# Machine Learning II

Nils Bertschinger
Matthias Kaschube

SoSe 2018

# Machine Learning II

Nils Bertschinger
Helmut O. Maucher Stiftungsjuniorprofessor für Systemische Risiken
Frankfurt Institute for Advanced Studies (Campus Riedberg)
Room 2.402
bertschinger@fias.uni-frankfurt.de

# What is "Machine Learning"?

- Learning according to Wikipedia:

  > *Learning is the act of acquiring new, or modifying and reinforcing, existing knowledge, behaviors, skills, values, or preferences and may involve synthesizing different types of information*

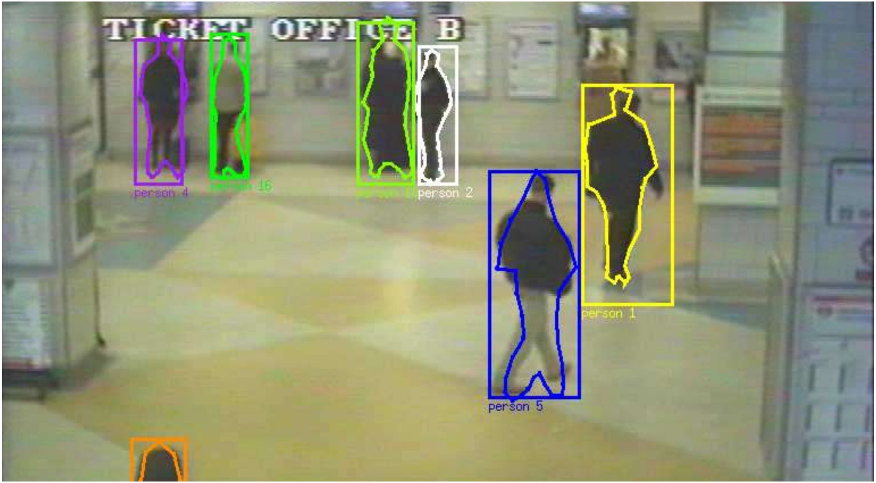
- In 1959 Arthur Samuel wrote the first self-learning *program* that could play Checkers. Accordingly he defined machine learning as

  > *Field of study that gives computers the ability to learn without being explicitly programmed.*

- More formal definition in operational terms by Tom Mitchell:

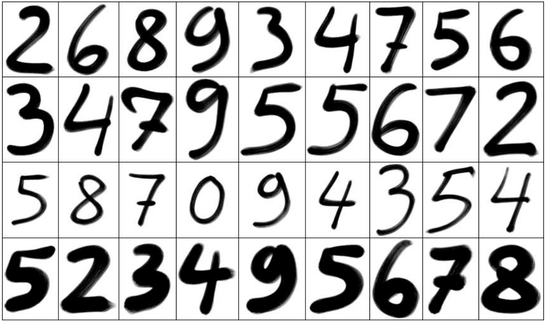
  > *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

# Surveillance



# Identity authentication

# Handwritten digit recognition

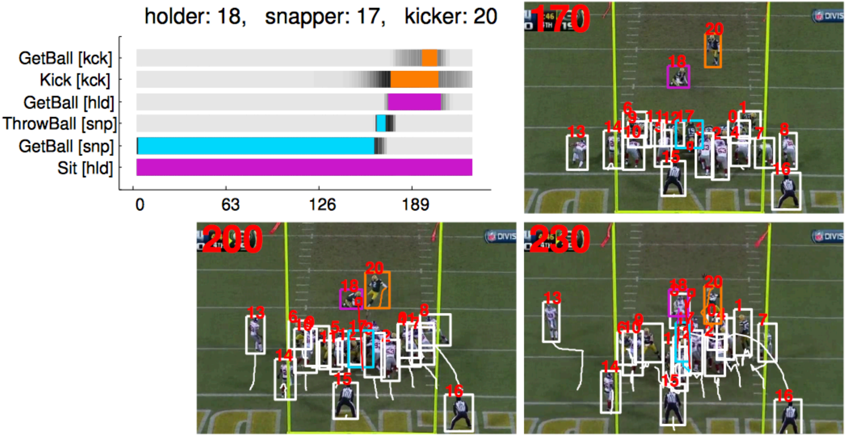



- MNIST database (Mixed National Institute of Standards and Technology database)

