Data mining with Machine Learning for the social sciences

Introduction, Challenges, the right & the wrong, Misunderstanding

Priv.-Doz. Dr. Stefan Bosse
University of Koblenz-Landau, Fac. Computer Science
University of Bremen, Dept. Mathematics & Informatics
18.5.2018
sbosse@uni-bremen.de

1. Inhalt

2. Introduction to Artificial Intelligence

2.1. Artificial Intelligence

In social science big data volumes must be handled.
But big do not mean helpful or important!
Data is noisy and uncertain!?

- One major task in data science is the derivation of fundamental mapping functions:

$$F(\textit{Input Data})\text{: } \textit{Input Data} \rightarrow \textit{Output Data}$$
$$\Leftrightarrow$$
$$F(\textit{Sensor Data})\text{: } \textit{Sensor Data} \rightarrow \textit{Knowledge}$$

- Such a function F performs **Feature Extraction**
- But often there are *no* or only partial numerical/mathematical *models* that can implement F!
- Usage of Artificial Intelligence and their methods can be helpful to derive such fundamental mapping functions - or at least an approximation: **Hypothesis**
- The **input data** is characterized commonly by a *high dimensionality* consisting of a vector of variables

$$[x_1, x_2, .., x_n],$$

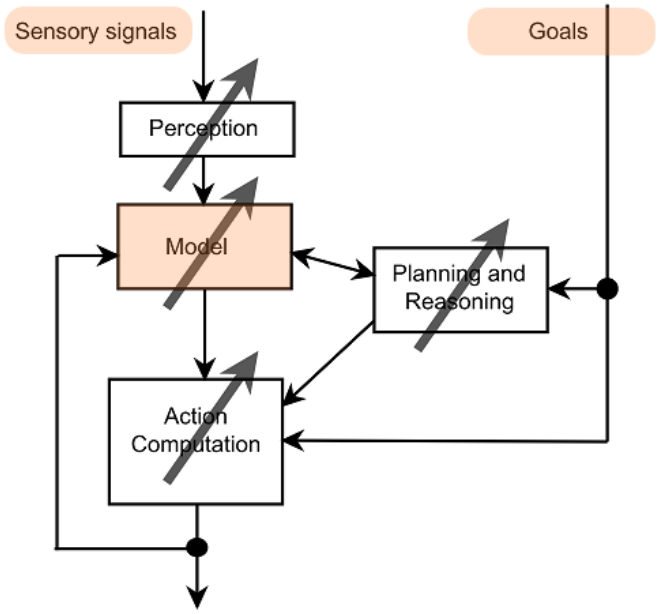
- whereby the **output data** (information) has a much *lower dimensionality* (data reduction!) consisting of the variable vector

$$[y_1,y_2,..,y_m]$$

- This means:

$$F\text{: } R^N \rightarrow R^M \text{ with } M \ll N$$

- **Data reduction** includes the pre-selection of suitable (high information entropy) data variables → **Feature Selection**



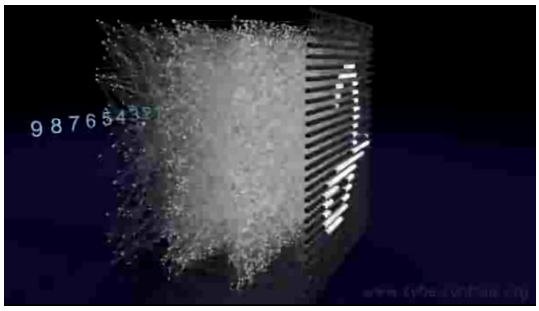
[3]

Fig. 1. A typical Aritificial Intelligence System

2.2. Machine Learning - Technical Sciences

- Often there are *no functional relations* between two variables **x** and **y**. In *technical applications* **x** can be a camera image with 1 Million pixels and **y** a figure from the set {0,1,2,..,9} that represent a hand written character. Generally:

$$f(\mathbf{x})\text{:}\mathbf{x} \rightarrow \mathbf{y}.$$



- *Machine Learning (ML) can be used to derive such relation from experimental/empirical training data!*
- Among the derivation of such functional relations the **prediction** of *what will happen next or in the future* is an important task of Machine Learning

2.3. Machine Learning - The Functional Approach

- Machine learning means the derivation of a **hypothesis** of a simple input-output function from training data provided by humans (statistical data!)!

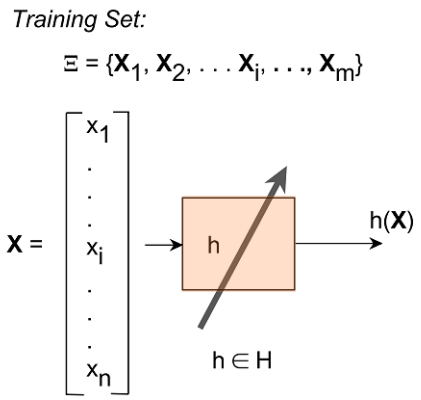


Fig. 2. A hypothesis of an input-output model function derived from training data

