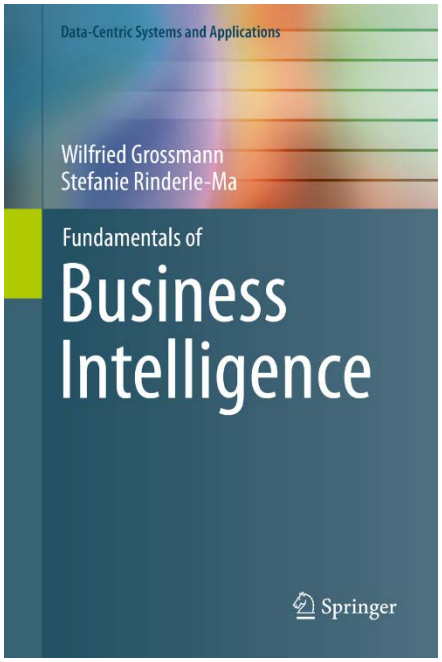


Chapter 5: Data Mining for Cross-Sectional Data

Contents

1 Introduction to Supervised Learning

- Starting point for supervised learning are data which are observations of random variables with two different interpretations:
  - *Explanatory variables*: $\mathrm{X}=(\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{p})$ also called *input variables* or *predictors*
  - *Response variable*: Y also called *output variable* or *explained variable*
- Notation for data from N observations:

  $N \times p$ matrix of input variables: $\boldsymbol{X}$

  $N \times 1$ vector of output variables: $\boldsymbol{Y}$

  - i-th observation: (x_i, y_i) Note that $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a p-dimensional vector

- Goal is to “learn” a function that allows prediction of the output variable from the input variables

- Distinguish two different types of Supervised Learning Problems
  - Regression Problems: Goal is prediction of a quantitative output variable from the input variable by a function

$$Y = f(X)$$

  - Examples:
    1. CRM Use case: Predict actual sales of customers dependent on characterizations of user profile like duration of customer relationship, sex, or sales in previous periods
    2. Used Car Prices: Predict the price of a used car from variables like age or mileage, etc.

- Classification Problems: In this case the output $Y \epsilon \{g_1, g_2, \ldots, g_k\}$ is a class identifier and the function defines a rule for class assignment based on the input variables:

$$Y = g_i = f(X)$$

  - Examples:
    1. CRM Use Case: Learning a classification rule which predicts the usage of a certain service from customers characteristics
    2. Credit Data: Find a rule, which allows a decision whether a credit defaults ($g = 1$) or not ($g = 0$) based on attributes of the customer like age, sex, family status, etc.

- Modeling task in supervised learning:
  - Definition of a class $\mathfrak{F}$ of functions, for example linear functions, decision trees, probability distributions,...
  - Specification of an appropriate *loss function* $L\big(Y, f(X)\big)$ which measures the error when the output Y is predicted by $f(X)$, for example quadratic loss:

$$L\big(Y, f(X)\big) = \big(Y - f(X)\big)^2$$

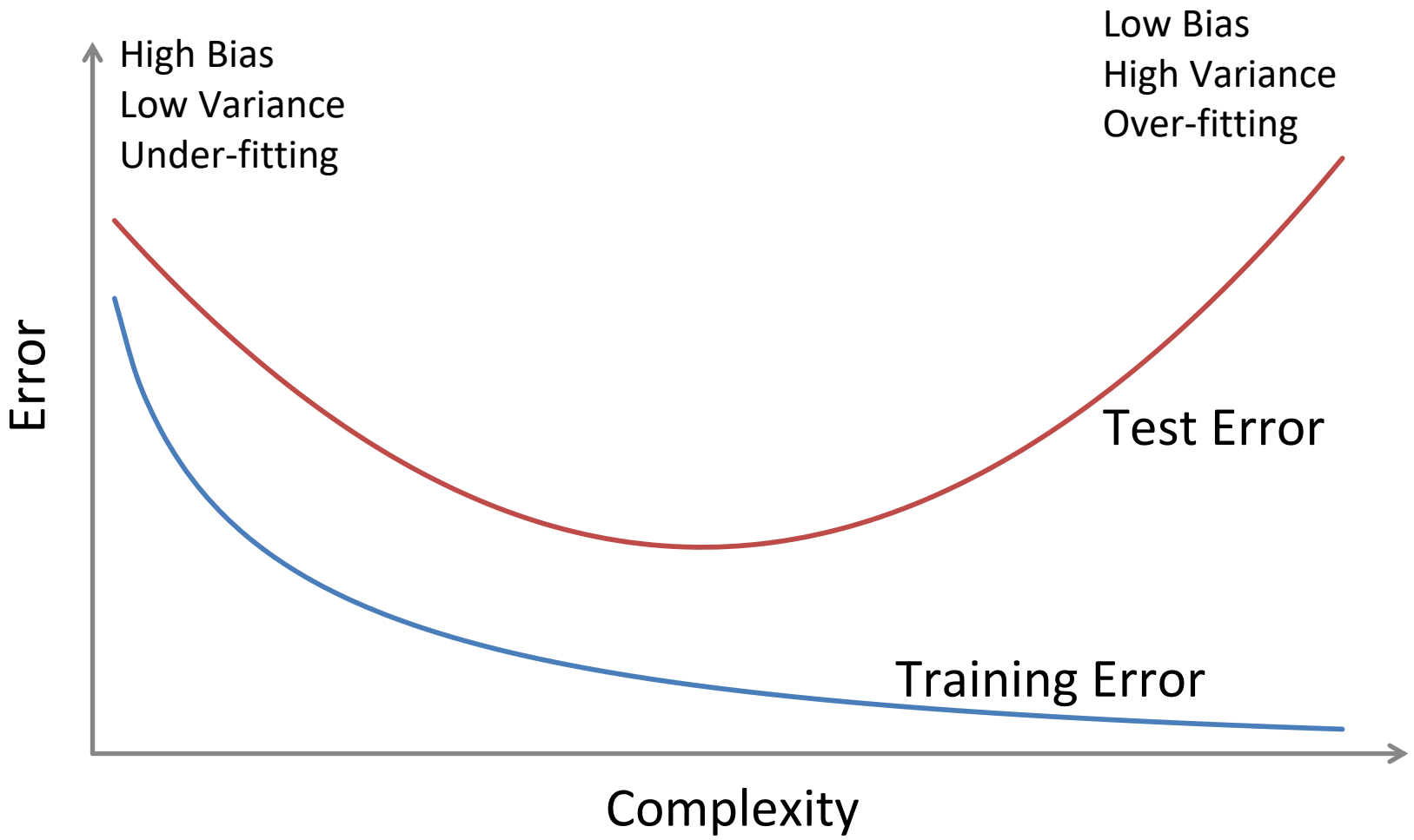
- Analysis task in supervised learning:
  - Determination of a function $\hat{f}(X)$ which has high predictive power
  - The predictive power is measured by the risk defined as the expected loss:

$$R = E\left[L\left(Y, \hat{f}(X)\right)\right]$$

- The risk is also called *generalization error* or *test error*
- The goal in supervised learning is to learn a function which minimizes the generalization error
- Components of the generalization error in the case of quadratic loss:
  - Irreducible error caused by the variability of the data
  - Variance caused by the fact that only a sample of all data is available
  - Squared bias caused by the choice of the model class
- The irreducible error is independent from the learning method and cannot be avoided
- Variance and bias depend on the specification of the model class $\mathfrak{F}$

- For understanding the role of the model class for variance and bias the idea of complexity of a model is used
- Usually the complexity of a model is measured by the number of the free model parameters which are estimated:
  - A simple model with low complexity will result in *underfitting*: the parameters are estimated with low variance but the model has a large bias
  - A model with high complexity will result in *overfitting*: the bias of the model is small but the parameters are estimated with high variance

  $\Rightarrow$ In both cases the test error will be rather large as shown in the figure on the next slide


High Bias
Low Variance
Under-fitting
Low Bias
High Variance
Over-fitting
Error
Test Error
Training Error
Complexity

- The graphic also shows that the *training error* also called *empirical risk*

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

is not a appropriate estimate for the test error because usually it decreases with model complexity and will underestimate the test error

– Due to the fact that the model class and the details of the model are unknown in supervised learning one has to solve three goals simultaneously:

- Estimation of the parameters for the models under consideration which minimize the empirical risk
- Selection the best model within the models under consideration
- Estimation of the test error (generalization error)

- The standard procedure for solving these tasks is splitting the data and proceed as follows:
  - Use one part of the data for estimating parameters of the models under consideration; using 50% of the data is recommended.
  - Use a second part of the data for selecting the appropriate model; 25% of the data is recommended.
  - Use the third part of the data for estimation of the test error; using 25% of the data is recommended.
- In this context the partition of the data are called *training data, validation data, and test data*
- For many learning procedures the first and the second step can be done simultaneously using theoretical considerations

- If no theoretical criteria for model selection are applicable one can use k-fold cross-validation for evaluation of the test error:
  - Schematic representation of 5-fold cross-validation

- Algorithmic description of k-fold cross-validation
  1. Divide the training data into k disjoint validation samples of roughly equal size
  2. For each validation sample use the remaining data to learn the decision function and estimate the empirical risk for the left out data
  3. Compute the overall prediction error by averaging the empirical risk of the validation data
- It should be noted that cross-validation gives an estimate of the average prediction error of the algorithm producing the learning rule and not the prediction error of the rule itself

Contents

2 Regression Models – Introduction

Problem formulation and terminology

- Regression models are used for prediction of a metric response variable Y from a number of predictors $X = (X_1, X_2, \ldots, X_p)$
- The model is defined by the equation:

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

f is a deterministic function from a model class $\mathfrak{F}$ and ε is the unexplained random component of the observations

  - Important model classes are parametric models, in particular linear functions
  - More general are nonparametric models, in particular neural nets
    - Nonparametric functions in one variable are often called *smoothers*