# Speech recognition

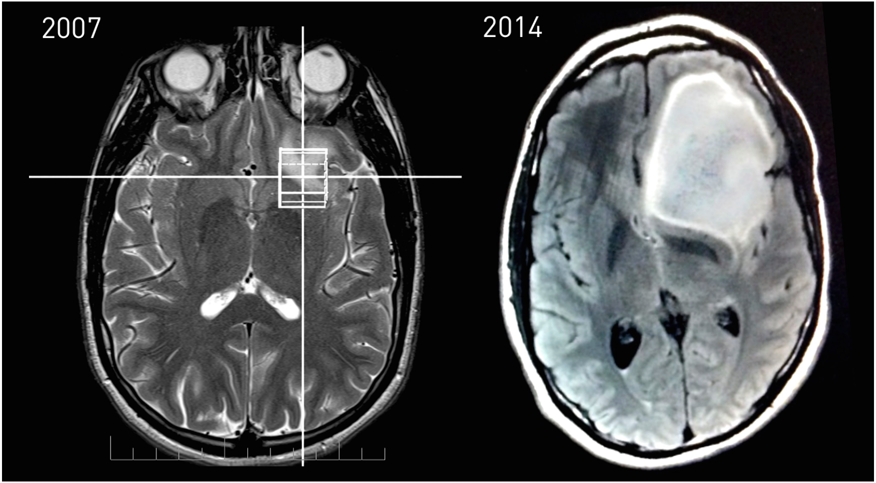
# Object Recognition in Sports

Holder, snapper, and kicker are denoted by purple, cyan, and orange, respectively while outsiders are denoted by white boxes.




(Kwak et al., CVPR 2013)

# Magnetic Resonance Imaging (MRI)




(Keating, 2015)

# Two approaches to Machine Learning

- Data-driven
  - Very large data sets ... "Big Data"
  - Non-parametric models, e.g. k-NN
- Model-driven
  - Can be used for small data sets
  - Parametric models

Note: As models become more complex any data set is "small"

$\Longrightarrow$ Recent rise of model based machine learning

# Model based machine learning

General setup of model based ML:

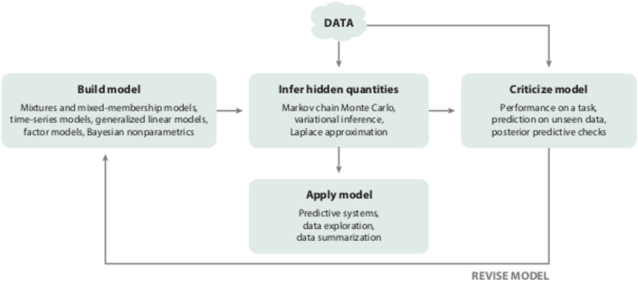



Fig. from: David M. Blei, *Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models*, Annu. Rev. Stat. Appl. 2014. 1:20332

# Basic terminology

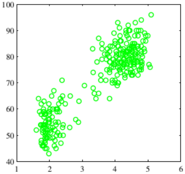
- **Supervised:** Patterns whose class/output is known a-priori are used for training (*labelled training data*)
  - *Regression:* Real-valued output
    Typical examples: Interpolation, (Time-series) Prediction
  - *Classification:* Categorical output
    Typical examples: Face recognition, Identity authentification, Speech recognition
- **Unsupervised:** Number of classes is (in general) unknown and no labelled data are available
  Typical examples: Cluster analysis, Recommendation systems

# Machine Learning as statistics

Classical example: **Clustering**

The Old Faithful geyser in Yellowstone National Park. ©Bruce T. Gourley www.brucegourley.com.

Plot of the time to the next eruption in minutes (vertical axis) versus the duration of the eruption in minutes (horizontal axis) for the Old Faithful data set.




- Which data points belong together?
- How many groups/cluster are there?