2.4. Machine Learning - Medicine

Diagnosis of Appendicitis from medicine and personal data

Input Data x

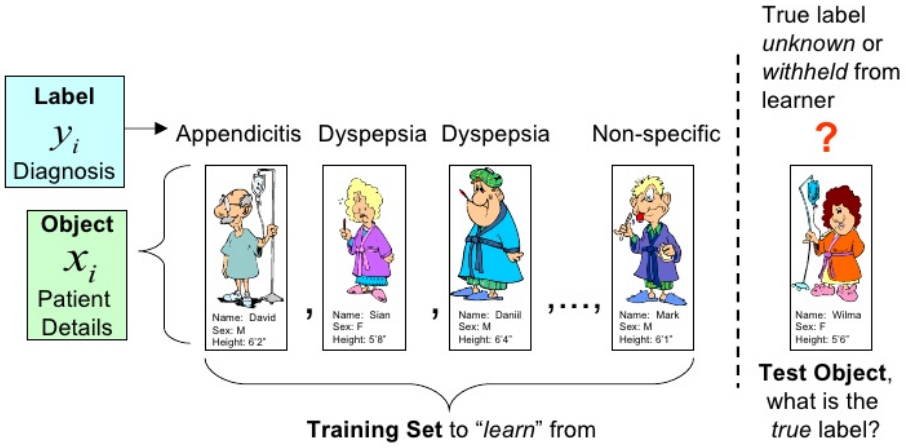
Patient Details [*weight,age,sex,pain left, pain right, temperature, ..*]

Output Data y

Diagnosis Label {*Appendicitis, Dyspepsia, Unknown, ..* }

Decision Learner

Returns one of the labels matching a new input vector **x** (the test object)

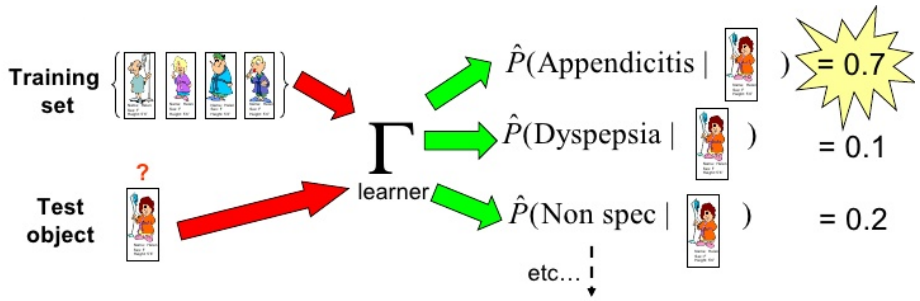


- ➤ Decision classifiers only return one (good or bad) matching label
- ➤ No information about *matching probability*

Probalistic Learner (Bayes Theorem)

Feature: Probability forecast estimating the conditional *probability of best matching* (or all) label(s) with a given observed object **x**

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$



2.5. Machine Learning - Social Sciences

Whats the problem with ML in Social Sciences?

- ➤ Beyond the considerations of effort, even though we can build a hypothesis for some model (e.g., in social science a code) with high accuracy,
  - ❑ the results may not be very informative or reliable from a social scientist point of view
  - ❑ because most ML methods work like a **black box** and do not offer **explanation** for *how decisions* are made.

2.6. Machine Learning - Applications

History of some selected application areas

a. Rule discovery using a variant of the ID3 for a printing industry problem [Evans & Fisher 1992]

b. Electric power load forecasting using a k-nearest-neighbor rule system [Jabbour , K- et al. 1997]

c. Automatic help desk assistant using a nearest-neighbor system [Acorn % Walden 1992]

d. Planning and scheduling for a steel mill using ExpertEase a marketed (ID3-like) system [Michie, 1992]

e. Classification of stars and galaxies [Fayyad et al., 1993]

⇒ Technical Problems!

f. Learning From Crowds - A probalistic approach for supervised learning [Raykar et al., 2010]

- ➤ Machine learning is having a substantial effect on many areas of *technology and science*; examples of recent applied success stories include [6]
  - ❑ robotics and autonomous vehicle control (top left),
  - ❑ speech processing and natural language processing (top left),
  - ❑ neuroscience research (bottom, left),
  - ❑ and applications in computer vision (right).

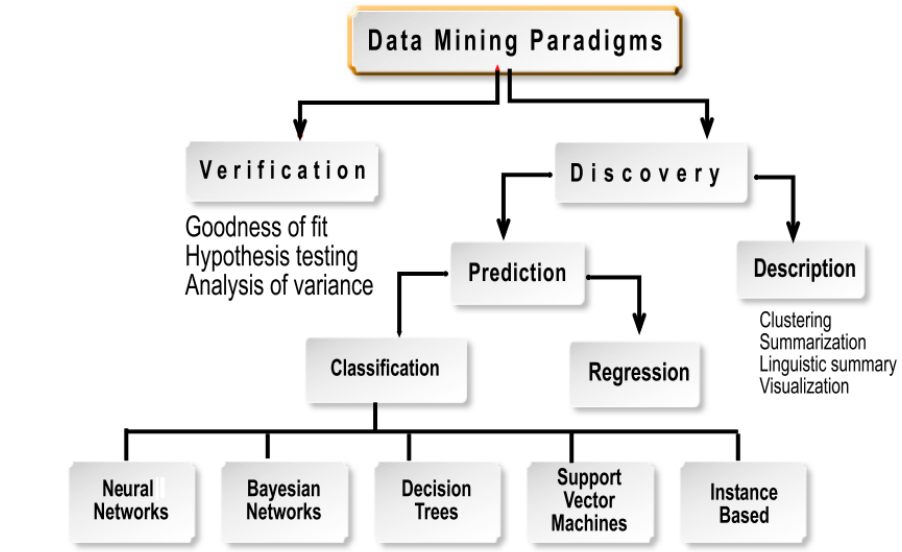
2.7. Data Mining

Data mining (DM) is the more general name given to a variety of computer-intensive techniques for discovering structure and for analyzing patterns in data.