2 Regression Models – Introduction

Analysis template:

- *Relevant Business and Data*: Customer behavior represented by cross-sectional data for process instances
- *Analytical Goals*:
  - Estimation of the response function describing the relationship between input variables and output variables
  - Prediction of the output for new input values
- *Modeling Tasks*:
  - Definition of a modelling class, for example, linear models

- *Analysis Tasks*:
  - Split data randomly into one set for training and one set for testing the model
  - Estimate candidate models by solving the minimization problem for the empirical risk for the training data
  - Asses the quality of the model using residual analysis
  - Select the best model from the candidates using either theoretical considerations or data oriented methods (cross-validation)

– *Evaluation and Reporting Task*:

- Evaluate the selected model using the empirical risk for the training data defined by:

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i)) = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2$$

- Evaluate the generalization error for new test data (y_j^{test}, x_j^{test}), $j = 1, \dots, M$ defined by :

$$MSE_{emp} = \frac{1}{M}\sum_{j=1}^{M} L(y_j^{test}, \hat{f}(x_j^{test}))$$

- Notation:
  - The values of the estimated model are denoted by $\hat{y}_i, i = 1, \dots, N$
  - The estimated errors are called residuals and are denoted by $r_i = y_i - \hat{y}_i$

The subsection gives only a brief introduction into linear regression[1]

– Modeling task in linear regression

- Given the input variables $X = (X_1, X_2, \ldots, X_p)$ and the output variable Y the model class is defined by linear functions in the input variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_p X_p + \varepsilon$$

ε denotes the random component of the data not explained by the model

– Model estimation in linear regression

- Given data from N observations $(\mathbf{Y}, \mathbf{X})$ where $\mathbf{Y}$ is a N-dimensional vector and $\mathbf{X}$ is a $\mathrm{N} \times \mathrm{p}$ dimensional matrix the estimates of the parameters are obtained by the following minimization task

$$\sum_{i=1}^{N} (y_1 - \hat{y}_i)^2 = \min_{\beta} \sum_{i=1}^{N} \left(y_1 - {x_i}^T \beta\right)^2$$

[1]A detailed description with practical appications with R is: Faraway, JJ (2014): Linear models with R. Chapman & Hall/CRC.

– Assessment of the explanatory power in linear regression

- Assessment of the explanatory power is based on the decomposition of the total sum of squares (SST) in the explained sum of squares (SSE) and residual sum of squares (SSR):

$$\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

Symbolic notation: $\mathrm{SST}=\mathrm{SSE}+\mathrm{SSR}$

- The overall assessment:
  - is the *multiple R-squared* defined as the ratio $\mathrm{R}=\mathrm{SSE}/\mathrm{SST}$, which can be interpreted as percentage of variability in the data explained by the model
  - The second important measure is the *F-ratio* defined as $\mathrm{F}=\frac{SSE/p}{SSR/(n-p-1}$, which can be used for testing the null hypothesis that the model gives no significant explanation against a model defined by the mean value

– Assessment of model assumptions in linear regression

- A formal justification of the explanatory power depends on the following model assumptions:
  - Model specification: The deterministic component is correctly represented by the model
  - Independence: The observations are independent random variables
  - Homogeneity of variances: The variances of the error terms are the same for all observations
  - Normal distribution: The error terms are normally distributed
- Checking of these assumptions is based on the analysis and visualization of the residuals[1]

$$r_i = y_i - \hat{y}_i$$

1 See for example: Faraway, JJ (2014): Linear models with R, Chapter 6. Chapman & Hall/CRC

– Model selection in linear regression

- Model selection in linear regression is finding those input variables which define a model with high predictive power, i.e., a bias-variance trade-off which minimizes the test error
- The following methods are frequently used:
  - Variable selection methods: Different search strategies can be used for finding appropriate subsets of variables
  - Penalization methods: These methods add a penalty term to the loss and minimize the penalized loss.
  - Frequently used penalizations are the Akaike information criterion (AIC), Bayes information criterion (BIC)[3], Ridge regression, and Lasso[4]

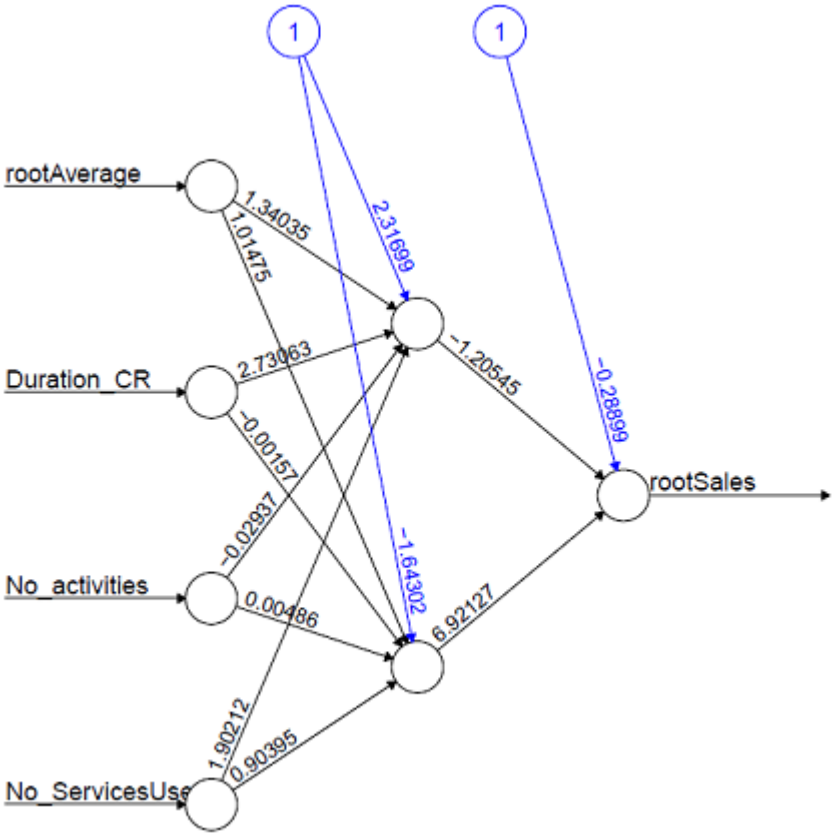
[3] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, Springer pp 203-208

[4] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, Springer pp 59-65

– Model formulation and terminology in neural networks

- Neural networks are of interest for modeling the relation between the input variables and the output in cases where only limited information about the relation between input and output is available
- This section outlines the basic ideas for using backpropagation networks for modeling regression problems[1]
- A neural network for a regression is a layered network defined by:
  - An input layer L_0 with p nodes corresponding to the number of input variables $X_1, X_2, \ldots, X_p$
  - A number of hidden layers $L_1, L_2, \ldots, L_K$ comprising n_k nodes in layer L_k
  - An output layer L_{K+1} with one node corresponding to the output variable Y
  - Edges e_{ij} connecting nodes $i \in L_k$ and $j \epsilon L_{k+1}$
  - Weight variables w_{ij} for the edges e_{ij}

- Example of a network with four input variables and one hidden layer with two nodes for prediction of sales in the CRM use case



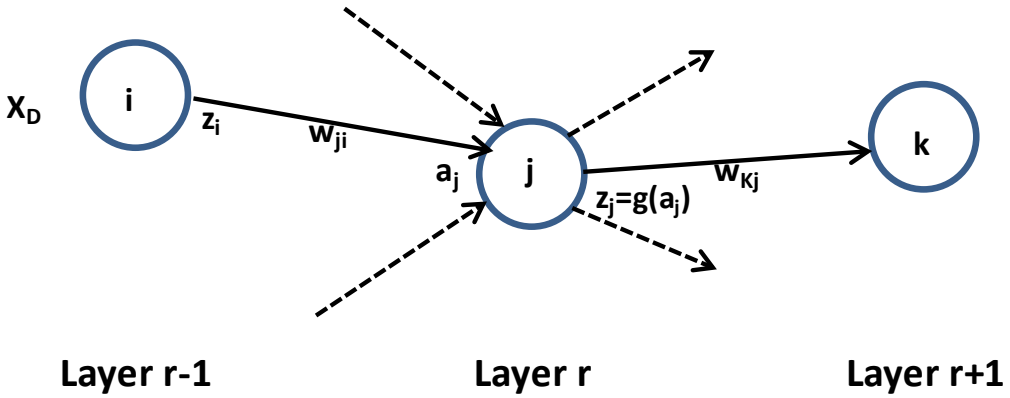
R package neuralnet

- The output of the network is obtained by propagating the input values in the following way:
  - The output values for the input layer L_0 are the values of the input variables $x_1, x, \ldots, x_p$
  - Using the output values z_i of the nodes in the layer L_{k-1} the input values a_j for the nodes in layer L_k are computed by

$$a_j = \sum_{i \in L_{k-1}} w_{ji} z_i + w_{j0}$$

    The weight w_{j0} is interpreted as constant input from a so called bias node
  - The output values are defined by the activation function from the input values in the node: $z_j = g_k(a_j)$

- Graphical representation of the calculation in a neural network



- Modeling task for neural networks:
  - Specification of the number of hidden layers
  - Number of nodes in the hidden layers
  - Definition of the activation functions:
    - For the hidden layers the most popular choice is the sigmoid function $g(t) = 1/(1 + e^{-t})$ for all layers
    - For the output layer the activation function is the identical function in the case of regression models

– Model estimation for neural networks

- The task of model estimation is finding values for the weights such that the empirical risk for the training data is minimized using the Backpropagation algorithm:
  1. Initialize the weights for the edges by random values
  2. Forward path: Propagate the input values of the training data to the output layer
  3. Backward path: Calculate the partial derivatives of the errors $E_i = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ with respect to the weights and adapt the weights in direction of the negative gradient using a step size parameter
  4. Repeat steps 2 and 3 until convergence of the errors is reached