Fig. from: Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006.

# Clustering

Classical algorithm: **K-means**

- Input: Data $x_1, \ldots, x_N \in \mathbb{R}^D$; Number of clusters $K$
  Output: Cluster centers $\mu_1, \ldots, \mu_K$

**Step 1** Choose "arbitrary" initial cluster centers $\mu_1, \ldots, \mu_K$

**Step 2** Assign each data point $x_i$ to its nearest cluster:

$$c_i = \mathrm{argmin}_c ||x_i - \mu_c||^2$$

**Step 3** Update the cluster centers:

$$\mu_c = \frac{1}{\#\{i|c_i = c\}} \sum_{\substack{i=1 \\ c_i=c}}^{N} x_i$$

Repeat from step 2 until assignment is stable.

# K-means

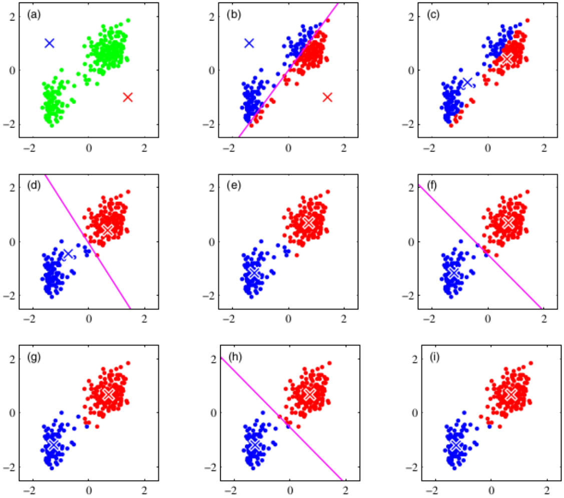



Fig. from: Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006.

# K-means

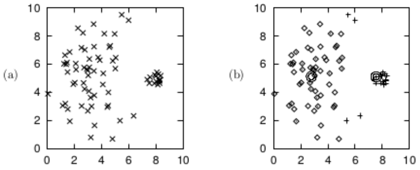



Figure 20.5. K-means algorithm for a case with two dissimilar clusters. (a) The "little 'n' large" data. (b) A stable set of assignments and means. Note that four points belonging to the broad cluster have been incorrectly assigned to the narrower cluster. (Points assigned to the right-hand cluster are shown by plus signs.)

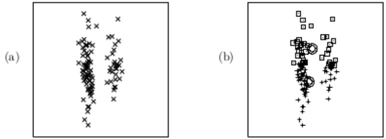



Figure 20.6. Two elongated clusters, and the stable solution found by the K-means algorithm.

Problems of K-means:

- Hard cluster assignment
- (Implicit) assumptions about cluster shape

Fig. from: David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

# Clustering ... rethinking

Clustering as a statistical problem:

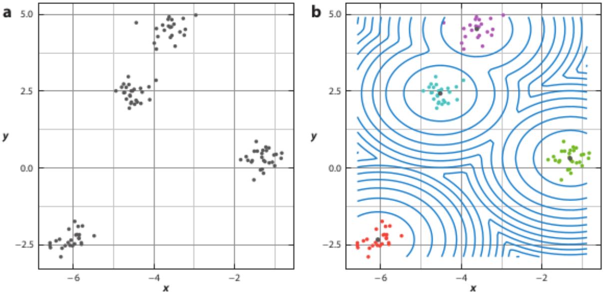
- Assume that data are drawn from probability distribution $p(\mathbf{x})$
- Data point $\mathbf{x}_n$ could have been generated as follows:
  1. Draw (hidden) class assignment $c_n \in \{1, \ldots, K\}$
  2. Draw data point from class-conditional distribution $p(\mathbf{x}_i | c_i)$
- *Mixture model:* Natural generative model for clustered data

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(c_k) p(\mathbf{x} | c_k)$$

Unobserved class assignment $c$ is *latent variable*

# Clustering

Gaussian mixture model:




- Takes uncertainty into account $\implies$ soft clustering
- Possible to predict new data points
- *Model selection:* Principled way to discover number of clusters