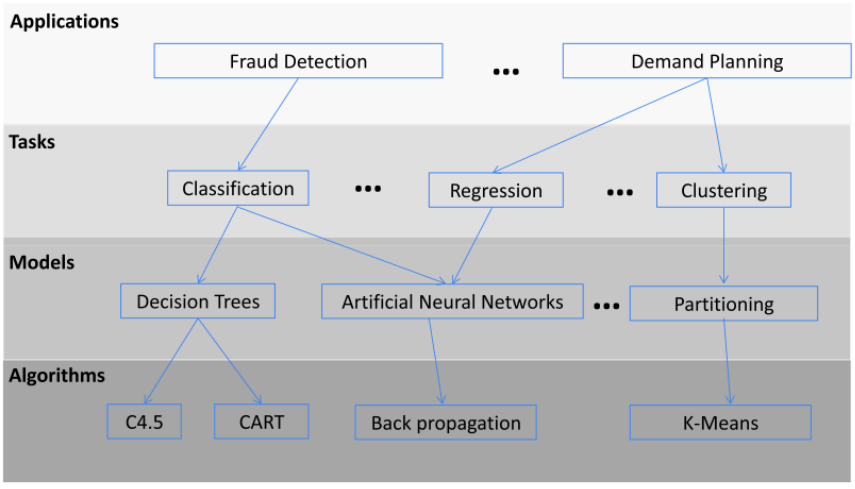
- Using those patterns, DM can
  - create **predictive models**, or
  - **classify** things, or
  - **identify** different groups or clusters of cases within data.
- Data mining, machine learning and predictive analytics are already widely used in business and are starting to spread into social science and other areas of research. [Atte15]
- A partial list of current data mining methods includes:
  - **Association rules**
  - Recursive partitioning or **decision trees**, including CART (classification and regression trees) and CHAID (chi-squared automatic interaction detection), boosted trees, forests, and bootstrap forests
  - Multi-layer **neural network** models and "deep learning" methods
  - Naive **Bayes classifiers** and Bayesian networks
  - **Clustering** methods, including hierarchical, k-means, nearest neighbor, linear and nonlinear manifold clustering
  - **Support vector machines**

- ❏ "Soft modeling" or partial least squares latent variable modeling

Taxonomy [1]



Layers [1]



3. Challenges in Data Mining

3.1. Challenges of Applying Machine Learning to Qualitative Coding

- Coding is an important part of qualitative analysis in many fields in social science.
- Most applications of qualitative coding require **detailed, line-by-line examination of the data**.
- **Such analysis can quickly become very time-consuming even on a moderately sized dataset!**
- Machine learning techniques could potentially extend the principles of qualitative analysis to the **whole dataset** given proper **guidance** from a qualitative **coder**.

Consequently these techniques offer a promising approach to ***scale up the coding process****.*

- *Machine learning* has emerged in the past decades, but its application in qualitative analysis is *still very limited*.
- One common reason: People who use qualitative methods usually do not have **background in ML**.
- The **complexity of selecting features, building models, and tuning parameters**, can prevent the construction of *acceptable models*.

On the other hand, an ML expert might be able to take the codes that a social scientist has applied to part of a dataset *in order to train a classifier to label the* whole dataset*.*

- **However**, since very few ML experts have background in social science, they do not have contextual information to engineer good features and to prevent issues like overfitting.

In addition, machine learning usually performs better on categories that have more instances, but those codes may not be the most interesting to a social scientist.

For example:

To build a good classifier,
we usually need **predefined categories** and a **large amount of corresponding labeled data** (in supervised learning), or the distributions of the datasets must have some **distinct separation** (in unsupervised learning).

➤ **However**, neither of these is the case in coding!

3.2. Labeling

➤ Supervised **Machine Learning requires labeled training data**.

➤ But: Who or how are labels assigned to data sets?

➤ For many supervised learning tasks it may be infeasible (or very expensive) to obtain **objective and reliable labels**.

➤ Instead, **subjective (possibly noisy) labels** from multiple experts or annotators are collected.

➤ In practice, there is a substantial amount of *disagreement* among the annotators, and hence it is of great practical interest to use and improve conventional supervised learning problems for such case.

3.3. Size is not everything

➤ Having **more rows is nice** (increases statistical power) → more training data

➤ Having **more columns is nice**, (estimate heterogenous effects, interactions) → more feature variables

➤ But with *limited research resources*:

❑ If the choice is between creating a counterfactual or more data, the counterfactual will mostly win!

Big but not correlated

- Big Data is usually collected from a **variety of sources with unknown models** how the data was generated.
- It can be **sparse data**: Weak and probably unknown correlation between data variables
- It is statistically variant and **noisy** data!

