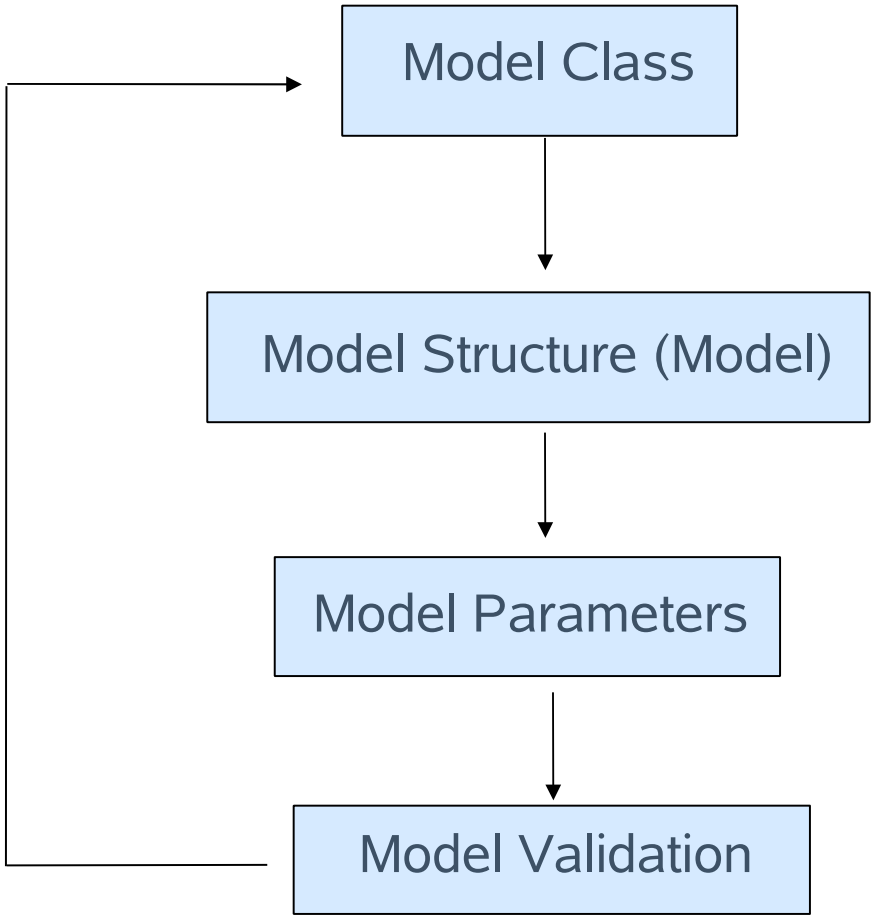
Analysis of Microarray Data with Methods from Machine Learning and Network Theory

Summer Lecture 2015

Prof. Dr. A. B. Cremers
Dr. Jörg Zimmermann

Machine Learning: The model building process



For example:

all polynomials

polynomials of specific degree, e.g.:

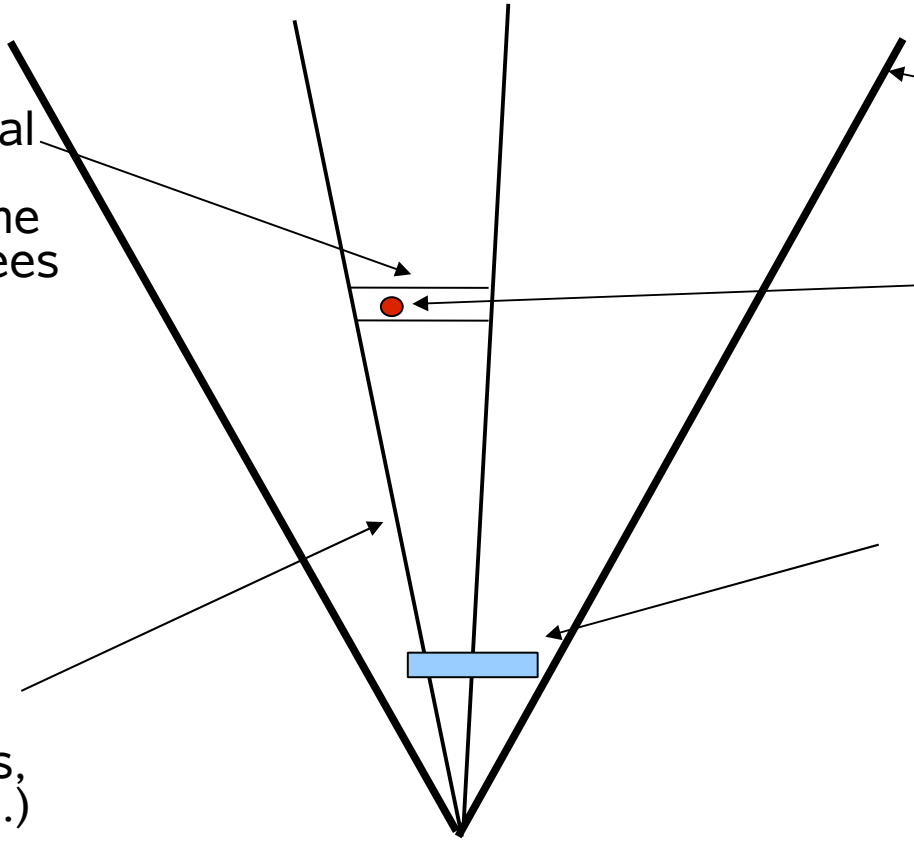
$a*x^2+b*x+c$

specific polynomial, e.g.:

$3*x^2+2*x-1$

Test set method, Cross-validation

The Universe of Models



segment of models with similar structural complexity (polynomials of same degree, decision trees with same number of nodes, ...)

all (computable) Models

concrete model with specified structure and parameters

flat model class (e.g. straight lines)

deep model class (e.g. all polynomials, all decision trees, ...)

(A model class is deep if for every finite data set there is a model in the class which describes the data set perfectly.)

Machine Learning

The most important and the most difficult task in machine learning is to select (or design) the “right” model class for your inference problem. That is the primary task of a data analyst and model builder, using his/her knowledge of the application area.

An important goal is to understand the possibilities and limitations of currently known deep model classes and to develop new model classes with efficient learnable and efficient evaluable models.

Important topics not covered

Confidence Intervals (Classical Statistics), Credal Intervals (Bayesian Statistics): estimation of intervals (not points), which likely contain the true parameter.

Sensitivity Analysis of Estimators: are they robust against small perturbations and outliers in the data set?

Experimental Design (statistical jargon), Active Learning (Machine Learning jargon): How should one devise a study? How should one collect data in an "optimal" fashion?

Curse of dimensionality: many methods, which work well for low dimensional problems (≤5 dim.), fail badly for high dimensional problems (≥15 dim).

Take Home Messages

Know your data.

Don't trust results of statistical analysis (too many hidden assumptions, software bugs, ...). Always crosscheck with your background knowledge and common sense.

Statistical results are starting points for thought, not end points.

Revisiting the most essential topics and notions of statistics and machine learning

Correlation is not Causation

It is important to note that correlation between two variables implies by no means causation!

The analysis of causation needs to address the elimination of confounders (hidden causes), e.g. by appropriate experimental design (double-blind studies, randomization of test groups, ...).

Maximum Likelihood Estimator

If L_x assumes a maximal value for θ*, i.e. if:

$$L_x(\theta^*) = sup\{L_x(\theta) : \theta \in \Theta\}$$

then we call θ* a maximum likelihood estimator of parameter θ.

In many practical situations there is a unique maximum likelihood estimator, and it is usually a good estimator.

Bayesian Statistics

Bayesian statistics is an alternative approach to estimation and decision problems.

The main difference to classical statistics is that model parameters like θ are now treated as random variables, too.

A probability distribution on model parameters is interpreted as a knowledge state of an observer about the true model parameters, thus quantifying the uncertainty about the true model parameters.

In Bayesian statistics it is possible to make probabilistic statements about the unknown (and unobservable) model parameters (in contrast to classical statistics).

The knowledge state of an observer can change if new data is observed. Probability theory implies that there is exactly one way to do this: Bayes' rule.

Bayes' Rule (or Bayes Theorem)

posterior probability of θ

prior probability of θ

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)} \cdot p(\theta)$$

update factor

Bayes Rule states how to learn from data. It tells you the revised probability of a model θ after seeing data x.

Remember: p(x|θ) is the likelihood function, i.e. the family of stochastic models one has to choose to define the inference problem (e.g. a Bernoulli likelihood or a Gaussian likelihood)..

Bayes' Rule: prior probability

The prior probabilities for θ (the prior distribution, or just prior) are interpreted as the knowledge state of an observer before seeing data x. How to choose a prior for a specific inference problem is in general a difficult problem and is treated extensively in the literature (e.g. R. Yang, J. Berger: A Catalog of Noninformative Priors, where you can look up priors for many standard inference problems).