- Details of backward path:
  1. For each observation the backward path starts from the error in the output layer $E_i = (y_i - \hat{y}_i)^2$ and calculates the derivatives of the error with respect to the weights between the output layer and the layer L_K
  2. Using the chain rule of calculus the derivatives of the error with respect to the weights between layer L_k and L_{k-1} can be calculated recursively from the errors in the layer L_k
- The calculations for all N observations can be organized efficiently as matrix computation

- Model assessment for neural networks
  - For model assessment the same methods as in the case of linear regression can be used
- Model selection and model evaluation for neural networks
  - Model selection is often experimental and encompasses definition of the numbers of layers and nodes in the layers
  - A formal method for avoiding overfitting is known as weight decay which is in line with the concept of penalization discussed in the section about linear regression
  - Importance of input variables can be measured by so-called generalized weights
  - The standard procedure for model evaluation is splitting in training and test data

2 Regression Models – Summary

- Regression models are used for the prediction of a quantitative response variable from a number of input variables
- Assessment of the quality of the data and relations between variables by descriptive summary measures and visualization tools discussed in chapter 4 is essential
- The most frequently used model class are linear models
- The standard method for estimation of the parameters is the method of least squares which minimizes the empirical risk defined by the quadratic loss
- The assessment of the model is based on the analysis of the residuals
  - If model assumptions are violated alternative models or robust estimation methods should be used[1]

[1]Maronna RR, Martin D, Yohai V (2006) Robust statistics, - theory and methods , Wiley

2 Regression Models – Summary

- Selection of an appropriate model can be done by different methods
  - The use of penalization methods is advisable
  - Alternatives are variable selection techniques or the method of cross-validation
- In BI applications testing the predictive power of the model for independent data is important (data splitting)
- In case of limited knowledge about the functional relationship between input and output variables neural networks offer an alternative
- Modeling with neural networks should use the same work flow as linear regression: specification of the network, estimation using backpropagation, model assessment, and model evaluation

Contents

Problem Formulation and Terminology

- In classification problems the output $Y \epsilon \{g_1, g_2, \ldots, g_k\}$ is a class identifier and the function defines a rule for class assignment based on the input variables:

$$Y = g_i = f(X)$$

- This goal can be achieved in two different ways
  - Class assignment: estimate the classification function
  - Class probabilities: estimate for each class the probability of the class membership for the observations given the input variables

$$p_g(x) = P(Y = g | X = x)$$

    Afterwards assign the observation to the class with highest probability
  - Usually both methods are computed

- For determination of the empirical risk the most important loss is the 0-1 loss defined by

$$L\left(y, \hat{f}(x)\right) = \begin{cases} 0 & if\ y = \hat{f}(x) \\ 1 & if\ y \neq \hat{f}(x) \end{cases}$$

- If misclassification causes different costs one can use a weighed loss
  - Example: Two classes, costs for wrong decision for class 1 is 5 times as costly as wrong decision for class 2

Loss function: $L\left(1, \hat{f}(x)\right) = 5, \quad L\left(0, \hat{f}(x)\right) = 1$

– Representation of the empirical risk by the *confusion matrix:*

$$C = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1k} \\ n_{21} & n_{22} & \cdots & n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ n_{k1} & n_{k2} & \cdots & n_{kk} \end{bmatrix}$$

- The diagonal represents the number of correct decisions, the other elements the number of wrong decisions
- The empirical risk is based on the sum of the off-diagonal elements of the confusion matrix
- In the case of different costs for misclassification the empirical risk is the weighted sum of the off-diagonal elements

- A number of classification methods are developed for the case of two classes
  - Generalizations to more than two classes can be obtained by one of the following methods
    - One versus the rest: For each class compute a classification against all other classes and classify the observations to the class with the highest probability
    - Classification of all pairs: Compute $\binom{K}{2}$ classifications for all pairs of classes, for each observation rank the classes according to the number of assignments in the classifications, and choose the class with highest rank

– In connection with classification problems for two classes the following terminology is used:

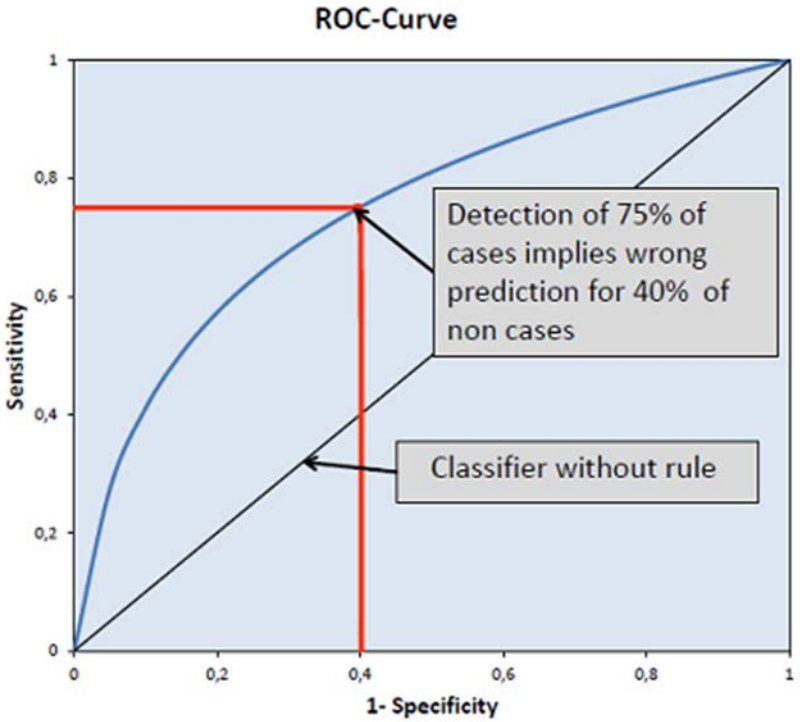
Prediction	Actual Class		
	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	Precision = TP/(TP+FP)
Negative	False Negative (FN)	True Negative (TN)	Negative Predicted Value =TN/(TN
	Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

– Further terminology:

- Precision = Positive predictive value
- Recall = Sensitivity
- False discovery rate = 1 – precision
- Accuracy = (TP + TN)/(TP + TN + FP + FN)

- In the case of two classes one can parametrize the decision rule class for class membership with a threshold value for the class probabilities
  - This parametrization implies a parametrization for sensitivity and specificity
  - The Receiver-Operator-Characteristic curve (*ROC curve*) is a plot of the sensitivity against 1- specificity for the different values of the threshold
  - The area under the curve is a measure for the quality of the decision rule
  - In case of random assignment of the classes the ROC curve is a straight line and the area under the curve is 0.5

- Example of a ROC curve for area under the curve of 0.73



Analysis template for Classification

- *Relevant Business and Data*:
  - Customer behavior represented by cross-sectional data for process instances $(\boldsymbol{Y}, \boldsymbol{X})$
- *Analytical Goals:*
  - Determination of a function for class membership for the observations
  - Determination of the probability of class membership for the observations
- *Modeling Task:*
  - Definition of a model class for the classification function (usually more than one model class is considered)

3 Classification Models - Introduction

- *Analysis Task:*
  - Random data splitting into training and test data
  - Model estimation for the selected model classes
  - Model assessment based on the confusion matrix, and ROC curve
  - Model selection using training data, penalization or k-fold cross-validation
- *Evaluation and Reporting Task:*
  - Evaluate the selected model using the test data

3 Classification Models - Introduction

- Descriptive methods for the tasks “Relevant Business and Data”, “Model Selection” and “Evaluation and Reporting”:
  - Considerations about the data generating process (quality)
  - Type of explanatory variables X: (categorical-qualitative or numeric-quantitative)
  - Frequency distribution of the numeric variables by summary measures, histograms, considerations whether normality of the distributions is an issue
  - Identification of outliers and of missing values
  - Correlations for numeric variables, scatterplot matrix
  - Grouped box plots for numeric variables with categorical explanatory variables and with the class identifier Y
  - Tables for categorical input data and visualization by bar charts and mosaic plots

3 Classification Models – Method Classes

- From the numerous methods for classification the following ones are discussed:
  - Classification methods using probabilistic structures: Bayes classification and logistic regression
  - Tree based models
  - K-nearest neighbor classification
  - Support vector machines
  - Combination methods (boosting)
  - Another important method are neural networks described in the context of regression
- The methods are explained for problems with two classes

– Modeling task in Bayes classification:

- Main components of a Bayes classifier:
  - Prior probabilities for the two classes:

$$P(Y = 0) = p(0),\ P(Y = 1) = p(1)$$

  - Distribution of the explanatory variables in the two classes:

$$p(x|0),\ p(x|1)$$

  - Posterior probabilities for the classes:

$$p(g|x) = \frac{p(x|g)p(g)}{x}\ , g = 0{,}1$$

  - Decision rule for an observation by maximization of the posterior probability:

$$\hat{y} = \begin{cases} 1 & if\ \dfrac{p(1|x)}{p(0|x)} \geq 1 \\ 0 & if\ \dfrac{p(1|x)}{p(0|x)} < 1 \end{cases}$$

– Model estimation in Bayes classification:
  - Prior probabilities are either known from understanding of the problem or estimated according to the size of the classes in the data
  - Estimation of the probabilities of the input data assumes that the input variables are independent (naïve Bayes approach):

$$p(x|g) = p(x_1|g) \cdot p(x_2|g) \cdot \cdots p(x_p|g)$$

  - Based on this assumption the priors for the input variables in the classes can be estimated for each variable separately
  - For qualitative input variables frequencies of the training data are used
  - For quantitative input variables either density estimates or a defined distribution (normal distribution) are used

3 Classification Models – Bayes Classification

- Example:
  - For 11 customers in the CRM use case the duration of the relationship to the company, the sales volume, and the type of used service is known
  - The following table shows the data and the decision obtained by a Bayes classifier (data in the grey area, decision in white area of the table)

