ADVANCED MACHINE LEARNING

Introduction

Lecturer: Prof. Aude Billard (aude.billard@epfl.ch)

Teaching Assistants:
Guillaume de Chambrier, Nadia Figueroa,
Denys Lamotte, Nicola Sommer

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Course Format

Alternate between:

• Lectures + Exercises and Practice Sessions:
  9h15-11h00 (Thursdays)
  10h15-11h00 (Fridays)

• Check class timetable!!

http://lasa.epfl.ch/teaching/lectures/ML_MSc_Advanced/
# Class Timetable

<table>
<thead>
<tr>
<th>Date</th>
<th>Related Documentation</th>
<th>Course Topics</th>
<th>Exercises/TP Template</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td></td>
<td><strong>Concepts</strong> (classification, regression, pattern recognition, etc.)</td>
<td></td>
</tr>
<tr>
<td>Thursday (Feb 25th)</td>
<td></td>
<td><strong>Methodology</strong> (training/testing sets, leave one out, crossvalidation, ROC, how to measure significance, etc)</td>
<td></td>
</tr>
<tr>
<td>Friday (Feb 26th)</td>
<td></td>
<td>Recap Statistics (Probability Distribution, Likelihood, E-M, etc.)</td>
<td></td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Kernels</td>
<td></td>
</tr>
<tr>
<td>Thursday (March 3rd)</td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Kernel PCA</td>
<td></td>
</tr>
<tr>
<td>Friday (March 4th)</td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Kernel PCA</td>
<td></td>
</tr>
<tr>
<td>Friday (March 10th)</td>
<td></td>
<td>MLDemos and mini programming projects introduction</td>
<td></td>
</tr>
<tr>
<td><strong>Week 3</strong></td>
<td></td>
<td><strong>Practical session on computer</strong>: Kernel PCA</td>
<td></td>
</tr>
<tr>
<td>Thursday (March 11)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Week 4</strong></td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Kernel CCA, kernel K-means</td>
<td></td>
</tr>
<tr>
<td>Thursday (March 17th)</td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Spectral Clustering</td>
<td></td>
</tr>
<tr>
<td>Friday (March 18th)</td>
<td></td>
<td><strong>Spectral methods</strong>: Structure Discovery, Dim. Reduction Spectral Clustering</td>
<td></td>
</tr>
<tr>
<td><strong>Week 5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Grading

50% of the grade based on personal work. Choice between:

1. Mini-project implementing and evaluating the algorithm performance and sensibility to parameter choices (should be done individually preferably).

OR

2. A literature survey on a topic chosen among a list provided in class (can be done in team of two people, report conf. paper format, 8 pages, 10pt, double column format)

~25-30 hours of personal work, i.e. count one full week of work.

50% based on final oral exam

  20 minutes preparation
  20 minutes answer on the black board
  (closed book, but allowed to bring a recto-verso A4 page with personal notes)
Prerequisites

Linear Algebra, Probabilities and Statistics

Familiar with basic Machine Learning Techniques:
- Principal Component Analysis
- Clustering with K-means and Gaussian Mixture Models
- Linear and weighted regression
Today’s class content

- Taxonomy and basic concepts of ML
- Examples of ML applications
- Overview of class topics
Main Types of Learning Methods

• *Supervised learning* – where the algorithm learns a function or model that maps best a set of inputs to a set of desired outputs.

• *Reinforcement learning* – where the algorithm learns a policy or model of the set of transitions across a discrete set of input-output states (Markovian world) in order to maximize a reward value (external reinforcement).

• *Unsupervised learning* – where the algorithm learns a model that best represent a set of inputs without any feedback (no desired output, no external reinforcement).
Main Types of Learning Methods

- **Supervised learning** – where the algorithm learns a function or model that maps best a set of inputs to a set of desired outputs.

Classes of supervised algorithms:
- Classification
- Regression
Classification: Principle

Map N-dim. input $x \in \mathbb{R}^N$ to a set of class labels $y \in \mathbb{N}_{\{1,\ldots,C\}}$.

C: number of classes

If two classes (binary classification), learn a function of the type:

$h : \mathbb{R}^N \rightarrow \mathbb{R}$ and $y = \text{sgn}(h(x))$.

Train on subset of datapoints

Hope to generalize class labels to unseen datapoints
Classification: overview

Classification is a supervised clustering process.