3.4. Big is not big enough

- Data Mining suggests that the combination of *brute computing power* and *very large datasets* enables data miners to *discover structures* in data?
- Assumption: *Applying conventional statistical approaches to datasets containing smaller numbers of cases do not deliver the structure.* → **Wrong**
- Even the largest social science datasets are **not large enough** to allow a comprehensive or exhaustive search for structure.
  Example: The five-million-person, multi-year census files available from the American Community Survey
- Even the biggest, fastest computers find certain empirical tasks intractable.
- Data mining frequently has to make
  - **Simplifying** assumptions to keep problems tractable, or
  - Select **subsets** of variables,
- Data Mining cannot handle all the available measures in one model!
  It approximates and make compromises!

3.5. Crowd Sensing

- *The "wisdom of the crowd" effect refers to the phenomenon that the mean of estimates provided by a group of individuals is more accurate than most of the individual estimates.*
- It is well-known that in many forecasting scenarios [4], **averaging** the forecasts of a set of individuals yields a collective prediction that is **more accurate** than the majority of the individuals' forecasts—the so-called "wisdom of crowds" effect (Surowiecki 2004).
- However, there are smarter ways of aggregating forecasts than simple averaging.

- A **hybrid approach of (re)calibrating and aggregating probabilistic forecasts** about future events provided by experts who may exhibit systematic biases can improve model quality (Turner, 2010) → *such as overestimating the likelihood of rare events*
- *For many situations, such as when estimating the opinions of a group of individuals, it is desirable to learn a classification model but there is no underlying ground truth!*

4. Important stages in Data Mining

Having introduced the paradox of too little big data and noted the challenges caused by high-dimensional data, we can now discuss how a DM analysis typically proceeds.

4.1. Data Preprocessing

- Raw sample data has to be preprocessed. But this task requires sometimes statistical knowledege of the input data (noise …).

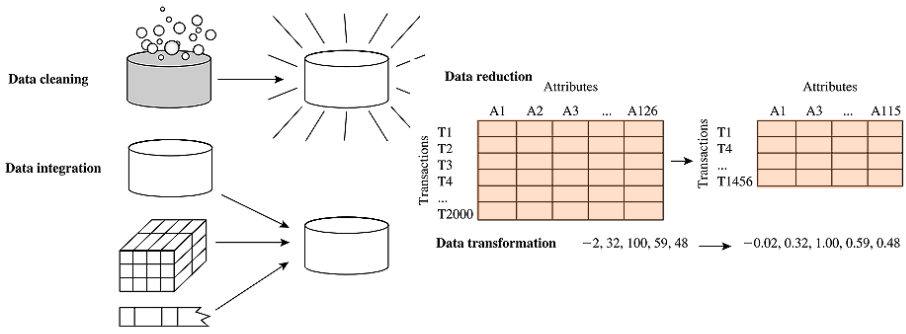


Fig. 3. Different forms of data preprocessing can be applied to raw measuring data

4.2. Feature Selection

- **Feature-selection** methods allow a researcher to **identify** which among many potential predictors those are **strongly associated with an outcome variable of interest**.
  - They also help avoid problems with multicollinearity among predictors.
- DM offers several alternatives for selecting a subset of independent variables that are the most effective predictors of a dependent variable.
  - **Manual variable selection**
  - **Automatic variable selection**

4.3. Feature Selection - Some Methods

Principle Component Analysis

- One example for a method used in dimesionality reduction of the input data → Feature Selection
  - But it can handle basically only two-dimensional data!
- PCA can be applied
  - to ordered and unordered attributes, and
  - can handle sparse data and skewed data.
- Multidimensional data can be handled by *reducing the problem to two dimensions*. Principal components may be used as inputs to multiple regression and cluster analysis.

Attribute Subset Selection

- Data sets for analysis may contain hundreds of attributes, many of them may be *irrelevant* to the mining task or *redundant*.
- Attribute subset selection **reduces the data set** size by **removing irrelevant** or **redundant attributes** (or dimensions).

4.4. Information Entropy

Feature selection using entropy

- Information entropy of a data set, i.e., value set c (column of data) of a variable x_i with $|o|$ as the number of values of a set o (x=v $\in$ V), p_v the occurence probability of a specific value v, and N a set of occurence counts:

$$entropy_N(N) = \sum_{i=1}^{n} -p_i \log_2 p_i, \text{with } p_i = \frac{N_i}{\sum N}$$

$$entropy(c) = \sum_{v \in V(c)} -p_v \log_2 p_v, \text{with } p_v = \frac{|c_v|}{|c|}$$

- Information entropy of associated data sets (x $\rightarrow$ y, x=v $\in$ V, y $\in$ T)

$$Entropy(c|T) = \sum_{v \in V(c)} p_v entropy_N(\{c_v|T\}) \ ,$$

with $p_v = \frac{|c_v|}{|c|}$ and $\{c_v|T\} = \{|c_v \text{ with y} = t_1|, |c_v \text{ with y} = t_2|, ..\}, t_i \in T$

Feature selection using information gain

- *Entropy is a measure of impurity in a collection of training examples*
- The effectniss of a feature variable x_i in classifying the training data is given by the *Information Gain*:

$$Gain(y, x_i) = entropy_N(\{y|T\}) - \sum_{v \in V(x_i)} p_v entropy_N(\{c(x_i)_v|T\})$$

with $p_v = \frac{|c_v|}{|c|}$ and $\{c_v|T\} = \{|c_v \text{ with y} = t_1|, |c_v \text{ with y} = t_2|, ..\}, t_i \in T$

and $\{y|T\} = \{|y \text{ with y} = t_1|, |y \text{ with y} = t_2|, ..\}, t_i \in T$

Example: Playing Golf?

