Machine Learning I and II

N. Bertschinger
M. Kaschube
V. Ramesh
Organization

- Lecture:
  Monday 10:00-12:00, SR 9

- Tutorial & problem session:
  Monday 12:00-14:00, SR 9

Weekly or biweekly problem sets
Tutorials on special topics

My email: kaschube@fias.uni-frankfurt.de
Schedule

WiSe 16/17
• 17. 10.
  Kaschube
• 24. 10.
  Bertschinger
• 31. 10.
  Ramesh
• 07. 11.
  Kaschube (Methods)
• 14. 11.
  Kaschube (Methods / applications)
• ...
• Last part of course: Ramesh (Applications / methods)

SoSe 17:
Bertschinger (Methods / applications)
...
Ramesh (Applications / methods)
Machine learning and pattern recognition
Stock market
Surveillance
Surveillance
Surveillance
Identity authentication
Handwritten digit recognition

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

• It was created by "re-mixing" the samples from NIST's original datasets.
• The creators felt that since NIST's training dataset was taken from American Census Bureau employees, while the testing dataset was taken from American high school students, NIST's complete dataset was too hard.
• The database contains 60,000 training images and 10,000 testing images.
## Performance

<table>
<thead>
<tr>
<th>Type of classifier</th>
<th>Authors</th>
<th>Error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear classifier</td>
<td>LeCun et. al, IEEE 1998</td>
<td>7.6</td>
</tr>
<tr>
<td>Non-Linear Classifier</td>
<td>LeCun et. al, IEEE 1998</td>
<td>3.3</td>
</tr>
<tr>
<td>Boosted Stumps</td>
<td>Kégl et. al, ICML 2009</td>
<td>0.87</td>
</tr>
<tr>
<td>Support vector machines</td>
<td>DeCoste &amp; Schölkopf, MLJ 2002</td>
<td>0.56</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>Keysers et. al, IEEE PAMI 2007</td>
<td>0.52</td>
</tr>
<tr>
<td>Neural network</td>
<td>Ciresan et. al, Neural Comput 2010</td>
<td>0.35</td>
</tr>
<tr>
<td>Convolutional neural network</td>
<td>Ciresan et. al, CVPR 2012</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Speech recognition

What can I help you with?
Characteristics of the data

• US English:
  – Voice Search, Voice Typing, Read data
• Billions of training samples
• Input: log-energy filter bank outputs
  – 40 frequency bands
  – 26 input frames
• Output: 8000 phone states

(Zeiler et al., ICASSP 2013; LeCunn & Ranzato, ICML 2013)
A typical speech recognition system

We focus only on the prediction of phone states from short time-windows of spectrogram.

(LeCunn & Ranzato, ICML 2013)
Note: The brain itself is a remarkable pattern recognition system - and it has provided significant inspiration on artificial recognition systems.
Krizhevsky et al., NIPS 2012

Deep Learning: Motivated from Human Learning

Learn massive data simple functions Multi-layered

(Krizhevsky et al., 2012)
Try it out for yourself!

http://www.image-net.org/
http://www.clarifai.com/
https://images.google.com/?gws_rd=ssl

Upload your favorite image and let it recognize its content and/or find similar images!
Measuring brain activity

Example: representation of space in rat hippocampus (Nobel Prize for Medicine, 2014)
• Raster plot of our 36 spike trains
• Each row corresponds to a spike train, with small, black, vertical lines indicating the time of individual spikes.
• The vertical red lines indicate the start and end of the maze exploration period.
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)
Medical applications
Computed Tomography (CT)
Computed Tomography (CT)

• A narrow beam of x-rays is aimed at a patient and quickly rotated around the body, producing cross-sectional images of the body.

• CT scans can be used to identify disease or injury within various regions of the body.

• CT image reconstruction is an ill-posed inverse problem. It may be solved using unsupervised learning techniques such as Filtered Back-Projection (FBP)
Magnetic Resonance Imaging (MRI)

(Keating, 2015)
Magnetic Resonance Imaging (MRI)

- MRI can detect tumors at early stages, revealing parts of the body that are not easily shown with other techniques.