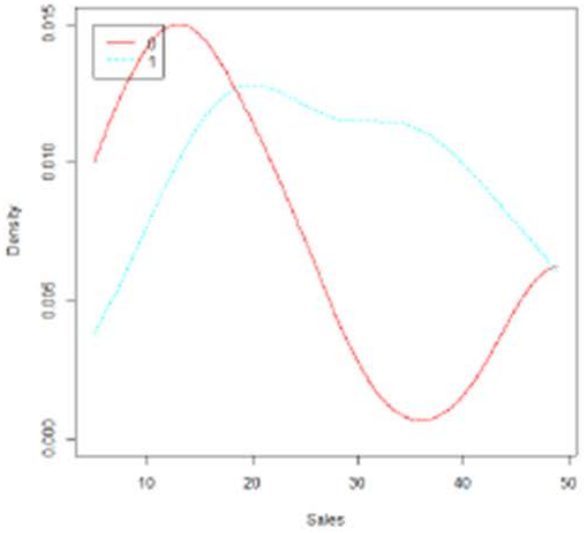
CR-Dur	Sales	User Type	UseService	P(no|x)	P(yes|x)	Decision
10	12	private	yes	0.5004	0.4996	no*
24	36	business	yes	0.0865	0.9135	yes
28	48	business	yes	0.5999	0.4001	no*
45	20	private	yes	0.3121	0.6879	yes
30	34	private	yes	0.0423	0.9577	yes
3	21	private	yes	0,3337	0.6663	yes
1	5	business	no	0.8300	0.1700	no
23	23	business	no	0.5414	0.4586	no
12	49	business	no	0.6672	0.3328	no
35	12	private	no	0.5080	0.4920	no
33	15	private	no	0.4389	0.5611	yes*
12	25	private	??	0.2804	0.7196	yes

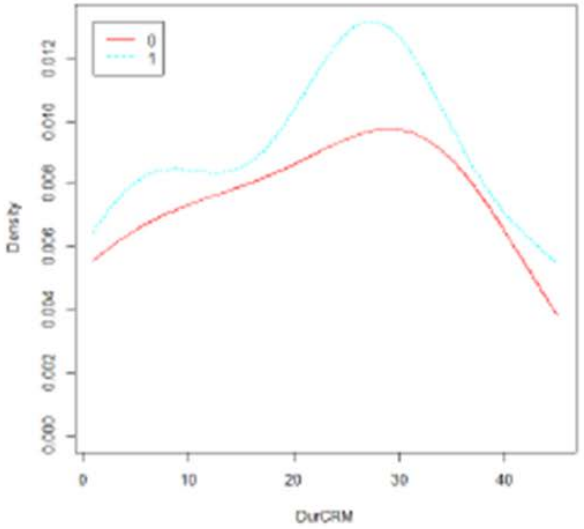
- The last line of the table shows the result for a new customer
- Due to the fact that there is only a small training set the accuracy is not very good

– Model assessment in Bayes classification

- For model assessment a plot of the estimated distributions of the variables as shown for the example for the variables duration and sales is recommended

R package e1071





- Model evaluation of Bayes classification
  - The advantages of naïve Bayes methods are:
    - Rather simple calculation
    - Rules for classification and probabilities
    - Simple adaptation to different misclassification costs
    - Straight forward generalization to more than two classes
  - A disadvantage is that there is no variable selection in the procedure
  - A well known successful application of Bayes classifiers are spam filters

– Modeling task in logistic regression:

- The probabilities for the two classes are defined as functions of the input variables by:

$$p(Y=1|x) = \frac{exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{1 + exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}, \qquad p(Y=0|x) = 1 - p(Y=1|x)$$

- This model implies that the logits. i.e. the logarithms of the odds for the two classes are a linear function of the input variables

$$logit = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_x$$

- The decision rule for an observation is defined by:

$$\hat{y} = \begin{cases} 1 & if\ p(Y=1|x) \geq tr \\ 0 & if\ p(Y=1|x) < tr \end{cases}$$

tr is an appropriately chosen threshold parameter (the standard case is $tr = 0.5$)

– Model estimation in logistic regression:

- The coefficients of the input variables in the logits are obtained by maximum likelihood estimation

– Model assessment in logistic regression

- For an overall assessment of the model the ROC curve is used

– Example:

- For 24 customers in the CRM use case it is known whether they have quit the relationship to the company;
- The goal is to predict quitting from the variables: “duration of the relationship with the company”, “activity index of the customer”, and “type of the user” (private or office)
- Application of logistic regression fro the data results in the following logit for churning (note that “duration of the relationship” is not used):

$$logit = 1.385 + 3.085 UserType - 0.577 ActInd$$

- Interpretation of the results:

  The risk of churning for a private customer is $\exp(3.058) = 21{,}3$ times the risk of an office customer, and increasing the activity index by one unit decreases the risk for churning by a factor $\exp(-0.577) = 0.56$

– Evaluation of logistic regression
  - Procedures for automatic selection of attributes can be used similar to regression
  - Many methods for model evaluation similar to linear regression can be used
  - Interpretation of the results as risk (see example)
  - Probabilities as well as class memberships are estimated
  - Adaptation to different costs of loss by using different thresholds
  - Generalization for more than two classes uses the method one versus the rest

– Modeling task in tree classification:
  - Given the training data $(\boldsymbol{Y}, \boldsymbol{X})$ from K different classes a tree model is defined by binary tree with nodes representing input variables and edges correspond to “yes” or “no” about the values of the input variables
  - The root node represents all observations in the training data which are split according to the decisions in the nodes into child nodes
  - The leaf nodes of the tree define a partition of the observations and are labeled with the class identifier to which the observations in the leaf nodes are assigned
  - The decision for a new instance is made by application of the decision rules to the input variables of this instance

– Estimation task in tree classification:
  - The estimation task has two components:
    - Splitting rules for construction of the tree
    - Pruning rules for simplification of the tree for avoiding overfitting
  - There exist different methods for construction of the tree
  - A frequently used method for building the tree is CART which stands for "Classification And Regression Trees"
    - The name indicates that CART can be also applied for regression models

Splitting rules in CART

- Splitting rules are defined by the the following conditions for variables:

  Quantitative variables: $X<tr\ or X>tr$, tr threshold value

  Qualitative variables: $X = ak\ \ or\ X \neq a_k$
- The choice of the variable used for split in a node is based on an impurity measure for the nodes, i.e., a numerical value which measures the mixture of different classes in the node t defined by the relative frequencies $\hat{p}(j|t)$ of the classes j in the node
- The following two impurity measures can be used in CART

  Gini Index: $Q(t) = \sum_i \sum_j \hat{p}(i|t) \cdot \hat{p}(j|t) = 1 - \sum_j \hat{p}(j|t)^2$

  Entropy: $Q(t) = -\sum_j \hat{p}(j|t) \cdot ln\left(\hat{p}(j|t)\right)$

- For each node the variable with minimal impurity measure after splitting is used as splitting variable (greedy search)
- Splitting stops if the node has impurity 0 or only a small number of data belonging to the node

Pruning rules

- Pruning of the tree is based on penalization of the empirical risk by the number of nodes |T |

$$R_{pen}(\alpha) = R_{emp} + \alpha|T|$$

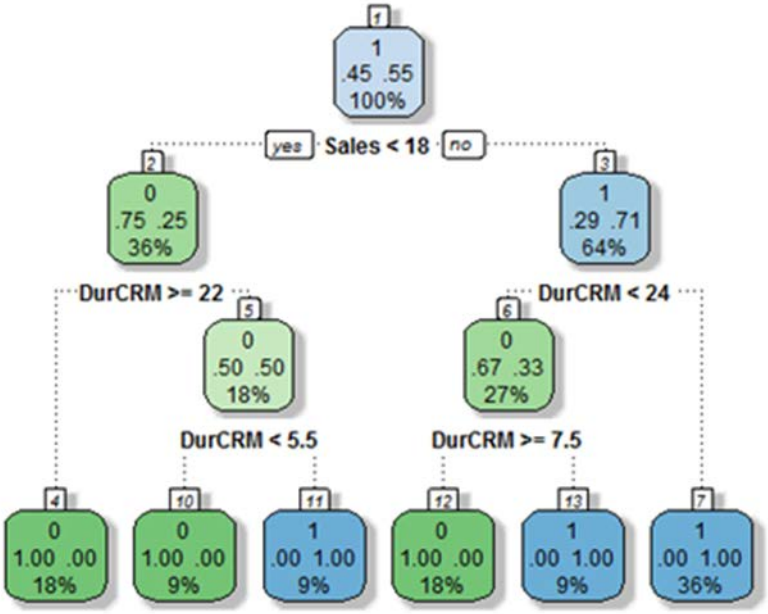
- It can be shown that only a final number of penalization parameters have to be checked and that there is a unique tree which minimizes the empirical risk
- The final tree is obtained by cross-validation from the candidate trees with the different penalization parameters

3 Classification Models – Tree Classification

– Example:

- Using the same data as in the example for Bayes classification the following tree is defined

R package rattle



– Evaluation of CART

- CART can be generalized to problems with different weights for misclassification costs
- Missing values can be handled within the procedure by so-called surrogate splits (alternative variable if the information for split is missing)
- CART is applicable to an arbitrary number of classes
- Estimation of class probabilities is rather simple
- CART is sensitive to the ordering of the variables and is dependent on the training data
- Splits do not take the dependency of the variables into account
- Tree classification is one of the most frequently used methods for classification

– Random Forests

- Random forests are a generalization of tree classifiers using the idea of bootstrapping in the following way:
  1. Generate a sample of size N from the training data by sampling with replacement
  2. Select randomly $r \ll p$ variables for the sample
  3. Build a classification tree from the data without pruning
  4. Repeat this process for M samples and define the classification rule by the majority of the vote of the classifiers