Fig. from: David M. Blei, *Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models*, Annu. Rev. Stat. Appl. 2014. 1:20332

# What is probability?

Two (competing) philosophies:

- **Frequentist:** Probability of an event is *relative frequency of occurrence* when experiment is repeated infinitely many times.
- **Bayesian:** Probability describes (subjective) degree of belief

# What is probability?

Two (competing) philosophies:

- **Frequentist:** Probability of an event is *relative frequency of occurrence* when experiment is repeated infinitely many times.
- **Bayesian:** Probability describes (subjective) degree of belief

Common wisdom

- Frequentist = objective
  Bayesian = subjective
- But:
  - Asymptotics vs finite sample
  - Repeated trials vs decision making

# Motivation

Hypothetical situation:

- Assume that a patient enters your office
- Her test result for a rare disease is *positive*

**Q:** Would you suggest an *expensive* treatment?

What do you need to know for an *informed* decision?

# Classical answer

**Classical** answer:

- Test between two hypothesis:
  - $H_0$: Patient is healthy (null hypothesis)
  - $H_1$: Patient is infected (alternative)
- Specificity of test, i.e.

$$P(Test = negative | H_0)$$

  Here: Specificity of 99.9%

# Classical answer

**Classical** answer:

- Test between two hypothesis:
  - $H_0$: Patient is healthy (null hypothesis)
  - $H_1$: Patient is infected (alternative)
- Specificity of test, i.e.

$$P(Test = negative | H_0)$$

  Here: Specificity of 99.9%

Reject null hypothesis at level $\alpha$ if

$$P(\underbrace{Test = positive}_{\text{or more extreme outcome}} | H_0) = 1 - P(Test = negative | H_0) = \frac{1}{1000} < \alpha$$

## Bayesian answer

*Basic idea:* Uncertainty about the status of the patient (healthy or infected) can be expressed as a probability distribution.
Combine two sources of information:

- **Prior:** How rare is the disease?
  Here: Rare disease can be found in 1 out of 10000 persons, i.e.

$$P(Status = infected) = \frac{1}{10000}$$

- **Likelihood:** How good is the test?
  - Specificity: $P(Test = negative|H_0)$
  - Sensitivity: $P(Test = positive|H_1)$
    Here: Sensitivity of 99.9%

Base your decisions on **posterior**, i.e.

$$P(Status|Test = positive)$$

# Bayes rule

Bayes rule is used to calculate posterior probability:

$$
\begin{aligned}
P(S|T) &= \frac{P(S)P(T|S)}{P(T)} \\
&= \frac{P(S)P(T|S)}{\sum_{S'} P(S')P(T|S')} \\
&\propto P(S)P(T|S) \\
\textit{posterior} &\propto \textit{prior} \times \textit{likelihood}
\end{aligned}
$$

Bayes rule is uncontroversial and follows from the product rule of probability theory:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

# Bayesian answer

Want to know posterior probabilities:

$$
\begin{aligned}
P(\mathit{Status} = \mathit{infected} | \mathit{Test} = \mathit{positive}) &= \frac{P(S = \mathit{inf})P(T = \mathit{pos}|S = \mathit{inf})}{P(S = \mathit{inf})P(T = \mathit{pos}|S = \mathit{inf}) + P(S = \mathit{hea})P(T = \mathit{pos}|S = \mathit{hea})} \\
&= \frac{\frac{1}{10000}\frac{999}{1000}}{\frac{1}{10000}\frac{999}{1000} + \frac{9999}{10000}\frac{1}{1000}} \\
&\approx 0.09
\end{aligned}
$$

Thus, patient is not very likely to be infected and we would need more evidence before suggesting the expensive treatment!

# Bayesian thinking

Bayesian statistics is conceptually simple

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood},$$

but can be computationally demanding
Every type of uncertainty is expressed in terms of probability distributions. This includes

- Statistical models of data, e.g.

$$P(\textit{Test} | \textit{Status} = \textit{infected})$$

- Plausibility of hypothesis, e.g.

$$P(\textit{Status} = \textit{healthy})$$

- Parameters ...