Classification is usually multi-class; Given a set of known classes, the algorithm learns to extract combinations of data features that best predict the true class of the data.

Original Data  After 4-class classification using Support Vector Machine (SVM)
Classification: example

Classification of finance data to assess solvability using Support Vector Machine (SVM).

| The set of attributes taken from the credit reports as the potential features. |
|---------------------------------|---------------------------------|---------------------------------|
| Consolidation class (three classes) | Total assets of the last year | Class of sales turnover of the last year |
| Profit before tax of the last year | Total liabilities at the end of the last year | Trend of sales turnover (the slope of linear model) |
| Current assets of the last year | Trend of profit after tax | Trend of shareholders’ equity |
| Shareholders’ equity | Age of company (in years) | Number of employees |
| Availability of company by phone | Number of employees for group | Class of sales turnover |
| Availability of company by fax | Class of positive profit after tax | Registry status |
| Legal form class | Credit limit | Current ratio |
| Is group employment | Share capital (authorized or issued/paid-up) |  

5-classes to measure insolvency risks:

- Excellent
- Good
- Satisfactory
- Passable
- Poor (credit)

Classification: issues

Classification of finance data to assess solvability using Support Vector Machine (SVM).


The features are crucial. If they are poor, poor learning

→ Usually determine features by hand (prior knowledge)
→ Novel approaches learn the features

5-classes to measure insolvency risks:
Excellent, good, satisfactory, passable, poor (credit)
Classification: issues

A recurrent problem when applying classification to real life problems is that classes are often unbalanced.

The quantity of cases belonging to different groups of insolvency risk.

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>211</td>
</tr>
<tr>
<td>Good</td>
<td>800</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>783</td>
</tr>
<tr>
<td>Passable</td>
<td>332</td>
</tr>
<tr>
<td>Poor</td>
<td>91</td>
</tr>
<tr>
<td>Total:</td>
<td>2217</td>
</tr>
</tbody>
</table>

Swiderski et al, 2012

More samples from positive class than from negative class

This can affect drastically classification performance, as classes with many data points have more influence on the error measure during training.
Classification Algorithms in this Course

- Support Vector Machine
- Relevance Vector Machine
- Boosting – random projections
- Boosting – random gaussians
- Random forest
- Gaussian Process
Regression: Principle

Map N-dim. input \( x \in \mathbb{R}^N \) to a continuous output \( y \in \mathbb{R} \).
Learn a function of the type:
\[
f : \mathbb{R}^N \to \mathbb{R} \text{ and } y = f(x).
\]

Estimate \( f \) that best predicts set of training points \( \left\{ x^i, y^i \right\}_{i=1,...,M} \)?
Regression: Issues

Map N-dim. input $x \in \mathbb{R}^N$ to a continuous output $y \in \mathbb{R}$. Learn a function of the type:

$$f : \mathbb{R}^N \rightarrow \mathbb{R} \quad \text{and} \quad y = f(x).$$

Estimate $f$ that best predict set of training points $\{x^i, y^i\}_{i=1,\ldots,M}$.

Fit strongly influenced by choice of:
- datapoints for training
- complexity of the model (interpolation)
Regression: example

Predicting the optimal position and orientation of the golf club to hit the ball so as to sink it into the goal.

Kronander, Khansari and Billard, JTSC award, IEEE Int. Conf. on Int. and Rob. Systems 2011.
Regression: example

Predicting the optimal position and orientation of the golf club to hit the ball so as to sink it into the goal.

Kronander, Khansari and Billard, JTSC award, IEEE Int. Conf. on Int. and Rob. Systems 2011.
Regression: example

Contrast prediction of two methods (Gaussian Process Regression and Gaussian Mixture Regression) in terms of precision and generalization.

Kronander, Khansari and Billard, JTSC award, IEEE Int. Conf. on Int. and Rob. Systems 2011.
Regression Algorithms in this Course

- Support vector regression
- Relevance vector regression
- Gaussian process regression
- Gradient boosting
- Locally weighted projected regression
Main Types of Learning Methods

- **Supervised learning** – where the algorithm learns a function or model that maps best a set of inputs to a set of desired outputs.

- **Reinforcement learning** – where the algorithm learns a policy or model of the set of transitions across a discrete set of input-output states (Markovian world) in order to maximize a reward value (external reinforcement).