A main goal of prior theory is the definition of “noninformative priors”, which represent maximal ignorance of model parameters with regard to a well-defined criterion (see A. R. Syversveen: Noninformative Bayesian Priors). Bayesian inference based on such noninformative priors can be regarded as “objective” in a well-defined manner, containing no hidden assumptions or subjective knowledge.

Bayes' Rule: update factor

The update factor is the ratio of the likelihood of the observed data x given a specific model parameter θ and the probability of data x (before you have seen the data).

If the likelihood of x given θ is greater than the (unconditioned) probability of data x, than this can be regarded as positive evidence for model parameter θ, resulting in an update factor greater than 1. Correspondingly, if the likelihood of x given θ is less than the probability of data x, than is negative evidence and the update factor will be less than 1.

The unconditional probability of data x can be reduced to known quantities. It holds:

conditionalization (3. axiom)

$$p(x) = \int_{\theta \in \Theta} p(x, \theta) d\theta = \int_{\theta \in \Theta} p(x|\theta) \cdot p(\theta) d\theta$$

marginalization (see next slide)

Bayes' Rule: Marginalization

The process of reducing a joint probability distribution p(x,θ) to only one variable by ignoring (or averaging) the other variable leads to the “marginal distribution”:

$$p(x) = \int_{\theta \in \Theta} p(x, \theta) d\theta$$

Continuous case

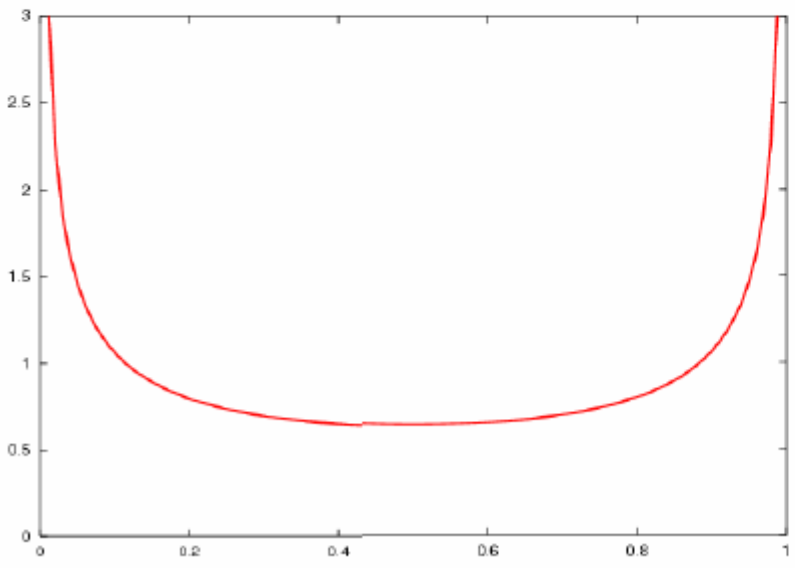
$$p(x) = \sum_{i \in I} p(x, \theta_i)$$

Discrete case

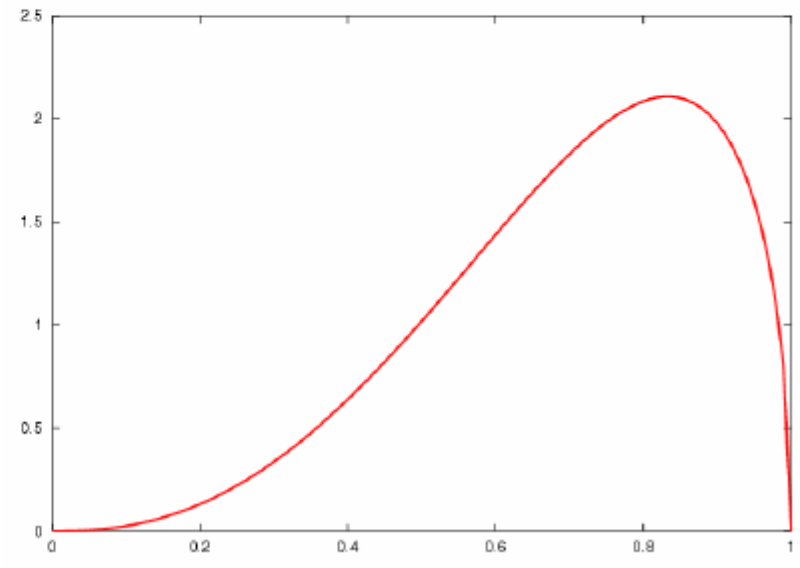
Marginalization can be justified via the 2. axiom (sum of probabilities).