- In the clinic, classification methods such as LDA, K-means, and SVMs are frequently used to detect tumors in early stages, providing an important support for radiologists’ diagnosis.
Slightly more abstract...
Two approaches to pattern recognition

- Data-driven
- Model-driven

Frequentist vs. Bayesian*

*Thomas Bayes 1701–1761
Classification

Linear Regression of 0/1 Response

FIGURE 2.1. A classification example in two dimensions. The classes are coded as a binary variable (\textcolor{blue}{BLUE} = 0, \textcolor{orange}{ORANGE} = 1), and then fit by linear regression. The line is the decision boundary defined by $x^T \hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as \textcolor{orange}{ORANGE}, while the blue region is classified as \textcolor{blue}{BLUE}. 
Classification

15-Nearest Neighbor Classifier

**FIGURE 2.2.** The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.
Classification

FIGURE 2.3. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then predicted by 1-nearest-neighbor classification.
Bayesian classification

FIGURE 2.5. The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).
Curve fitting (regression)
Curve fitting (regression)

1st Order Polynomial
Curve fitting (regression)

$3^{rd}$ Order Polynomial

$M = 3$
Curve fitting (regression)

Overfitting, a major issue!
Bayesian Curve Fitting

- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points
Bayesian Curve Fitting

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points
Supervised vs. unsupervised pattern recognition: The two major directions

- **Supervised**: Patterns whose class is known a-priori are used for training.
- **Unsupervised**: The number of classes is (in general) unknown and no training patterns are available.
Topics ML1
(WiSe16/17; Kaschube and Ramesh)

Topics:
• Math Recap (linear algebra, probability)
• Probability distributions
• Linear Regression, Bayesian linear Regression
• Classification
• Neural Networks, deep learning
• Latent variables
• Decision trees, randomized forests
• Clustering

Applications:
• Neuroscience and bio-imaging data
• Vision Applications for detection, tracking, recognition, segmentation, registration, geometry estimation, etc.
Topics ML2
(SoSe16/17; Bertschinger and Ramesh)

• Motivation: Bayesian vs frequentist statistics
• Decision theory: handling uncertainty, loss functions
• Probability theory: Conjugate priors
• Modeling: Latent variables, hierarchical models, Bayesian nonparametrics
• Model selection: Marginal likelihood, sparsity priors
• Algorithms: Variational Bayes (ELBO), sampling
• Mixture models, hmm's, EM algorithm
• Graphical model based inference, MCMC based sampling

Applications:
• Social data: Voting results, network models
• Economic data: GDP forecasting, volatility modeling
• Vision Applications for detection, tracking, recognition, segmentation, registration, geometry estimation, etc.
• Other applications in language and/or speech processing will be covered if time permits.
Books 1

Pattern Recognition and Machine Learning
Christopher M. Bishop
Springer
Books 2

An Introduction to Statistical Learning with Applications in R

Authors: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Springer
Books 3

Computer vision: models, learning and inference, Simon J.D. Prince, Cambridge University
Video lectures

E.g. online course by Andrew Ng (Stanford):

https://www.youtube.com/view_play_list?p=A89DCFA6ADACE599

https://www.youtube.com/watch?v=qeHZOdJvFU&index=1&list=PLZ9qNFMHZ-A4rycgrgOYma6zxF4BZGGPW
eduVote is an Audience Response System for the Academic Environment:

The use of ARS, also known as a TED-System or voting system, is considered very beneficial in large lecture halls and various other teaching arenas:

- Students are engaged through active participation in the material, thus increasing their attention span.
- As the students must give precise answers, they gain an awareness of where they may have knowledge gaps.
- In addition, the instructor is able to gain a quick overview of the audience’s current knowledge on the subject being discussed.

In comparison with other voting systems that require proprietary hardware (e.g., a hand-held clicker), eduVote is very cost-efficient since it provides Apps that run on a student’s laptop or smartphone. Thus minimizing the time and effort required to organize and distribute equipment and eradicating any purchase or maintenance costs.

In comparison to a web-based system, we take privacy extremely seriously. The eduVote server does not receive data regarding the instructor’s question or the student’s voting results. The question and answers are stored locally on the instructor’s local machine. We are aware and respect that instructors value this control over their questions and results.

Anything else? eduVote incurs no usage-based costs! eduVote can be integrated into PowerPoint for Windows! And, since 2011, eduVote has been successfully used at a number of universities across Germany, Austria and Switzerland.

eduVote Testimonials: evaluation and feedback on eduVote can be viewed [here](http://www.eduvote.de/en/).
Books 4

The Elements of Statistical Learning - Data Mining, Inference, and Prediction
Authors: Trevor Hastie, Robert Tibshirani, Jerome Friedman
Springer