- **Unsupervised learning** – where the algorithm learns a model that best represent a set of inputs without any feedback (no desired output, no external reinforcement).
Reinforcement Learning: Principle

Generate a series of input for \( t \) time steps by performing action \( a \) to go from state \( x_{t-1} \) to state \( x_t \):

\[
\begin{align*}
x_1 & \xrightarrow{a_1} x_2 \xrightarrow{a_2} \ldots \xrightarrow{a_{t-1}} x_t
\end{align*}
\]

Collect data \( D = \{x_1, x_2, \ldots, x_t, a_1, a_2, \ldots, a_{t-1}\} \)

Receive a reward at \( t \):

\( r_t = 1 \)

Map the reward to choices of actions \( a \) for each state (Credit Assignment Problem)

Repeat the procedure (generate a series of trials and errors) until you converge to a good map that represents the best choice of action in any state to maximize total rewards.
Reinforcement Learning: Issues

Generate a series of input for t time steps by performing action $a$ to go from state $x_{t-1}$ to state $x_t$:

$x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow x_t$

Collect data $D = \{x_1, x_2, \ldots, x_t, a_1, a_2, \ldots, a_{t-1}\}$

Receive a reward at $t$:

$r_t = 1$

Map the reward to choices of actions $a$ for each state (Credit Assignment Problem)

Repeat the procedure (generate a series of trials and errors) until you converge to a good map that represents the best choice of action in any state to maximize total rewards.

Number of states and actions may be infinite (continuous world).

In discrete world, visiting all action-state pairs may be prohibitive.
Reinforcement Learning: example

Goal: to stand up
Reward: 1 if up, 0 otherwise
Requires 750 trials in simulation + 170 on real robot to learn

First set of trials
At convergence

Morimoto and Doya, Robotics and Autonomous Systems, 2001
Reinforcement Learning Algorithms in this Course

Random walk

Dynamic programming

Power

Genetic algorithm

Discrete and continuous state reinforcement learning methods

Above: solutions found by 4 different techniques for learning
Main Types of Learning Methods

• **Supervised learning** – where the algorithm learns a function or model that maps best a set of inputs to a set of **desired** outputs.

• **Reinforcement learning** – where the algorithm learns a policy or model of the set of transitions across a **discrete** set of input-output states (Markovian world) in order to maximize a **reward** value (external reinforcement).

• **Unsupervised learning** – where the algorithm learns a model that best represent a set of inputs without any feedback (no desired output, no external reinforcement)
Unsupervised Learning: Principle

Observe a series of input: \( D = \{x_1, x_2, \ldots, x_M\}, \ x_1 \in \mathbb{R}^N \)

Find regularities in the input, e.g. correlations, patterns.

Use these to reduce the dimensionality of the data (with correlations) or to group datapoints (patterns).
Unsupervised learning ~ Structure discovery

Raw Data

Trying to find some structure in the data…..

Pattern Recognition
Pattern Recognition: Similar Shape

Results of decomposition with Principal Component Analysis: eigenvectors
Pattern Recognition: Similar Shape

Encapsulate main differences across groups of images (in the first eigenvectors)
Pattern Recognition: Similar Shape

Detailed features (glasses) get encapsulated next (in the following eigenvectors)
Pattern recognition techniques of the type of PCA can be used to reduce the dimensionality of the dataset. Data are classifiable in a lower dimensional space (here plotted on 1\textsuperscript{st} and 2\textsuperscript{nd} eigenvector).
Pattern recognition techniques of the type of PCA can be used to reduce the dimensionality of the dataset. Subclasses can be discovered by looking at other combinations of projections (here plotted on 2\textsuperscript{nd} and 3\textsuperscript{rd} eigenvector).
Pattern Recognition: Determining Trends

Goal: Predicting evolution of prices of stock market

Issues:
  - Data vary over time
  - Time is the input variable
Pattern Recognition: Determining Trends

Goal: Predicting evolution of prices of stock market

Challenges:
→ are previous data meaningful to predict future data?
• Data are very noisy; each datapoint is seen only once
  → Must decouple noise from temporal variations of time series
Pattern Recognition: Clustering

Clustering for automatic processing of medical images: enhance visibility

Multispectral medical image segmentation. (left: MRI-image from 1 channel) (right: classification from a 9-cluster semi-supervised learning); Clusters should identify patterns, such as cerebro-spinal fluid, white matter, striated muscle, tumor. (Lundervolt et al, 1996).