– Modeling task in k-nearest neighbor classification:
  - Given the training data $(\boldsymbol{Y}, \boldsymbol{X})$ the model uses a distance between the observations which is calculated from the input variables
  - For a new observation the distances of the new observation from the training data are computed and sorted in increasing order
  - The new observation is assigned to the class which is most frequent in the k nearest observations of the training data

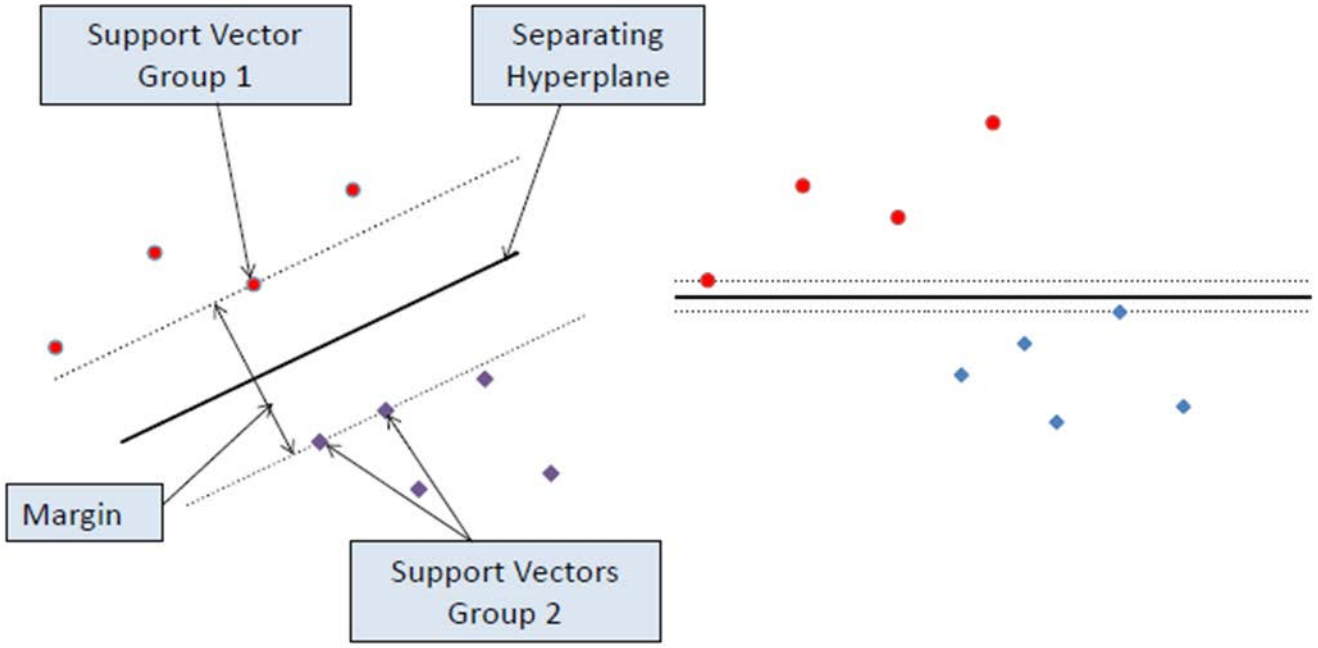
– Estimation task in k-nearest neighbor classification:
  - Definition of the distance between the observations:
    - In the case of quantitative variables the distance is usually the Euclidean distance
    - In the case of binary variables the Hamming distance is frequently used
    - Combination of qualitative and quantitative variables needs special consideration
      - The function daisy in the R package cluster supports calculation of distances for variables with different scales
  - Choice of k: The simplest case is k = 1, k = 5 is often recommended in the case of two classes

– Evaluation of k-nearest neighbor classification
  - Easy to calculate
  - From theoretical point of view the method is close to Bayes classifiers
  - No explicit learning step is necessary (instance based learning, lazy learning)
  - A disadvantage is that the method is sensitive to local structure of the data
  - Variable selection is not integrated in the method

– Modeling task for support vector machines
  - Given training data ($\boldsymbol{Y}$,$\boldsymbol{X}$) from two classes labeled by $y_i \epsilon \{-1,1\}$ a support vector machine separates the classes by a linear function (hyperplane)
  - The quality of separation is measured by the *margin*, i.e. the minimum distance of the the points from the hyperplane
  - Vectors which have minimum distance are called *support vectors*
  - The goal is to find a hyperplane for the training data which maximizes the margin

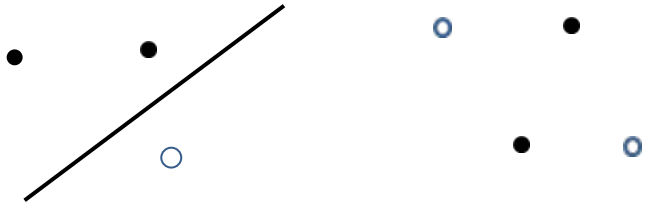
3 Classification Models – Support Vector Machines

- Graphical representation for separation of observations in two dimensions



- Main issue in modeling is the question whether it is always possible to separate observations from two classes perfectly by a function from a certain function class
- This maximum number of observations from two groups in any position which can be separated by a function class is called the *Vapnik-Chernovenkis dimension* (VC – dimension) of the functions class
- In two dimensions obviously only three points can be separated by linear functions

three points separated four points not separable

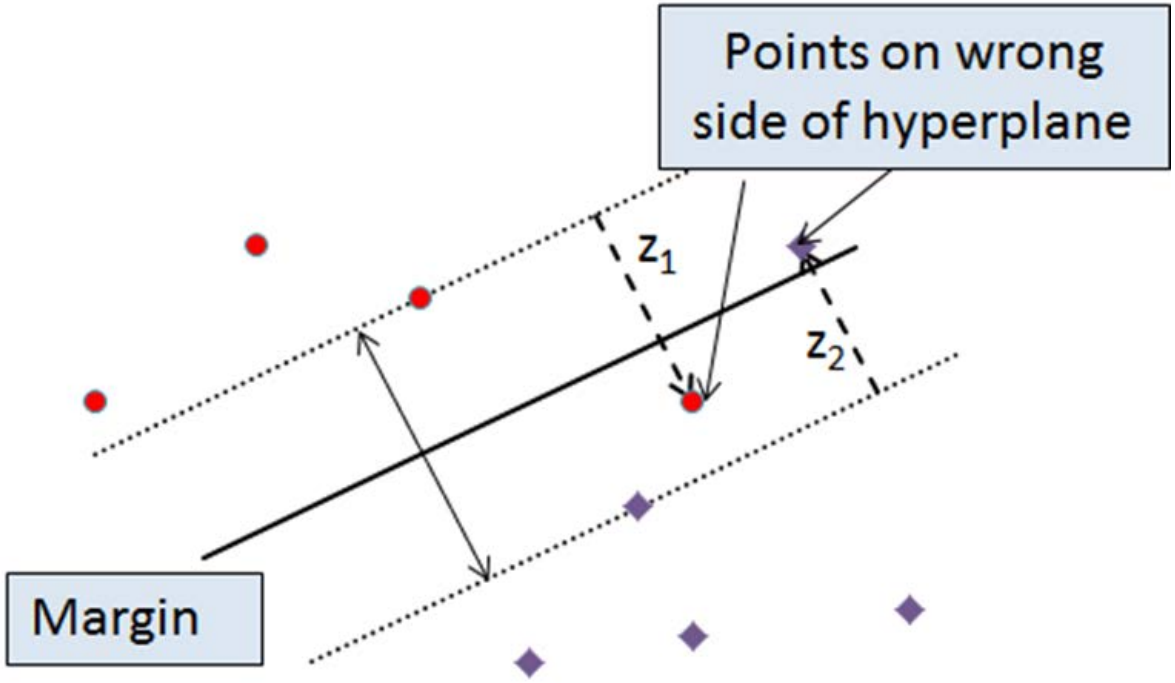


- More generally in case of linear functions in k dimensions the VC-dimension is $k + 1$

– Estimation task for Support Vector Machines:

- The algorithm for finding the best separating hyperplane is based on three ideas
  - First idea: If observations can be separated by a hyperplane the best separating hyperplane of the solution is obtained by solving a high dimensional quadratic optimization problem
  - Second idea: If perfect separation is not possible for the observations then misclassified observations are penalized by their distance from the margin defined (see graphic on the next slide)
  - This penalization of misclassified points is also known as *soft margin*
  - Algorithmic solution leads again to a quadratic optimization problem

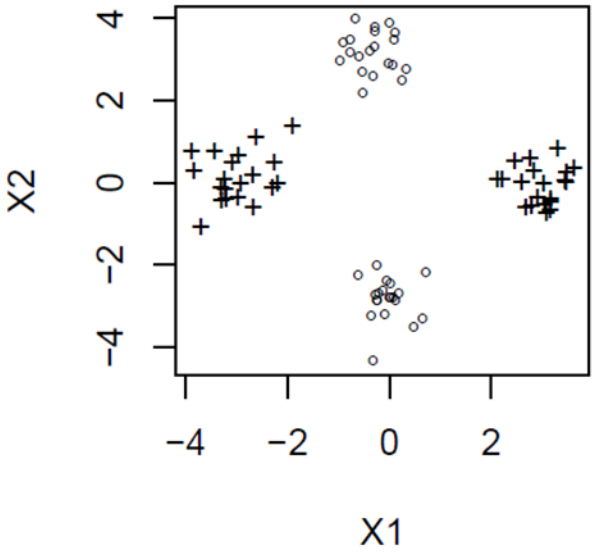
Support vector machine with soft margin



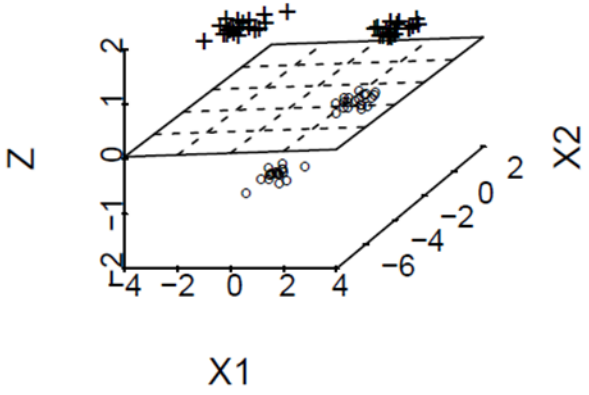
3 Classification Models – Support Vector Machines

