Business Intelligence
SS 2017

Cross-sectional Analysis 1

W. Grossmann

Content

- Types of Learning Problems
- Evaluation of a Learning Method
  - Supervised Learning
  - Model Selection in Supervised Learning
- Evaluation in Unsupervised Learning

Types of Learning Problems

- Starting point for learning are data from variables with two different interpretations:
  - Explanatory Variables : $X = (X_1, X_2, \ldots, X_p)$ also called input variables or predictors
  - Explained variable: Y also called output variable or response variable
- Notation for data from N observations:

$$Input : (x_1, x_2, \ldots, x_N)$$

$$Output : (y_1, y_2, \ldots, y_N)$$

  - Note that each input observation is a p-dimensional vector

Types of Learning Problems

- Our goal is to “learn” a function that allows prediction of the output variable from the input variables
- We distinguish different types of Learning Problems
  - Regression Problems
  - Classification Problems

Types of Learning Problems
Regression Problems

- Regression Problems: Goal is prediction of a quantitative output variable from the input variable by a function

$$Y = f(X)$$

  - Example Used Car Prices: Predict the price of the car from variables like age or mileage, etc.

Types of Learning Problems
Classification Problems

- Classification Problems: In this case the output is a class identifier, and we want to find a rule which allows for given input variables the decision about the class to which the data belong. Formally we denote the decision as function taking as values one of the group identifiers

$$Y = Group = f(X)$$

  - Example Credit Data: Find a rule, which allows a decision whether a credit defaults (Class 1) or not (Class 2) based on attributes for the customer like age, sex, family status, etc.

Types of Learning Problems
Classification Problems

- Two approaches towards classification problems:
  - Use an algorithm which determines the class membership

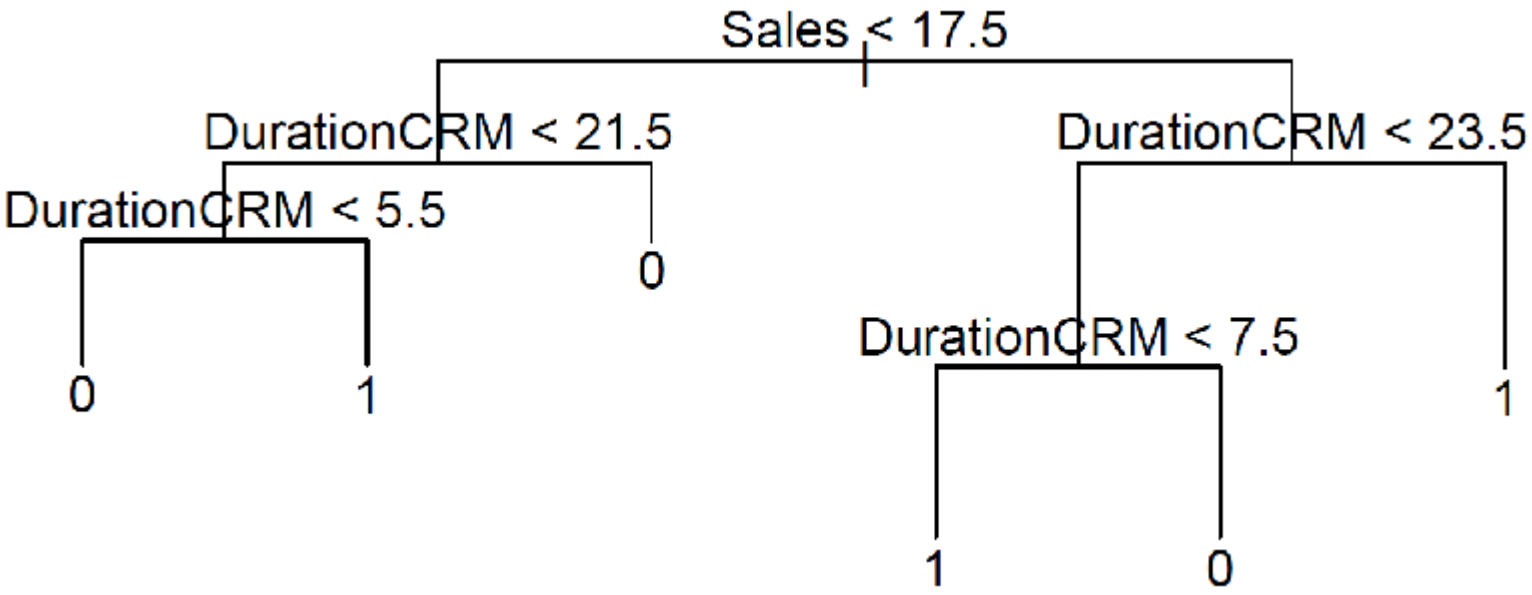
Example Decision trees:

Given the information from 11 customers about usage (1 = yes, 0 = no) of a certain service, find a rule which allows predict of usage for customer 12

CR-Dur	Sales	User Type	UseService
10	12	private	yes
24	36	business	yes
28	48	business	yes
45	20	private	yes
30	34	private	yes
3	21	private	yes
1	5	business	no
23	23	business	no
12	49	business	no
35	12	private	no
33	15	private	no
12	25	private	??

Types of Learning Problems
Classification Problems

Decision tree:



Types of Learning Problems Classification Problems

- Use an algorithm which determines the probability of class membership

  Example: Given 24 observations from customers from which 12 have quit their relation to the company in the last 12 month and 12 are still customers, learn a function which allows the computation of the probability that a customer quits the relation.

  Available information:

  - Duration of the relationship
  - Activity of the customer
  - User Type (private or office user)

Types of Learning Problems
Classification Problems

Data

Duration	ActInd	UserType	Quit	Duration	ActInd	UserType	Quit
5.63	1.93	office(0)	no (0)	6.43	7.6	office(0)	yes(1)
6.39	9.47	office(0)	no (0)	5.55	3.53	private(1)	yes(1)
5.31	9.23	office(0)	no (0)	6.68	3.6	private(1)	yes(1)
5.76	11.67	office(0)	no (0)	3.35	0.23	private(1)	yes(1)
7.12	8.9	office(0)	no (0)	4.31	0.53	private(1)	yes(1)
8.13	9.9	office(0)	no (0)	2.06	2.33	private(1)	yes(1)
4.1	7.27	office(0)	no (0)	3.03	2.5	private(1)	yes(1)
4.29	10.8	office(0)	no (0)	4.78	5.37	private(1)	yes(1)
1.55	4.97	office(0)	no (0)	5.89	1.13	private(1)	yes(1)
0.81	7.2	office(0)	no (0)	4.78	3.83	private(1)	yes(1)
5.25	9.0	private(1)	no (0)	3.83	1.47	private(1)	yes(1)
4.26	8.57	private(1)	no (0)	1.25	2.87	private(1)	yes(1)

Learned Function:

$$\text{logit}(Quit) = \log \frac{P(Quit=1)}{P(Quit=0)} = 1.385 + 3.058 UserType - 0.75 ActInd$$

Types of Learning Problems
Classification Problems

- Decision Rule:
  - Quit = yes if logit(Quit) < 0
  - Quit = no if logit(Quit) > 0

- In practice, both solutions are usually calculated

Types of Learning Problems Clustering

- Clustering : In this case the data contain no output variable and we want to find a group structure in the data, such that the observations in the groups are rather homogeneous with respect to the variables
  - Example: Different countries have different eating habits and consume protein in different form e.g. red meat, white meat, eggs, milk. Given the data about the average amounts of staple food per person per year in different countries, can we find groups of countries with similar consume of staple food? For example Mediterranean countries, northern countries,….

Types of Learning Problems Association Analysis

- Association Analysis: In this case we are interested in associations between consumer choices and we want to learn frequent patterns of consumptions
  - Example: Persons purchasing in a supermarket different goods (market basket). Can we learn rules how the likelihood of purchasing one good is increased by purchasing another good. For example buying beer increases the likelihood for buying snacks. (market basket analysis)

Types of Learning Problems supervised and unsupervised learning

- We call the first two learning problems (regression and classification) ***supervised learning***, because the data informs us about the values of the output variable given the input variables
- The third and the fourth problem (clustering and association analysis) are examples of ***unsupervised learning*** problems, because there is no output variable in the data which can be learned
  - Unsupervised learning is more descriptive and aims at learning a structure in the data

Evaluation in Supervised Learning

- In case of supervised learning the data offer a benchmark which allow a formal evaluation of a learning method, i.e. , we compare the quality of our learning method with the observed output in the data