	$\mathbf{x}_1$ **Outlook**	$\mathbf{x}_2$ **Temperat.**	$\mathbf{x}_3$ **Humidity**	$\mathbf{x}_4$ **Wind**	$\mathbf{y}$ **Playing?**
	Sunny	Hot	High	Weak	No
	Sunny	Hot	High	Strong	No
	Overcast	Hot	High	Weak	Yes
	Rain	Mild	High	Weak	Yes
	Rain	Cool	Normal	Weak	Yes
	Rain	Cool	Normal	Strong	No
	Overcast	Cool	Normal	Strong	Yes
	Sunny	Mild	High	Weak	No
	Sunny	Cool	Normal	Weak	Yes
	Rain	Mild	Normal	Weak	Yes
	Sunny	Mild	Normal	Strong	Yes
	Overcast	Mild	High	Strong	Yes
	Overcast	Hot	Normal	Weak	Yes
	Rain	Mild	High	Strong	No
entropy	1.58	1.56	1.0	0.99	0.94
Entropy	0.69	0.91	0.79	0.89	-
Gain	0.25	0.03	0.15	0.05	-

4.5. Training and Testing

- *Sample training data* is now used to **derive a model** as a hypothesis of the unknown input-output function.
- There are multiple **hypothesis functions H**=$\{h_1,h_2, ..\}$ with $h(x)$: $x \rightarrow y$ that can approximate the unknown function $f(x)$!

The challenge: Find the best hypothesis model function h_i by testing the derived model against the sample and specific *test data* (e.g., by maximizing R^2 measure and using *cross validation*)!!

Constructing a Model

- ➤ Once a researcher has constructed a dataset rich in relevant features or variables, modeling can begin.
- ➤ In DM different kind of models and learning approaches are used, but a specific choice is only temporary.
- ➤ Thus, the data will be analyzed using **several different kinds of models or approaches** and compare their prediction accuracy before settling on a final approach.

4.6. Cross Validation

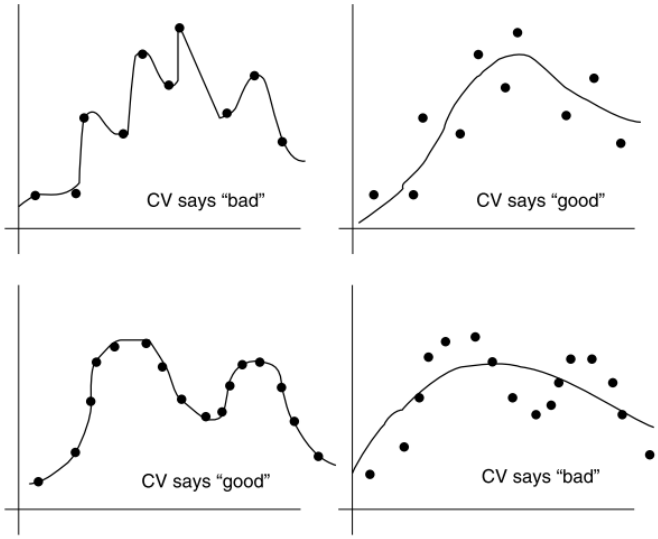


Fig. 4. Some examples of Cross validation of different model hypothesis and their fitting quality

Different sample data classes for generalization

1. One group or random subset of cases or observations is known as the **training sample** or estimation sample. This is the group of cases that will be analyzed first, to create a predictive model.
2. A second random sample can be created, known as the **tuning sample** (it is sometimes called the validation sample). It is used to estimate certain modeling parameters that will yield an optimal prediction.

3. A third randomly selected group of observations is central to cross-validation. This is the **test sample**, sometimes called the holdout sample. The test sample is not used in any way during the creation of the predictive model; it is deliberately kept completely separate (held back).

4.7. Calibrating

➤ **Calibrating** is another DM strategy for improving model prediction that differ from conventional practices, i.e., increasing the R^2 statistical measure of a model for a given sample dataset.

Comparison of conventional and DM approaches

Model type	Test sample R^2
OLS	.288
Partition tree	.442
Bootstrap forest	.438
Boosted tree	.436
Neural network	.481

[2]

Impact of Calibration on R^2 statistical measure

	R^2	RMSE
Basic OLS regression model	.5237	2.28
Above + quadratic term: $\hat{Y}^2$	.5929	2.11
Above + cubic term: $\hat{Y}^3$	.5939	2.11
Above + quartic term: $\hat{Y}^4$	.5949	2.11

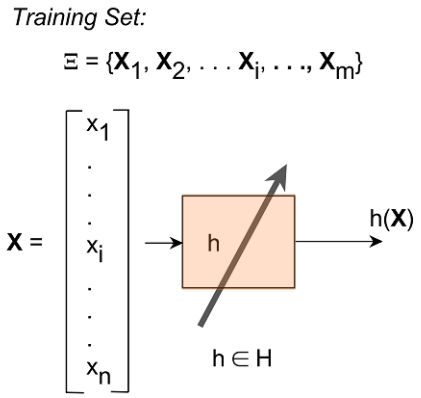
[2]

➤ DM methods outperforms conventional regression with better predictive power!