- Third idea: Transformation of the problem in a more complex space with higher VC-dimension with respect to linear separation
- Example: Transformation of the XOR-problem into three dimensions

Data



Transformed Data



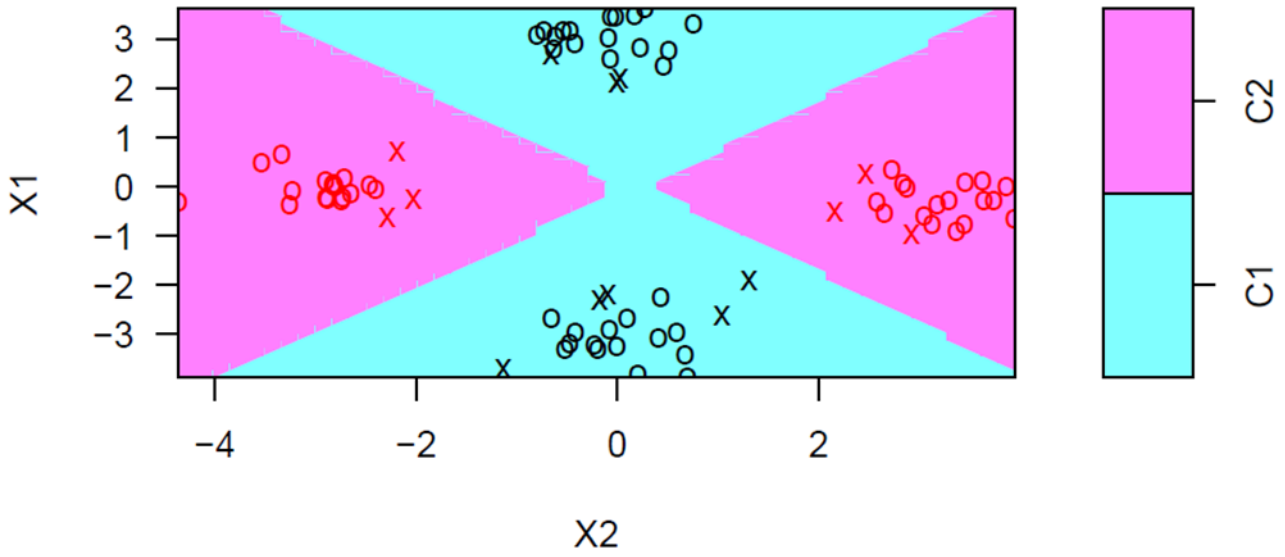
- Realization of the transformation is done by the so-called kernel trick
- Instead of explicit transformation a kernel function is used, which allows the calculation of inner products and distances in the new space
- The most frequently used kernel function for transformation of standard problems is the radial basis kernel

$$H(x, x_1) = exp\left\{\frac{-||x - x_1||^2|}{\sigma^2}\right\}$$

– Example:

- Visualization of the solution for the XOR problem with radial basis kernel in the two dimensional space (crosses mark support vectors)

R package e1071



– Evaluation of support vector machines
  - From a theoretical point of view the solution is of interest because it the empirical risk is directly minimized and a formal estimation of the generalization error is possible
  - The kernel trick allows application to non-numeric data, e.g., classification of graphs (graph kernels), or classification of text data (string kernels)
  - Contrary to logistic regression or trees interpretation of the solution is usually difficult
  - The probability for class assignment must be calculated separately, e.g., in R logistic regression is used
  - Classification of more than two classes can be done by comparing all pairs of classes and taking the majority vote
  - For evaluation of the method k-fold cross-validation can be used

– Modeling task for combination methods
  - Given training data ($\boldsymbol{Y}$,$\boldsymbol{X}$) from two classes labeled by $y_i \epsilon \{-1,1\}$ the goal is to find a classification rule by repeated application of weak classifiers
    - Weak classifiers are classifiers which are only slightly better than a naïve classification without using the input variables which would be 0.5

– Estimation task for combination methods

- The Adaboost algorithm computes the classifier in the following way
  1. Start with a weak classifier $f(x)$ for the training data using equal weights for all observations
  2. Compute modified weights in such a way that misclassified data get higher weight and correctly classified data lower weight
  3. Compute a new classifier $f^*(x)$ for the data with the new weights
  4. Repeat steps 2 and 3 T times and define the final classification function as a combination of the T classifiers

3 Classification Models – Combination Methods

- In most cases classification trees with only few nodes are used as weak classifiers, sometimes called stumps
- Calculation of the weights for the data and weighting of the different classifiers uses the concept of exponential loss of a classifier defined by

$$L_{exp} = exp\big(-yf(x)\big), y\epsilon\{-1,1\}$$

– Evaluation of combination methods

- The method allows theoretical estimation of the generalization error and has similarities with logistic regression
- Generalization for more than 2 classes is achieved by one versus the rest

3 Classification Models – Summary

- Goal in classification is learning a decision rule for class membership from training
- The decision rules are computed either directly or from probabilities for the classes
- Most important tool for evaluation of the solution is the confusion matrix
- In the case of two classes the ROC-curve is a useful tool for evaluation
- Independent from the chosen method the general analysis template defines the workflow for classification and appropriate tools for data understanding and data visualization must be used
- Some methods are developed for two classes and generalization for problems with more than two classes must be done

3 Classification Models – Summary

- Classification methods use different formal structures and the selection of the best model in a model class depends on the specification of the class
  - For some methods like logistic regression or CART integration in the estimation process is possible
  - For methods like support vector machines or combination methods k-fold cross-validation is usually applied
- Classification trees have an option to handle missing data in the estimation process

Contents

4 Introduction to Unsupervised Learning

- In unsupervised learning the data contain no output variable and the goal is to find a group structure in the data, such that the observations in the groups are rather homogeneous
- The groups are called clusters
- The available data are N observations of p variables:

$$X = (X_1, X_2, \ldots, X_p)$$

  Observation vectors are denoted by $x = (x_1, x_2, \ldots, x_p)$

- Two different approaches for construction of clusters:
  - Distance based clustering
  - Model based clustering (not considered in the slides)

– Definition of distances

- For metric variables the most important distance is the Euclidean distance

$$d(x,z) = \sum_{i=1}^{p} (x_i - z_i)^2$$

- Alternative distances:

  Absolute deviation: $d(x,z) = \sum_{i=1}^{p} |x_i - z_i|$

  Maximum distance: $d(x,z) = \max_{p} |x_i - z_i|$

- For binary variables the Hamming distance defined by the number of different values in the variables is frequently used
  - The Hamming distance is equivalent to the Euclidean distance

- Another measure frequently used is the *cosine similarity* defined by the angle between two vectors:

$$sim(x,z) = |cos(x^T z)|, \qquad d(x,z) = 1 - sim(x,z)$$

  - From statistical point of view the cosine similarity measures the correlation between two observation vectors
  - Cosine similarity is frequently used if the vectors are binary vectors
- For categorical (qualitative) variables one can define indicator variables for each attribute value and calculate distances for the binary variables
- Calculation of distances between objects characterized by variables with different scales needs special attention
  - Realized for example in the procedure daisy in R (package cluster)

- For more complex structures like string variables (text) or graphs the distance calculation can be based on kernels
  - String kernels: based on counting the simultaneous occurrence of substrings of certain length

– Standardization of variables

- In many it is useful to consider standardized
- Possible standardizations:
  - All variables are standardized with mean zero and variance 1
  - All variables are standardized such that the values are in the interval [0 , 1] , or [-1 , 1]

Analysis template for Cluster analysis

- *Relevant Business and Data*: Customer behavior represented by cross-sectional data for process instances with a matrix $\boldsymbol{X}$ of explanatory variables
- *Analytical Goals:*
  - Find a segmentation of the data into clusters which allows interpretation from business understanding
  - Determination of representatives for the clusters
- *Modeling Task:*
  - Definition of a model for the data either based on distances or on a mixture model

- *Analysis Task:*
  - Random data splitting into training and test data
  - Model estimation for the cluster solutions
  - Model assessment based on homogeneity of the clusters, and separation between the clusters
  - Model selection based on the number of clusters
- *Evaluation and Reporting Task:*
  - Using test data evaluate the solution with respect to validity and reliability

Contents

- From the numerous methods for clustering the following methods are discussed:
  - Hierarchical clustering, agglomerative methods
  - Partitioning clustering, k-means
  - Model based clustering
  - Other frequently used general methods are:
    - DBSCAN which groups points according to their density measured by the number of nearest neighbors
    - Self organizing maps (SOM) which are a special type of neural nets and can be interpreted as k-means clustering defined on a distorted grid
    - Two stage clustering combine the ideas of hierarchical clustering with the ideas of k-means (IBM/SPSS) by using a cluster-feature tree
    - Model based clustering which estimated a mixture distribution for the data using the EM-algorithm