Evaluation in Supervised Learning

- Evaluation of supervised learning is based on the concept of loss and risk
- If we learn a function $\hat{f}(x_i)$ for prediction of the observed output y_i, the ***loss*** measures the deviance of the learned value from the observed value
  - Notation: $L(y_i, \hat{f}(x_i))$
- The risk is defined as expected loss:

$$Risk(Y, \hat{f}) = E[L(y_i, \hat{f}(x_i))]$$

Evaluation in Supervised Learning Regression

- In case of regression the loss is usually defined by the squared loss

$$L(y_i, \hat{f}(x_i)) = (y_i - \hat{f}(x_i))^2$$

- Due to the fact that we do not know reality exactly we cannot calculate the risk exactly but have to use the empirical risk

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

Evaluation in Supervised Learning Classification

- In case of classification the loss is usually defined by the 0-1 loss function

$$L(y_i, \hat{f}(x_i)) = \begin{cases} 1 & \text{if decision about class is wrong} \\ 0 & \text{if decision about class is correct} \end{cases}$$

- Sometimes we use instead of 1 an 0 values which reflect the cost of misclassification
  - Example: Two classes, costs for wrong decision for class 1 is 5 times as costly as wrong decision for class 2
  - Loss function:

$$L(1, \hat{f}(x_i) = 0) = 5, L(2, \hat{f}(x_i) = 1) = 1$$

Evaluation in Supervised Learning Classification

- The empirical risk is usually represented 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 obtained by multiplying the coefficients with the corresponding costs

Evaluation in Supervised Learning Classification, 2 classes

- Confusion matrix and notation in case of classification problems with two classes

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)	

Evaluation in Supervised Learning Classification, 2 classes

Other terms used:

Precision = Positive predictive value

Recall = Sensitivity

False Positive Rate (False Alarm) = 1 – Specificity

False Discovery Rate = 1 – Precision

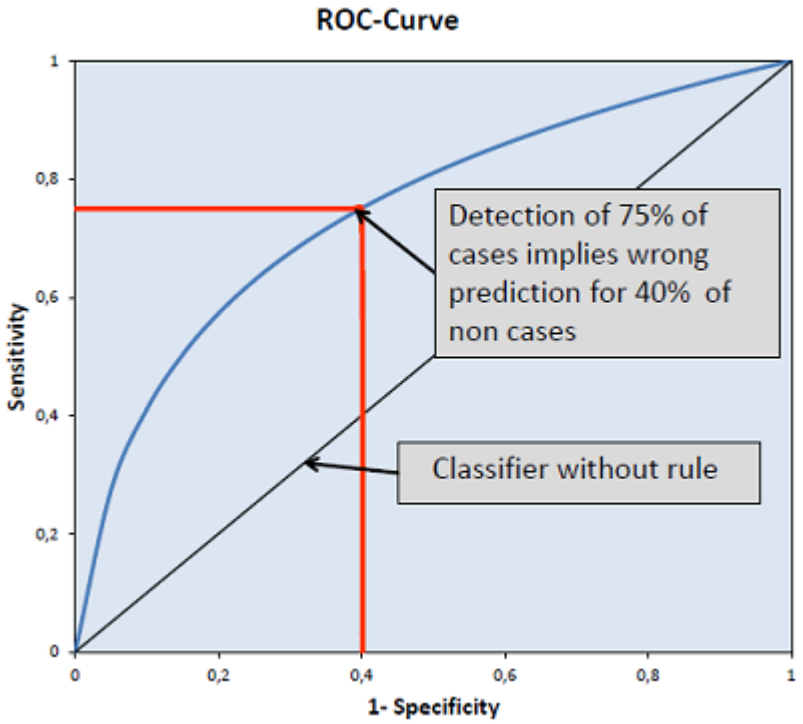
Accuracy = (TP + TN)/(TP + TN + FP + FN)

Evaluation in Supervised Learning Classification, 2 classes

- If we base our decision on a threshold probability for the class membership we can parameterize sensitivity and specificity with this threshold and visualize the decision rule with the Receiver-Operator characteristic (ROC curve)
- The ROC-curve is a plot of the sensitivity against 1- specificity for 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 area under the curve would be 0.5