Cluster groups of pixels by tissue type
→ use grey shadings to denote different groups
Pattern Recognition: Clustering

Clustering assume groups of points are similar according to the *same* metric of similarity.

Challenge:
Clustering may fail when metric changes depending on the region of space.

Jain, 2010, Data clustering: 50 years beyond K-means, Pattern Recognition Letters
Pattern Recognition: Clustering

Different techniques or heuristics can be developed to help the algorithm determine the right boundaries across clusters:

Add manually set of constraints across pairs of datapoints
  → determine links between datapoints in same/different clusters
  → use semi-supervised clustering

Jain, 2010, Data clustering: 50 years beyond K-means, Pattern Recognition Letters
Pattern Recognition: Clustering

Clustering encompasses a large set of methods that try to find patterns that are similar in some way.

Hierarchical clustering builds tree-like structure by pairing datapoints according to increasing levels of similarity.
Hierarchical clustering can be used with arbitrary sets of data.

**Example:**

Hierarchical clustering to discover similar temporal pattern of crimes across districts in India.


**Fig. 3.** Dendrogram of Crime Against Body for 2002-2006 using DTW with Parametric Minkowski Model
Unsupervised Learning Algorithms in this Course

- PCA
- Kernel PCA
- Genetic algorithm
- Kernel K-means
- Isomap
- Laplacian map
- Swissroll example
Main Types of Learning Methods

- **Supervised learning** – where the algorithm learns a function or model that maps best a set of inputs to a set of desired outputs.

- **Unsupervised learning** – where the algorithm learns a model that best represent a set of inputs without any feedback (no desired output, no external reinforcement).
Data Mining

Pattern recognition with very large amount of high-dimensional data

(Tens of thousands to billions) (Several hundreds and more)
Data Mining: examples

Mining webpages

- Cluster groups of webpage by topics
- Cluster links across webpages

Other algorithms required:
- Fast methods for crawling the web
- Text processing (Natural Language Processing)
- Understanding semantics

Issues:
- Domain-specific language / terminology
- Foreign languages
- Dynamics of web (pages disappear / get created)
Data Mining: examples

Mining webpages

Personalizing search
• Improvising user’s experience
• Enhance commercial market value

→ Build user-specific data
• country of origin
• languages spoken
• list of websites visited in the past days/weeks
• frequency of occurrence of some queries

→ Link user-specific data to other groups of data
• e.g. robotics in Switzerland
• academic institutions in Europe
Data Mining Main Issue: Curse of Dimensionality

Computational Costs

Computational costs may also grow as a function of number of dimensions

ML techniques seek linear growth whenever possible

Apply methods for dimensionality reduction whenever possible
Some Machine Learning Resources

http://www.machinelearning.org/index.html
• http://www.pascal-network.org/ Network of excellence on Pattern Recognition, Statistical Modelling and Computational Learning (summer schools and workshops)

Databases:
• http://archive.ics.uci.edu/ml/

Journals:
• Machine Learning Journal, Kluwer Publisher
• IEEE Transactions on Signal processing
• IEEE Transactions on Pattern Analysis
• IEEE Transactions on Pattern Recognition
• The Journal of Machine Learning Research

Conferences:
• ICML: int. conf. on machine learning
• Neural Information Processing Conference – on-line repository of all research papers, www.nips.org
Topics for Literature survey and Mini-Projects

The exact list of topics for lit. survey and mini-project will be posted by March 4

Topics for survey will entail:

• Data mining methods for crawling mailboxes
• Data mining methods for crawling git-hub
• Ethical issues on data mining
• Method for learning the features
• Metrics for clustering

Topics for mini-project will entail implementing either of these:

- Manifold learning (Isomap, Eigenmaps, Local Linear Embedding)
- Classification (Random Forest)
- Regression (Relevance Vector Regression)
- Non-Parametric Approximations Techniques for Mixture Models

Programming language open: matlab, C/C++, Python, or embed into MLDemo software
Overview of Practicals
Summary of today’s class content

• Taxonomy and basic concepts of ML
  • Structure discovery, classification, regression
  • Supervised, unsupervised and reinforcement learning
  • Data mining and curse of dimensionality

• Examples of ML applications
  • Pattern recognition in images
  • Data mining of the web

• Overview of projects and survey topics