Binomial Model

Prior



Posterior (n = 4, x = 3)



Bayesian learning as the transformation of knowledge states (as probability distributions on hypotheses space) driven by data.

p-values

If two experimenters use different values as significance niveaus, say 0.05 and 0.01, then the same experimental outcome could lead to rejection of the null-hypothesis in one case and acceptance (better: non-rejection) in the other case.

If a concrete result of an experiment is available, one could ask what the critical significance level is:

The critical significance level is the least value which allows a rejection of the null hypothesis.

This critical level is also called the observed significance niveau or p-value.

p-values

In the case of our tea tasting lady, the p-value for n=20 and

x=14	is	p-value = 0.0577
x=15	is	p-value = 0.0207
x=16	is	p-value = 0.0059
x=17	is	p-value = 0.0013
x=18	is	p-value = 0.0002

We see that in fact the observed significance niveau in the case x=15 is much better than 0.05.

p-values are widely used in reporting statistical results. Generally, a p-value less than 0.05 is regarded as a significant result.

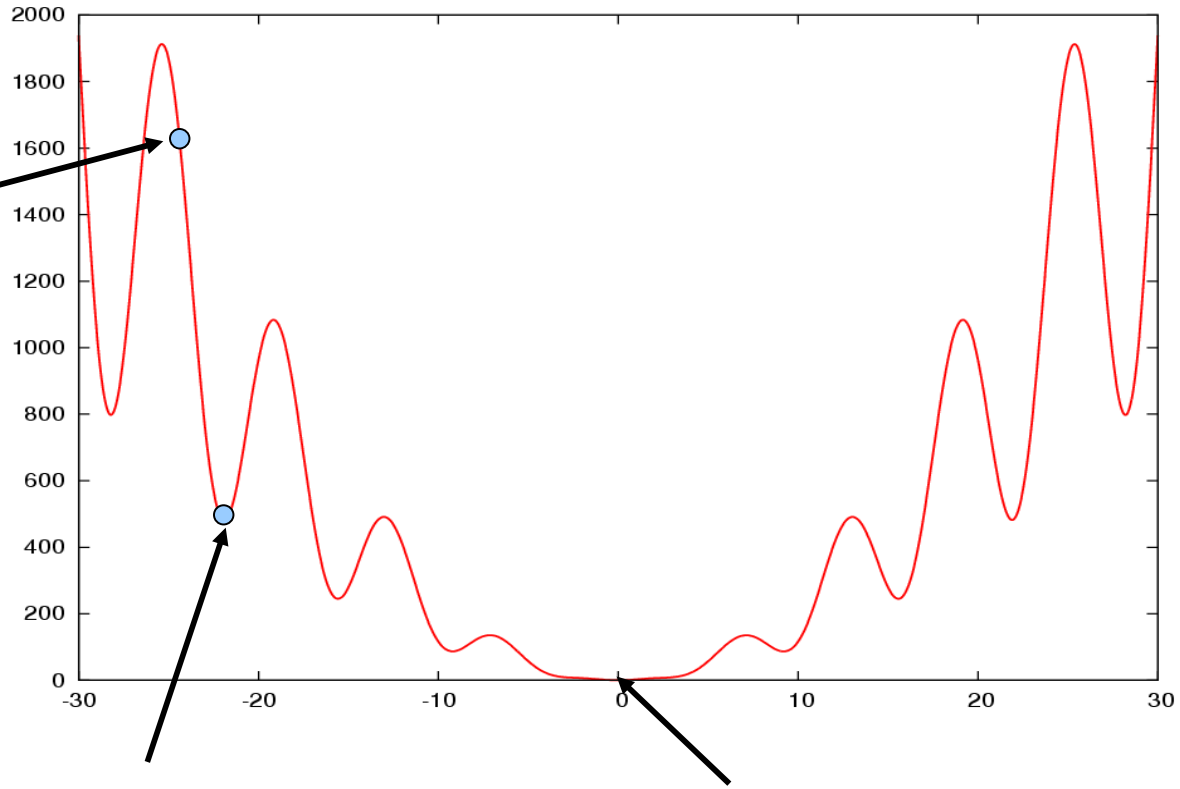
K-Means Clustering

In some cases, the randomly chosen starting points for cluster centers can lead to the following converged configuration, which has not minimal total cluster variance:

K-Means Clustering

General fact: local optimization algorithms can get stuck in local optima

If you start here



... you will end there

global minimum