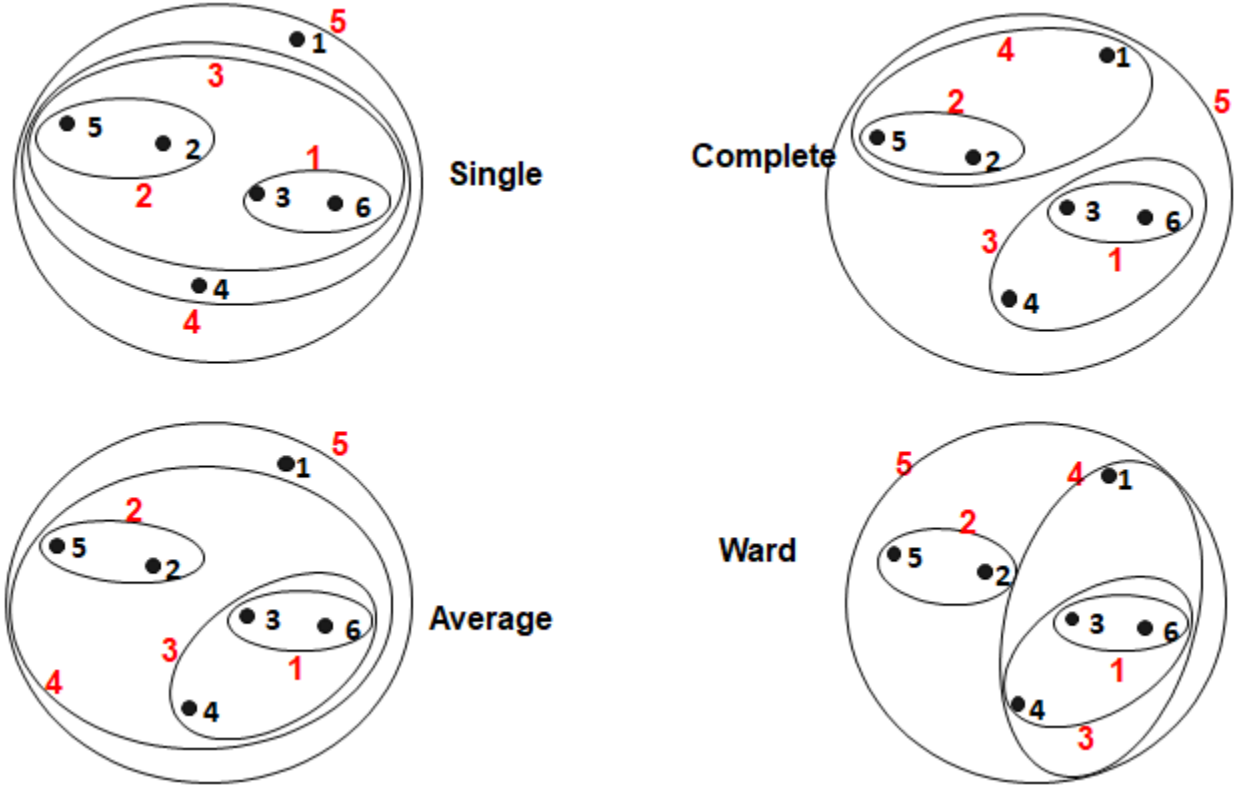
- Modeling task in hierarchical clustering:
  - Given the training data $\boldsymbol{X}$ hierarchical agglomerative clustering defines stepwise a cluster tree using a bottom up method:
    1. Define N initial clusters by the observations $N_{cl} = N$
    2. **for** $k = 1$ **to** $N - 1$ **do**

       merge clusters $C_r \; and \; C_s \; if \; d(C_r, C_s) = \min_{(i,j)} d(C_i, C_j)$

       $N_{cl} = N_{cl} - 1$
  - Main modeling activity is the definition of the distance between the clusters based on the distance between the objects, and determination of number of clusters
  - The distance between the clusters is called the linkage

- Different specifications for linkage
  - Single linkage: Distance between clusters is the distance of the closest points (minimum spanning tree)
  - Complete linkage: Distance between clusters is the distance of the farthest points
  - Average linkage: mean distance between all the points in the two clusters
  - Ward distance: difference between the total within cluster sum of squares for the two clusters separately, and the within cluster sum of squares resulting from merging the two clusters in one cluster
- In general complete linkage, average linkage, or Ward's method are recommended
- Single linkage usually not recommended because the clusters have often a chain form and are not spherical

- Visualization of the different linkages for 6 observations



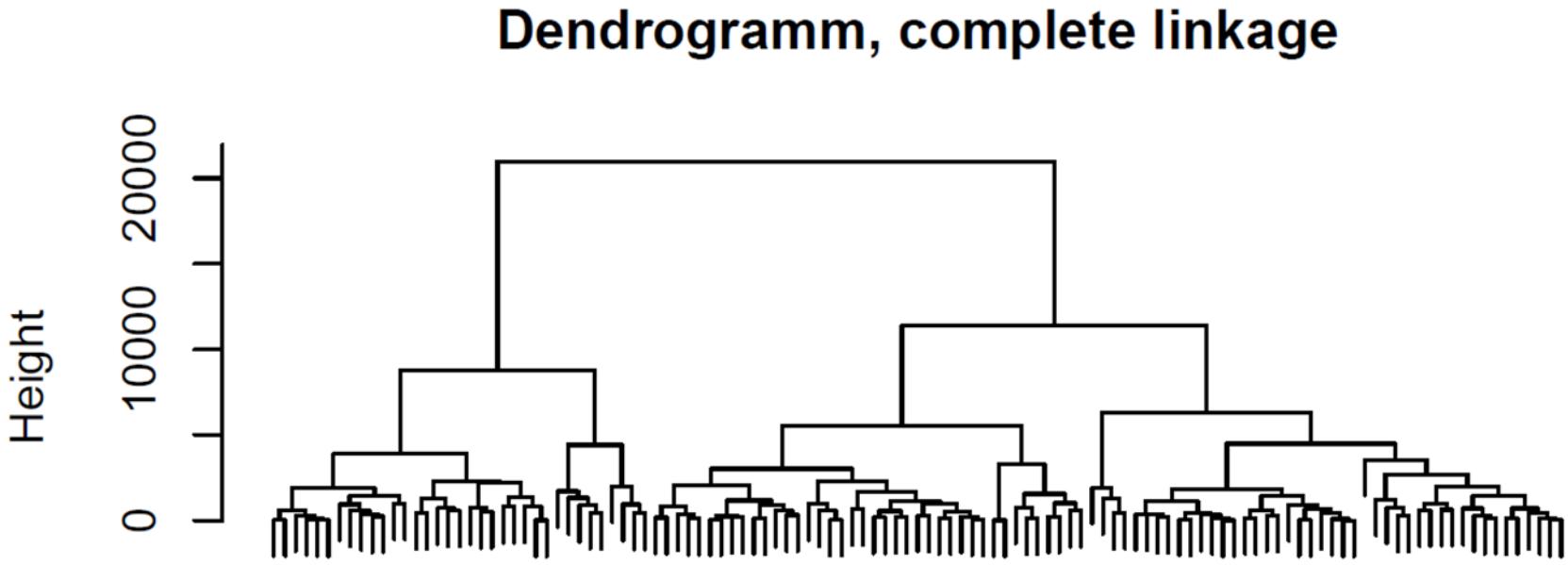
– Estimation task in hierarchical clustering:
  - The main estimation task is specification of the number of clusters
  - Most popular method is using the dendrogram which gives a visual representation of the aggregation process as a tree:
    - The leaves of the tree are defined by the objects
    - Other nodes are formed according to the aggregation process
    - The heights of branches is defined by the distance between the clusters

  - An alternative to the dendrogram is a scree plot of the distance between the merged classes in dependence of the number of classes
    - The decision about the number of classes is defined by the elbow of the scree plot