In general, probabilities are assigned to logical statements.
Conclusions are derived by computing their posterior probabailities.

# Decision theory

Bayesian statistics is deeply rooted in decision making
Subjective probabilities can be recovered from betting odds:

- Assume you are willing to accept a bet at 1:19, i.e.
  - You pay 1$ if you loose
  - You get 19$ if you win

  Your subjective probability of winning is then $\frac{1}{20}$ as it leaves you indifferent between accepting the bet or not, i.e.

$$\mathbb{E}[payout] = (1 - \frac{1}{20})(-1\$) + \frac{1}{20}19\$ = 0$$

*Dutch book argument*:

- A dutch book is a set of bets such that you loose money no matter what happens.
- Coherent betting odds have to fullfil the laws of probability

# Dutch book

Dutch book coherence:

- Consider a bet at odds 1:a, i.e. with subjective probability $q = \frac{1}{1+a}$
- Equivalent to a lottery ticket which costs $q$ and pays 1 on winning

Now, the total price of tickets winning on disjoint events $A$ and $B$ respectively, must equal the price of ticket winning on $A \cup B$, i.e.

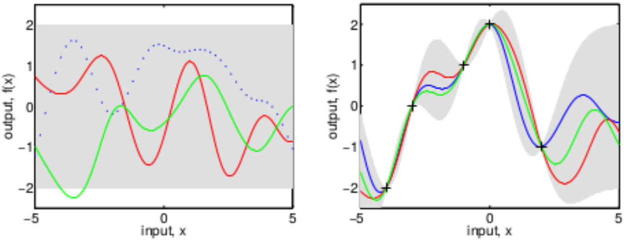
$$q(A) + q(B) = q(A \cup B) \quad \text{for } A \cap B = \emptyset$$

stating that $q$ obeys the sum rule of probability.
Similarly, other laws are proved for coherent bets.

# Bayesian machine learning

*Bayesian statistics*:

- Principled and logically consistent way to reason under uncertainty
  Prior $\xrightarrow{\textit{Data}}$ Posterior (belief update)




(a), prior (b), posterior

- Especially useful when taking decisions or making predictions

Fig. from: C. E. Rasmussen & C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006.

# Bayesian machine learning

*Bayesian statistics*:

- Principled and logically consistent way to reason under uncertainty
  Prior $\xrightarrow{\textit{Data}}$ Posterior (belief update)
- Especially useful when taking decisions or making predictions

*Bayesian machine learning*:

- Statistical modeling:

$$p(\underbrace{\mathbf{x}}_{\text{Data}}, \underbrace{\mathbf{z}}_{\substack{\text{Latent variables}\\ \text{Parameters}}}) = \underbrace{p(\mathbf{z})}_{\textit{prior}} \underbrace{p(\mathbf{x}|\mathbf{z})}_{\textit{likelihood}}$$

- Conceptually simple, but computationally challenging

# Bayesian machine learning

Bayesian machine learning:

- Bayesian modeling requires prior assumptions:
  - Parametric models, e.g. linear regression
  - Bayesian non-parametrics:
    - Flexible models with infinite-dimensional parameter spaces
    - Effective number of parameters grow with amount of data

  But, explicit about prior assumptions
  - *No free lunch theorem*: Assumption-free learning is impossible!
- Takes uncertainty into account
  *Bayesian Occam's razor:* Automatic penalty for model complexity
- Computational challenge: Posterior $p(\mathbf{z}|\mathbf{x})$ often intractable
  - Sampling algorithms
  - Variational approximations

# Machine Learning II

Machine Learning II course ... Focus on Bayesian methods

- Motivation: Bayesian vs frequentist statistics
- Probability theory: Conjugate priors
- Model selection: Marginal likelihood, sparsity priors
- Modeling:
  - Latent variable models
  - Bayesian non-parametrics: Gaussian processes
  - Deep neural networks
- Algorithms: EM, sampling methods

Potential applications

- Social data: Voting results, network models
- Economic data: GDP forecasting, volatility modeling
- Computer vision: Detection, tracking, recognition, segmentation
- ...