➤ Calibration improves predictive accuracy!

4.8. Application

Now the learned and mined model function can be used for the prediction of ***unknown*** *input data!*



5. Machine Learning Algorithms and Models

5.1. Machine Learning Classes

Supervised Learning

The *training data* consists of *input data* (**x**, sensor variables, structure parameters, ..) with the associated *output data* (**y**, so-called labels, eg material parameters). The output data is commonly assigned by *experts* (humans), but can also be fed back through an *automatic* evaluation (reinforcement learning).

Unsupervised Learning

In so-called clustering, patterns in the input data are automatically recognized, i.e., the training data consists only of the *input data* **x**.

Reinforcement Learning

This learning class is closely related to *autonomous agents* interacting in an environment with the behaviour: *action* → *perception* → *decicion*

It is a sequential decision-making problem with delayed reward. *Reinforcement learning* algorithms seek to learn a policy (mapping from states to actions) that maximize the reward received over time.
The unlabeled training data $\mathbf{x}(t)$ is provided sequentially as a stream!

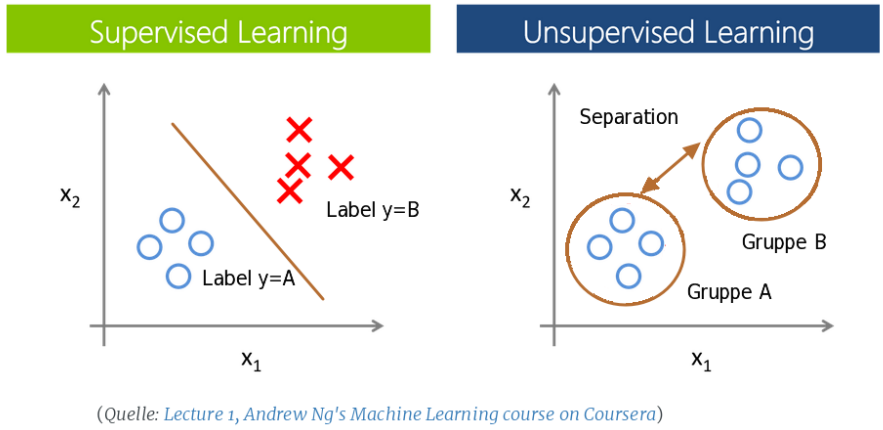


Fig. 5. Comaprison of supervised and unsupervised learning

5.2. Example 1

Do I will play tennis today?

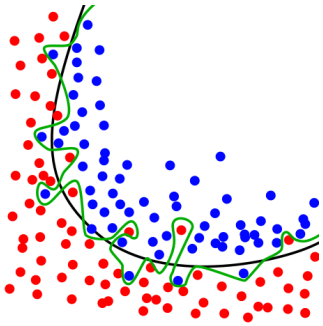
- ➤ Supervised learning with **Decision Tree Model**
- ➤ Trainer works determinstc: *data* → *model*

```
var features = ['Outlook','Temperature','Humidity','Wind']
var target = 'Play Tennis';
var data = [
 ['Sunny',    'Hot',  'High',   'Weak',   'No'],
 ['Sunny',    'Hot',  'High',   'Strong', 'No'],
 ['Overcast', 'Hot',  'High',   'Weak',   'Yes'],
 ['Rain',     'Mild', 'High',   'Weak',   'Yes'],
 ['Rain',     'Cool', 'Normal', 'Weak',   'Yes'],
 ['Rain',     'Cool', 'Normal', 'Strong', 'No'],
 ['Overcast', 'Cool', 'Normal', 'Strong', 'Yes'],
 ['Sunny',    'Mild', 'High',   'Weak',   'No'],
 ['Sunny',    'Cool', 'Normal', 'Weak',   'Yes'],
 ['Rain',     'Mild', 'Normal', 'Weak',   'Yes'],
 ['Sunny',    'Mild', 'Normal', 'Strong', 'Yes'],
 ['Overcast', 'Mild', 'High',   'Strong', 'Yes'],
 ['Overcast', 'Hot',  'Normal', 'Weak',   'Yes'],
 ['Rain',     'Mild', 'High',   'Strong', 'No'],
]
model1 = ML.learn({algorithm:ML.ML.C45,
 data:data, features:features});
ML.classify(model1,[
  ['Overcast', 'Cool', 'Normal', 'Strong'],
  ['Sunny',    'Hot',  'High',   'Strong'],
  ['Overcast', 'Hot',  'Normal', 'Strong'],
  ['Rain', 'Hot', 'Normal', 'Strong']]);
>>
[ 'Yes', 'No', 'unknown', 'unknown' ]
```

5.3. Machine Learning - Noise

- ➤ Noise and measurement uncertainty can affect the derived model until it is useless !!!
- ➤ This concerns both phases:
  - ❑ Learning process (training)
  - ❑ Classification process (application of the learned model)
- ➤ Also erroneous assignment of labels (classes) is noise!