3 Clustering Algorithms – Hierarchical Clustering

- Example of a dendrogram

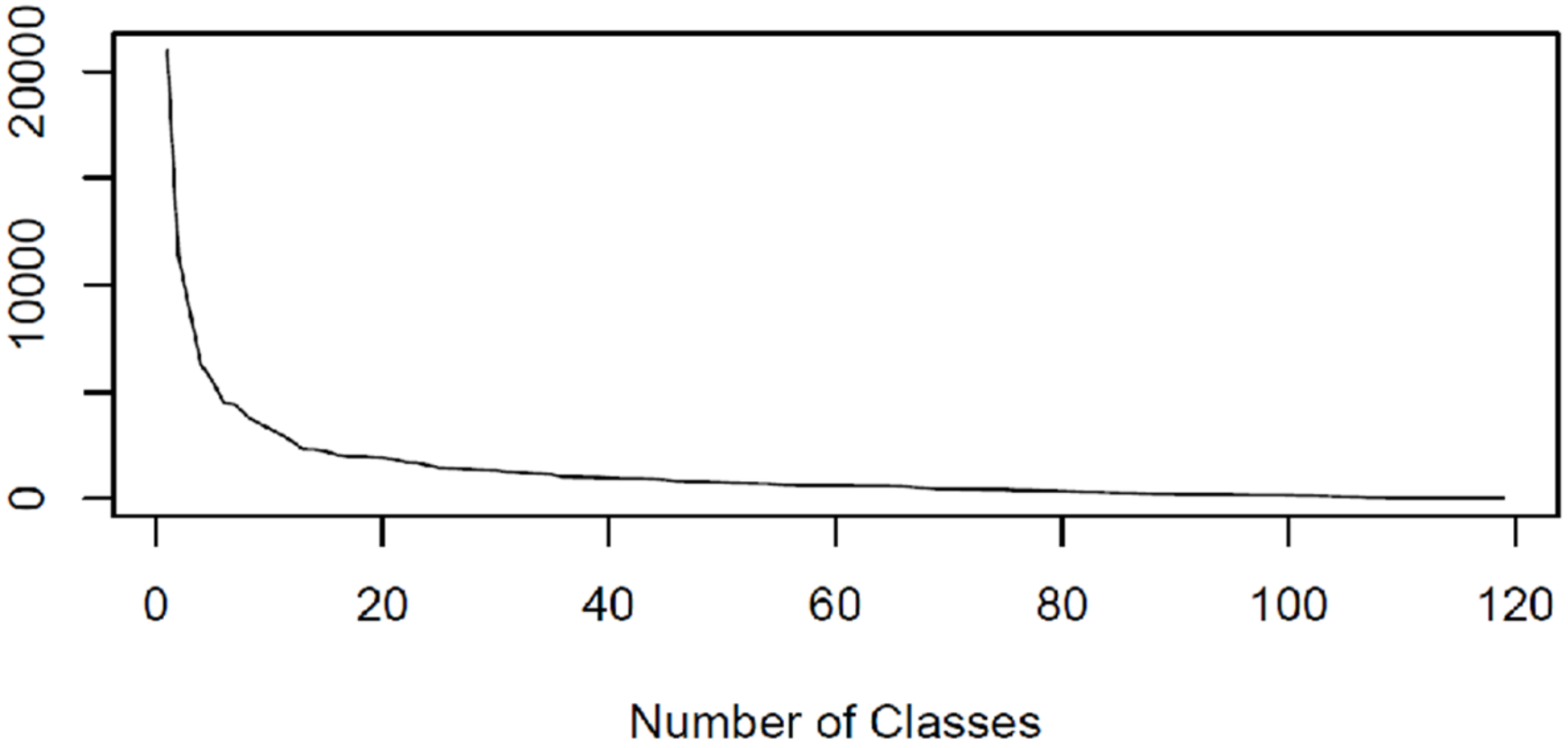
R package cluster



3 Clustering Algorithms – Hierarchical Clustering

- Example of a screeplot for clusters

R package cluster



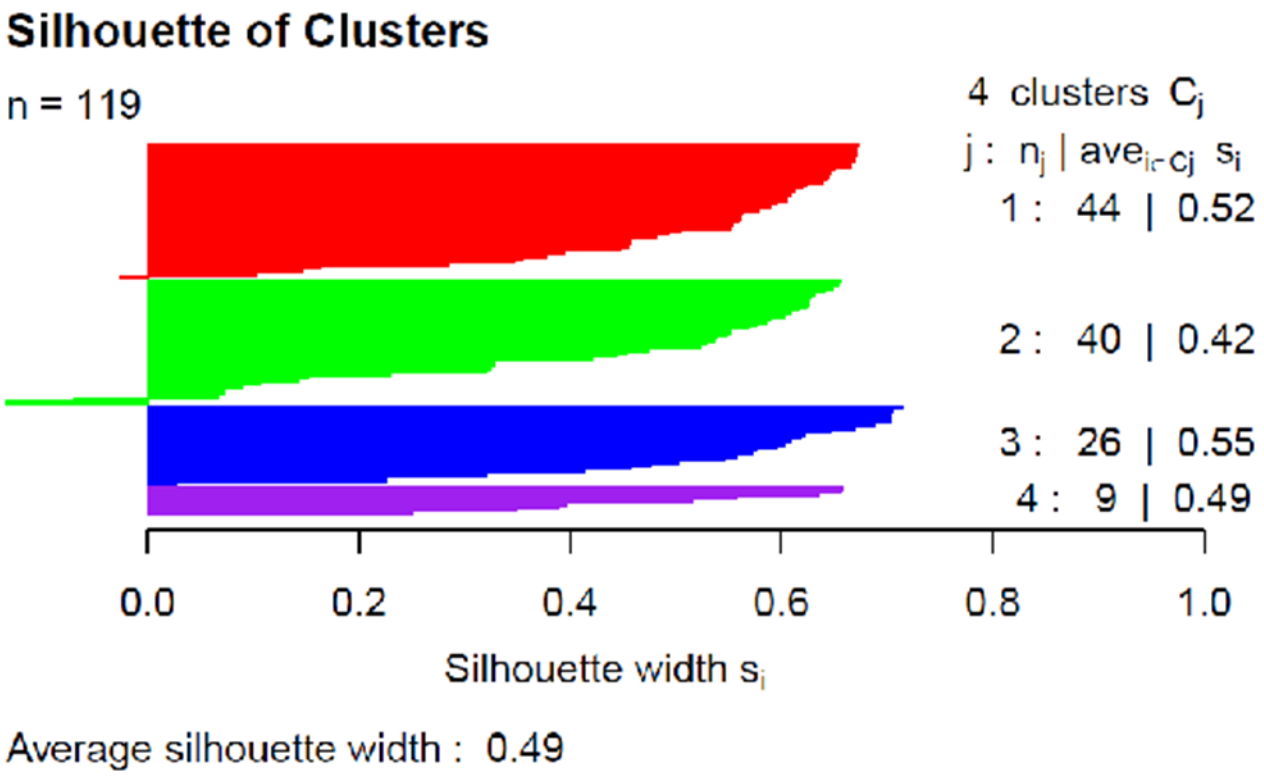
– Model assessment in hierarchical clustering:

- Model assessment uses descriptive methods for analyzing the homogeneity of the cluster, e.g., distribution of the variables within the clusters and grouped boxplots for the variables in the clusters
- For showing the differences between the clusters a frequently used method is representation of the data in a principal component plot
- An alternative to principal components is using a representation based on multidimensional scaling

- A specific visualization method for assessment of the cluster solution is the silhouette plot
  - The silhouette shows for each point how well the point is located in the cluster
  - A value close to 1 shows that the point fits well to the cluster
  - A negative value indicates that the point is not well assigned
  - The silhouette can also be used for comparing different cluster solutions

- Example of a silhouette plot

R package cluster



– Evaluation of hierarchical clustering:
  - Hierarchical clustering allows an intuitive visualization of the solution
  - Decision about the number of clusters can be done after analysis
  - Hierarchical clustering is limited to small number of observations
  - The decision about the cluster assignment cannot be changed in the algorithm
  - Hierarchical clustering supports the analysis of the validity of the solution with respect to a subject matter explanation

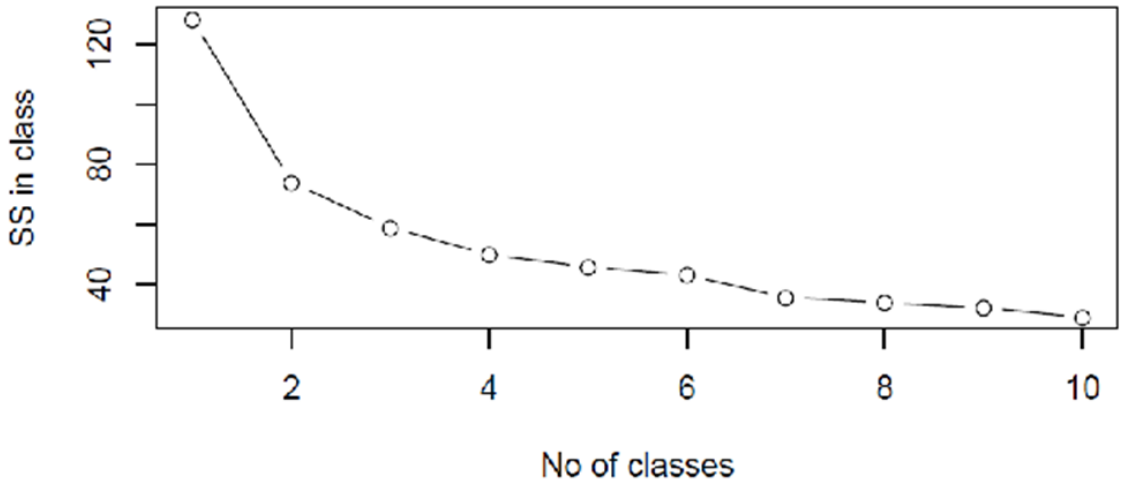
– Modeling task in K-means clustering:
  - Given the training data $\boldsymbol{X}$ K-means clustering defines a cluster solution with K clusters for the data in the following way:
    1. Define initial solution for cluster centers $(c_1, c_2, \ldots, c_K) \quad = N$
    2. Assign each observation to the cluster with center closest to the observation
    3. Compute new centers for the clusters as means of the assigned observations
    4. Repeat steps 2 and 3 as long as there is a significant change in the centers
  - Main modeling task is the determination of the number of clusters and definition of the initial cluster centers

– Estimation task in K-means clustering:

- Determination of the number of clusters can be based on a visualization of the sum of squares within the clusters for a solution against the number of clusters
- Using this plot the number of clusters can be determined by the elbow criterion

Example of a scree plot for K-means clustering

R package cluster



- With respect to the initial solution the standard procedure is choosing the centers randomly and trying different solutions

– Model assessment in K-means clustering

- Model assessment in k-means clustering uses the same techniques as hierarchical clustering

– Evaluation of K-means clustering

- K-means clustering is a fast method from computational point of view
- K-means clustering can be applied to large data sets
- New observations can be assigned easily to the clusters (cf. k-nearest neighbor classification)

5 Unsupervised Learning – Summary

- Goal in unsupervised learning is finding groups in data and characterization of the groups by a group representative
- Most cluster methods use the definition of a distance between observations
- Standardization of variables should be taken into account
- The number of clusters can be determined by descriptive analysis of the cluster solution
- For validation of a cluster solution interpretation from a business perspective is important

Contents

6 Summary & Outlook

- In data mining for cross-sectional data considers two different types of problems: supervised learning based on input data which explain the output data, and unsupervised learning using only potential explanatory input data
- Goals in supervised learning are predictive goals and are formulated either as regression problems or classification problems, goals in unsupervised learning are more descriptive goals for grouping data in homogeneous groups
- For supervised learning the definition of a loss function and the evaluation of the predictive power of the learned prediction using test data is essential (balancing between overfitting and underfitting)
- Most methods in unsupervised learning are based on the definition of a distance between the observations and assessing the validity of the cluster solution from a business perspective is essential

Refernces

- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer
- Linoff GS, Berry MJA (2011) Data mining techniques for marketing, sales, and customer relationship management. Wiley
- Marsland S (2009) Machine learning – an algorithmic perspective. CRC Press
- Wu X, Kumar V (2009) The top ten algorithms in data mining. CRC Press

6 Summary & outlook