Evaluation in Supervised Learning Classification, 2 classes

- Example: ROC-curve with area 0.73



Evaluation in Supervised Learning and Model Complexity

- Due to the fact that we have not complete information but only the data in the sample, we cannot calculate the risk from the data but only the empirical risk

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

Evaluation in Supervised Learning and Model Complexity

- The risk minimization considers only the question how well the method works for the training data
- If we apply the method for new data we want to know the prediction quality of the model for new data
  - Such new data are called test data
- Prediction quality for test data is measured by the Generalization error

Evaluation in Supervised Learning and Model Complexity

- The generalization error depends on
  - The number of available training data
  - The complexity of the model
    - A models may be simple, for example we use for prediction of a regression problem the mean, or it may be rather complex using many variables for prediction
    - In extreme case we use the data itself as prediction function
- A useful model must balance these two effects (Bias-Variance Trade off)

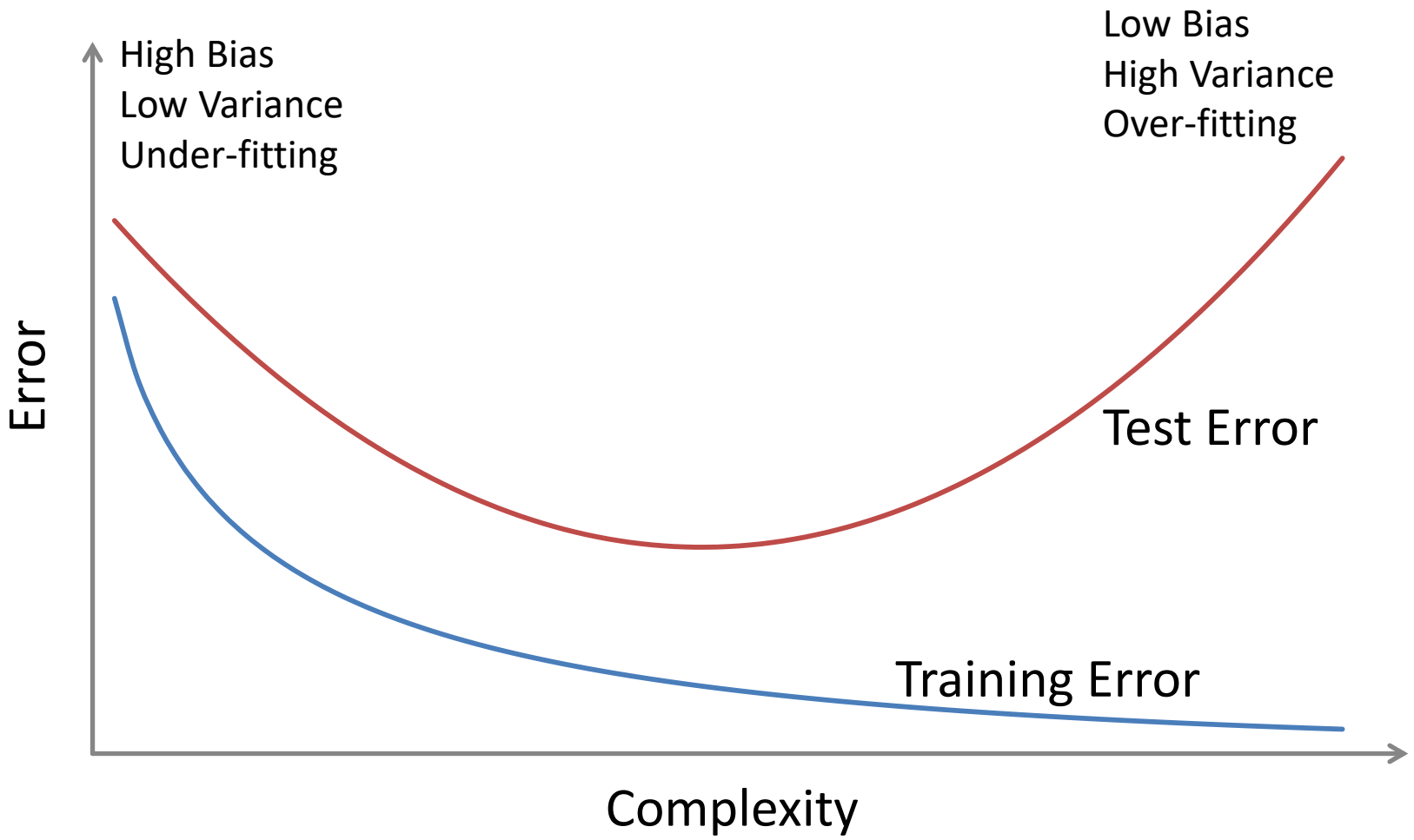
Model selection in Supervised and Model Complexity