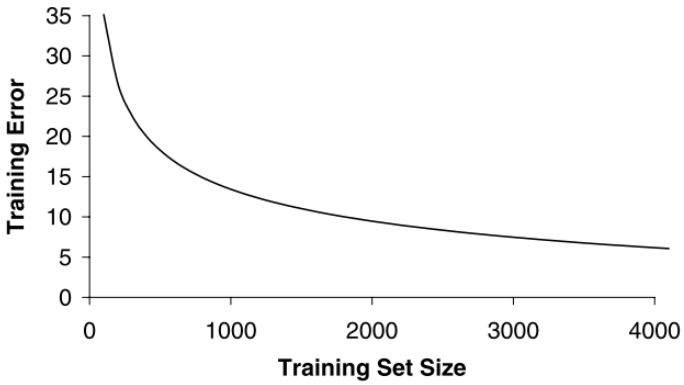
Impact of noise on prediction accuracy depends on the algorithm and models that are used



5.4. Machine Learning - Overfitting

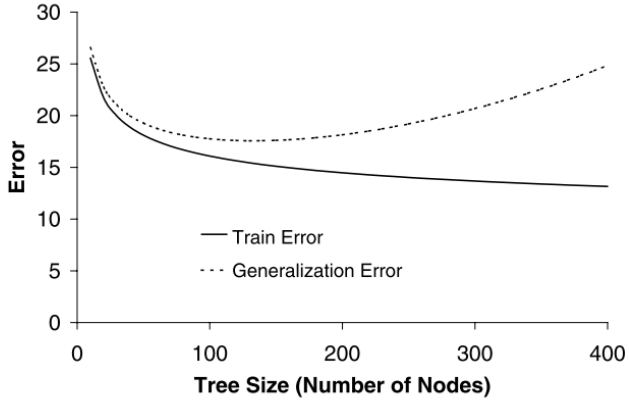
Overfitting

- The generalization error decreases with the training set size
- But the training and generalization errors decreases more slowly for larger and grwoing training data sets

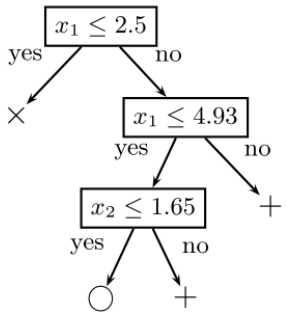


Overfitting in Decision Trees

- Overfitting *decreases* prediction accuracy in decision trees!
- The training error continues to decline as the tree become bigger.
- The generalization error declines at first then at some point starts to increase due to overfitting



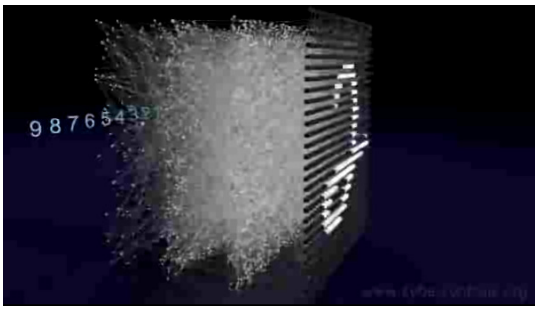
5.5. Model: Decision Tree

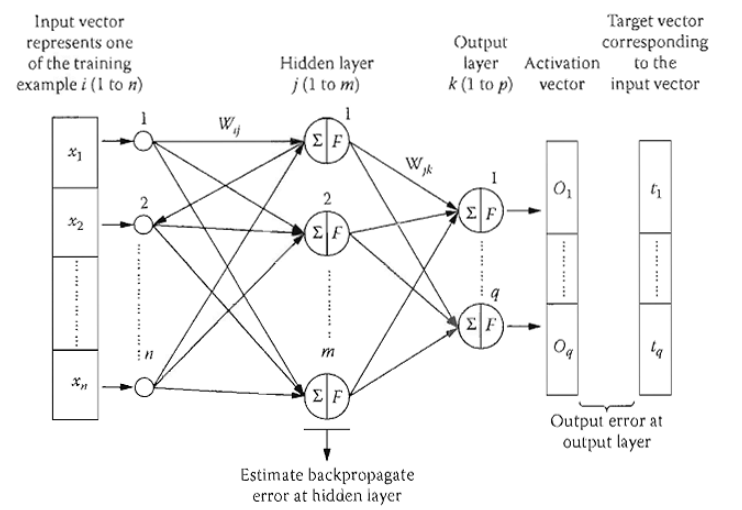


if $x_1 \leq 2.5$ then $\times$
if $x_1 \in (2.5, 4.93]$ and $x_2 \leq 1.65$ then $\bigcirc$
else $+$

- A decision tree consists of nodes that are linked to and evaluate a given input variable x_i [3].
- There are one or more edges to successor nodes depending on the variable value (relational or interval).
- Leaves are linked to the classification result. Both graph, table, and program notation are available as representation of the learned model? **Compact model**!

