# The Four Steps of Modeling

##### Select the model class

* + General structure of the analysis result
  + "Architecture" or "model class"
  + Example: Linear or quadratic functions for regression problem

##### Select the score function

* + Evaluate possible "models" using a score function
* Apply the **algorithm**
  + Compare models through the score function
  + But: How do we find the models?

##### Validate the results

* + We know: Best model among the chose ones
  + But: Is this the best among very good or very

bad choices?

* Model = The form or structure of the analysis result
* Here the parameters are not defined. Only the type is selected
* Examples
  + Linear models, e.g., *y* = *ax*  *b*
  + Constant values (e.g. mean)

− Rule based models (if A buys product one, then weather is sunny)

##### Global models provide a (not necessarily

good) description for the whole data set

* + Example: Regression line

##### Local models or patterns provide a description for only a part or subset of the data set.

* + Example: Association rules

##### find an objective function 𝑓: 𝑀 → 𝑅

* Which, evaluates the quality of your model
* In order to detect the "best" model

• Given dataset 𝐷 = {𝑑1, 𝑑2, … , 𝑑�} ∈ 𝑅 and

Model 𝑀: 𝑅� → 𝑅�

*n*

##### Mean squared error:

*f* (*x*) = 1 

*x*  M(*x*)2

* Mean absolute error:

*f* (*x*) =

*n i*=1

*n* *i*=1

*n*

1

*x*  M(*x*)

##### Short comment : What is classification?

Imagine a cup factory,

which wants to classify their cups as good or broken.



**Error functions for classification problems**

How to set up an error function for those

classification problems?

* Very common misclassification rate =

# wrongclassified

# totalclassified

* A low misclassification rate does not

necessarily tell anything about the quality of a classifier.

* when classes are unbalanced (e.g. When 99% of the production are ok, a classifier always predicting ok will have a misclassification rate of 1%.)

### The objective function (scoring function)

for models

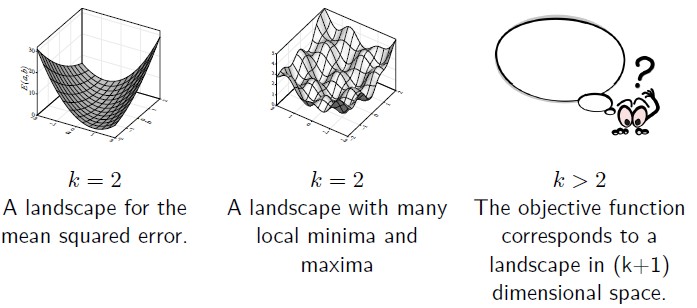
##### does not tell us how to find the best or a good model,

* it only provides a means for comparing

models.

Optimisation algorithms to find the best or at

least a good model are needed.

For differentiable score functions, a gradient methods can be applied. (*f*: 𝑅𝑘 → 𝑅 , k – real-valued parameters)

* Will only find local optima.
* Parameters (step width) must be adjusted or

computed in each iteration step.

* For discrete problems with a finite search space (like finding association rules), combinatorial optimization strategies are needed.
* In principle, an exhaustive search of the finite domain 𝑀 is possible, however, in most cases it is not feasible, since 𝑀 is much too large..
* Finding the best possible association rules with an underlying set of1000 items (products).
* Every combination of items, i.e. every nonempty subset is a possible candidate set from which several rules may be constructed.
* The number of nonempty subsets alone contains

21000 − 1 > 10300 elements.

**Heuristic strategies are therefore needed**

* There are 4 parts of potential error origins, which sum up to the overall error measure
* Error = Experimental error

+ Sample error

+ Model error

+ Algorithmic error

* The **pure error** or **experimental error** is inherent in the data and due to noise, random variations, imprecise measurements or the influence of hidden variables that cannot be observed.
* It is impossible to overcome this error by the

choice of a suitable model.

* Also called **intrinsic error**.
* In the context of classification problems it is also

called **Bayes error**.

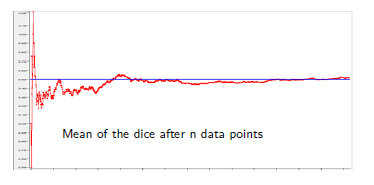
* A confusion matrix is a table where the rows represent the true classes and the columns the predicted classes. Each entry species how many objects from a given class are classified into the class of the corresponding column.

|  |  |  |  |
| --- | --- | --- | --- |
| **True classes** | **Predicted classes** | | |
|  | Iris setosa | Iris versicolor | Iris verginica |
| Iris setosa | 50 |  |  |
| Iris versicolor |  | 47 | 3 |
| Iris verginica |  | 2 | 48 |

A possible confusion matrix for the Iris data set

* The data is not a perfect representation of the underlying data
* The smaller the sample the smaller the probability

for a perfect model.



s Edited from "Guide to Intelligent Data Analysis", By M. Berthold et al

* Throw a dice
* Sample Bias: what if the dice was not fair?

There are different models for the data

* simpler model → bigger error
* more complex model → overfitting and larger error

on new data

* type of model → different “fit" to data
* Based on the selected algorithm
* For example
  + Gradient descend → local minima
  + Randomized method → too much randomness
* The algorithmic error can often not be measured

(several runs of similarly biased)

* Normality: we assume that our algorithm is good enough (otherwise : choose another)

. Berthold et al

* The error for unseen data will most probably always be bigger than for the data used for training.
* How do we find out which model is actually suited best to our problem?
* Split data into two subsets: training and test data
* Train your model on the training data and measure the model quality on the test data
* Typically 2/3 training 1/3 test (usually more training)
* Splitting strategies
  + Random (distribution in both sets should be roughly same)
  + Stratification (i.e. the distribution of one class should remain)
* Split into training, test and validation
  + Choose for each model kind the best based on the test

Data

Ts Eedistetd ftrohme"Gubidee tso ItntemlligeontdDeatalsAnaolynsis",tBhyeM. Bvearthloildd eat atlion data set.

* Split the data multiple times to validate the results, to reduce the effect of the single estimation
* Use a combination of the received models, or the best.
* K-fold Cross-Validation (e.g. k=10):
  + Divide into k subsets
  + Test data is always another subset, training data the rest
  + Average of the k model error is supposed as the model error
* Leave-One-Out Method:
  + For very small data sets
  + Use everything except of one data point for training
  + This single data point is used for testing.

#### Summary

* Select the Model Class
  + Which kind of model you are looking for?
* Select the Score Function
  + Which function can be used to evaluate the possible

“models”?

* Apply the Algorithm
  + How to fine the model?
* Validate the Results
  + Is the model found the best?

## References

* M. Berthold, C. Borgelt, F. Hoppner and F. Klawonn

Guide to Intelligent Data Analysis, Spriner-Verlag, 2010

– Chapter 5, Section 5.1, 5.2, 5.3,5.4, 5.5.1, 5.5.2

* Online materials