- Minimizing the empirical risk for evaluation of a model is not a good strategy for evaluation of the predictive power of a model, because we have to take into account the effect of under-fitting and over-fitting
- ***Under-fittin***g: If we estimate a simple model statistics tells us that such a simple model can be estimated with high accuracy from the data, but the model describes reality not sufficiently, we "underfit" the data
- Consequently we have a large model bias and the results of prediction for new cases are poor, i.e. we have a large test error

Evaluation in Supervised Learning and Model Complexity

- ***Overfitting:*** If we estimate a complex model, the model would fit the data quite well, but accuracy of the estimate would be rather low because too many parameters are estimated with high variance, we "overfit" the data
- Consequently we have a rather large variance and the results of prediction for new cases are also poor, i.e. we have a large test error
- The following figure shows this problem, also known as bias-variance trade-off schematically

Evaluation in Supervised Learning and Model Complexity



Evaluation in Supervised Learning and Model Selection

- In practice we do not know
  - Which attributes are of importance for the model
  - What is the functional form of the relationship between input variables and output variables
- Consequently supervised learning has to solve three goals simultaneously
  - Find the parameters for the models under consideration which minimize the empirical risk
  - Selection the best model within the models under consideration
  - Estimation of the generalization error

Evaluation in Supervised Learning and Model Selection

- The standard procedure for solving these tasks is splitting the data and proceed as follows:
  - Use a one part of the data for estimating parameters of the models under consideration. Frequently , using 50% of the data is recommended.
  - Use a second part of the data for selecting the appropriate model. Frequently, using 25% of the data is recommended.
  - Use the third part of the data for estimation of the test error. Frequently , using 25% of the data is recommended.

Evaluation in Supervised Learning and Model Selection

- In this context the partition of the data are called ***training data, validation data, and test data***
- Many times the first and the second step can be done simultaneously using theoretical considerations for validation
  - Variable selection in case of linear regression based on Akaike Information criterion (AIC)

$$\min_{p} AIC(p) = -2\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + p/N$$

  - AIC measures information loss compared to true model

Evaluation in Supervised Learning and Model Selection

- AIC is a special case of the penalization principle
- Penalization methods use instead of the risk a combination of the risk and a complexity measure

$$R_{pen}(Y, \hat{f}(X)) = R_{emp}(Y, \hat{f}(X)) + \alpha \cdot comp(\hat{f}(X))$$

Evaluation in Supervised Learning and Model Selection

- A data based method for model selection is k-fold cross validation

```
1 begin
2 |   Divide the training data into k disjoint samples of roughly equal size;
3 |   For each validation sample use the remaining data to construct the
  |   estimate and estimate the empirical risk for the left out data;
4 |   Compute the prediction error by averaging the empirical risk of the
  |   validation data;
5 end
```

Evaluation in Unsupervised Learning

- In case of unsupervised learning we have no output data and use more descriptive methods for evaluation of our learning method
- Hence we cannot define a loss function and a risk
- The goal is definition of formal criteria for model reliability and model validity

Evaluation in Unsupervised Learning

- In case of Clustering the evaluation of a computed group structure is usually defined by a homogeneity measure applied to the data
  - In case of quantitative variables a standard approach for measuring homogeneity is looking at the decomposition of the sum of squares of deviations from the group means well known from analysis of variance

$$SST = SSW + SSB$$

Sum of squared deviations from the overall mean =

Sum of squared deviations from group means within groups +

Sum of squared deviation of group means from overall mean

Evaluation of Unsupervised Learning

- In case of association analysis an association rule can be evaluated according to the frequency of occurrence and the confidence of the rule
  - These concepts will be discussed in connection with the methods for finding such rules

Evaluation of Unsupervised Learning and Model Selection

- In case of unsupervised learning we have only descriptive procedures for model selection
- But splitting into training data and test data is recommended also in case of unsupervised learning