN*FP)$)$.
- These quantities are obtained using the same data for fitting the predictive model and for evaluating the predictive accuracy. A less optimistic, and more realistic, evaluation can be obtained using CV.

ROC curve

- The choice of a classifying threshold equal to 0.5 is one possible choice. It could be useful to evaluate the global performance of a binary classifier taking into account all possible thresholds.
- The **Receiver Operating Characteristic** (**ROC**) **curve** is a popular graphic created by plotting the *true positive rate* against the *false positive rate* (given by $1-$*true negative rate*) at various threshold settings.
- The overall performance of a classifier can be described by the area under the (ROC) curve (**AUC**).
- An ideal ROC curve will pass near the top left corner, so that the larger the AUC the better the classifier.
- Roc curves are useful for comparing alternative binary classifiers, since they take into account all possible thresholds.

Example: credit scoring

A multiple logistic regression model is fitted, including all the five predictors, and it is used for classification on the same training data.

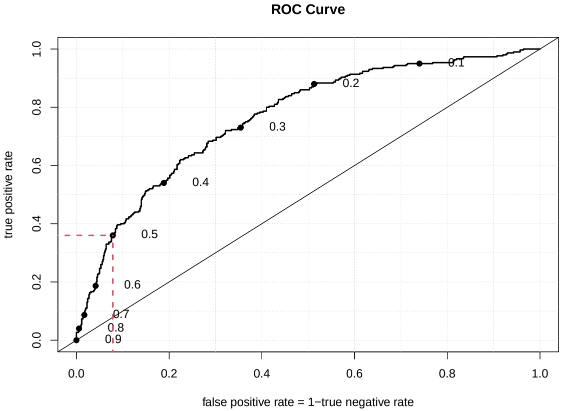
To summarize its predictive performance, the confusion matrix is calculated

	not defaulting (obs)	defaulting (obs)
not defaulting (pred)	645	192
defaulting (pred)	55	108
Total	700	300

The predictive model seems to be satisfactory only for predicting the customers not at risk of defaulting: the **true negative rate** and **true positive rate** are $645/700 = 0.921$ and $108/300 = 0.360$, respectively; the **total accuracy rate** is $(645 + 108)/1000 = 0.753$.

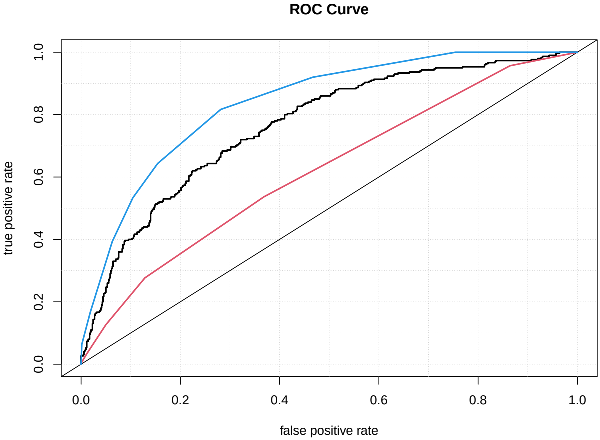
A more reliable evaluation is obtained via (10-folds) cross-validation, but in this case it makes little difference: the rates are 0.920, 0.350 and 0.749, respectively.

The threshold 0.5 for classifying an observation is not necessarily the best choice, especially if the two errors have different cost. The **ROC curve** illustrates the performance of the binary classifier for all possible thresholds.



ROC curves are useful for comparison, with the best classifier producing the highest curve.

Here the comparison regards the classifiers based on **logistic regression**, on LDA using only the numerical predictors and kNN with optimal K.



With regard to the credit scoring data, the optimal K value for the kNN classifier is $K = 9$.

This value has been chosen by a leave-one-out cross-validation procedure.

For such a non-linear classifier, the assessment based only on the training data is indeed over-optimistic.

	tot. accuracy rate	true negative rate	true positive rate
training data	0.79	0.90	0.53
cross-validated	0.75	0.89	0.43

A further nonlinear classifier could be obtained by extending logistic regression, including nonlinear terms for the continuous predictors in the same way illustrated for linear regression models. The additional gain of such extension is, however, rather limited in this case.

Further methods

- Classification is a very broad area, and the methods presented here are just the simplest ones. A (partial) list of other useful methods may include:
  - **Naive Bayes**: simple classifier which treats the predictors as independent random variables
  - **Classification trees** and **decision stumps**: very general, simple methods providing results easy to understand
  - Ensemble methods (such as **Boosting**, **Bagging**, **Random forests**): usually based on iterated applications of a simple classifier
  - **Support vector machines**: combine linear models with instance-based methods.
- Classification is a fundamental problem in statistical learning and there are multiple methods available, each with some pros and cons.
- No matter which method is used, it is always essential to estimate the classification accuracy, avoiding over-optimistic assessments based only on the training data.