5.6. Model: Artificial Neural Network





- An artificial neural network (ANN) is a network of neurons (perceptrons) [1].
- Each neuron n_i has one or more inputs $\boldsymbol{x}_i$ and an output y_i.
- There is an activation function: $f_i(\boldsymbol{x}_i)$: $\boldsymbol{x}_i \rightarrow y_i$, $y \in [0,1]$, $x_{i,j} \in [a,b]$.
- The training of the network is iterative (probalistic) and is done by 1. distribution of weights of the edges; 2. Configuration until the entire classification with minimal error is possible.

5.7. Example 2

Do I will play tennis today?

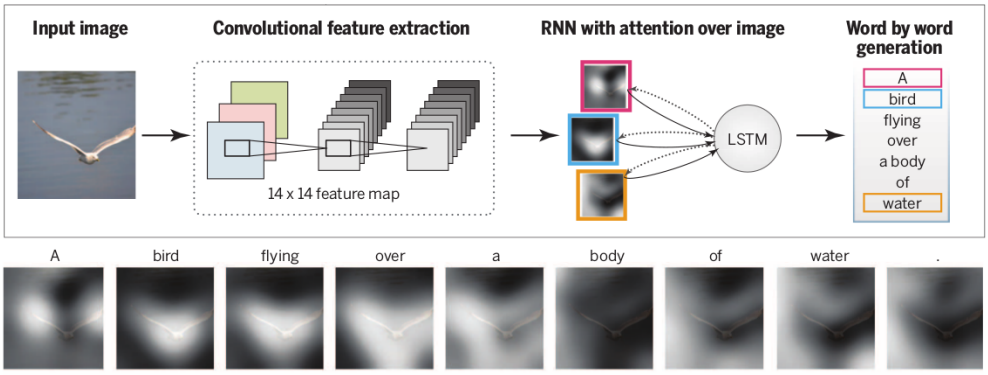
- Supervised learning with an **Artificial Neural Network**
- Trainer works randomized: $data \rightarrow model_1|model_2|,..$

```
x = [
  [ 1, 1, 1, 0 ], [ 1, 1, 1, 1 ],
  [ 0, 1, 1, 0 ], [ -1, 0, 1, 0 ],
  [ -1, -1, 0, 0 ], [ -1, -1, 0, 1 ],
  [ 0, -1, 0, 1 ], [ 1, 0, 1, 0 ],
  [ 1, -1, 0, 0 ], [ -1, 0, 0, 0 ],
  [ 1, 0, 0, 1 ], [ 0, 0, 1, 1 ],
  [ 0, 1, 0, 0 ], [ -1, 0, 1, 1 ] ];
y = [ [ 0, 1 ], [ 0, 1 ], [ 1, 0 ], [ 1, 0 ], [ 1, 0 ],
      [ 0, 1 ], [ 1, 0 ], [ 0, 1 ], [ 1, 0 ], [ 1, 0 ],
      [ 1, 0 ], [ 1, 0 ], [ 1, 0 ], [ 0, 1 ] ];
model2 = ML.learn({
  algorithm:ML.ML.MLP,
  x:x, y:y,
  hidden_layers : [4,4,5],  // network structure
  lr : 0.6,  // magic parameters
  epochs : 20000,  // itertations
});
ML.classify(model2,[
  [0,-1,0,1],
  [1,1,1,1],
  [0,1,0,1],
  [-1,1,0,1]);
>>
[ [ 0.9912471319019274, 0.008733940677322432 ],
  [ 0.006654940681834191, 0.9933810393277491 ],
  [ 0.9953821975343892, 0.004588392283525628 ],
  [ 0.767924178373238, 0.23242211270157057 ] ]
```

5.8. Deep Networks

Example: Automatic generation of text captions for images with a deep neuronal network [6]

- A convolutional neural network is trained to interpret images, and its output is then used by a recurrent neural network trained to generate a text caption (top).
- The sequence at the bottom shows the word-by-word focus of the network on different parts of input image while it generates the caption word-by-word.



6. Conclusions

Using Data Mining

- creates **predictive models**, or
- **classifies** things, or
- **identifies** different groups or clusters of cases within data.

The challenge: Find the best hypothesis model function h_i by testing the derived model against the sample and specific *test data* (e.g., by maximizing R^2 measure and using *cross validation*)!!

- **Data**: Big is not big enough $\Rightarrow$ But size is not everything!
- **Labeling**: Accurate assigning of labels to training data is crucial!
- **Feature Selection**: Selecting the right (the best/relevant) feature attributes is crucial!
- **Training and Testing**: The data will be analyzed and tested using *several different kinds of models or approaches* and compare their prediction accuracy before settling on a final approach.

7. References

Literature

1. L. Rokach and O. Maimon, DATA MINING WITH DECISION TREES Theory and Applications. World Scientific Publishing, 2015.

2. P. Attewell and D. B. Monaghan, Data mining for the social sciences : an introduction. University of California Press, 2015.
3. N. J. Nilsson, Introduction To Machine Learning. 1996.
4. W. Mason, J. Wortman Vaughan, and H. Wallach, "Computational social science and social computing," Machine Learning, vol. 95, 2014.
5. V. C. Raykar et al., "Learning From Crowds," Journal ofMachine Learning Research, vol. 11, 2010.
6. M. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, 2015.

Videos

A. Neural Network 3D Simulation, www.youtube.com/watch?v=3JQ3hYko51Y