T-61.3050 Machine Learning: Basic Principles

Introduction

Kai Puolamäki

Laboratory of Computer and Information Science (CIS)
Department of Computer Science and Engineering
Helsinki University of Technology (TKK)

Autumn 2007

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course
2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course

2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

People and Locations

- People:
  - Kai Puolamäki, PhD, lecturing researcher, lecturer.
  - Antti Ukkonen, MSc, course assistant.
- Please see the course web site at `http://www.cis.hut.fi/Opinnot/T-61.3050/2007/` for current information.
- If you want to send email related to the course please use the email alias `t613050@james.hut.fi` (not personal addresses).
- Lectures: in T1 on Tuesdays at 10–12 (11 September to 11 December 2007, no lecture on 30 October).
- Problem sessions: in T1 on Fridays at 10–12 (from 14 September to 7 December, no problem session on 26 October; problem sessions not every week).

Participating

- To participate to this course you need to be a registered student at TKK (that is, you need a student number).
- You must sign in to course using WebTOPI, `https://webtopi.tkk.fi/` Please sign in today, if you have not already done it.
- You will need to have an addresses of form `12345X@students.hut.fi`, where `12345X` is your student number (for exam results, exercise work feedback etc.). Check that this address works (if not, you should contact the student registry and update your email address there!).

Prerequisites

- To participate to this course you need to have the following prerequisite knowledge:
  - basic mathematics and probability courses (Mat-1.1010, Mat-1.1020, Mat-1.1031/1032 and Mat-1.2600/2620; or equivalent);
  - basics of programming (T-106.1200/1203/1206/1207 or equivalent); and
  - data structures and algorithms (T-106.1220/1223 or equivalent).
- If you lack this prerequisite knowledge we strongly encourage you to take the above mentioned courses before participating to this course!
- You should be able to complete the problems in the prerequisite knowledge test (problem 1) for the first problem session next Friday (see the instructions in the problem sheet).

How to Pass the Course

- You will get 5 cr for passing this course.
- Requirements for passing the course:
    - Pass the exercise work. The exercise work should be submitted by 2 January 2008. More instructions will appear in a few weeks time.
    - Pass the examination. You can participate to the examination after passing the exercise work (exception: you can participate to the December examination before passing the exercise work; you'll then pass the course if you pass the exercise work).
- Optional, but useful:
    - Lectures.
    - Problem sessions.
    - Reading the book and other material.

About Exercise Work

- Detailed instructions for the exercise work will be announced within a couple of weeks.
- The exercise work will include a data analysis challenge.
- The final report, which should describe the methods you have used and your results, should be submitted at 2 January 2008, at latest.
- You can submit the results of the data analysis challenge by 1 December 2007.
- You must pass the exercise work to pass the course. You will get an increase to your grade if your report is well done. You get some extra points if you additionally perform well in the data analysis challenge.

About Examination

- The examinations are currently scheduled as follows:
  - In B at 16–19 on 19 December 2007.
  - In * at 10–13 on 2 February 2008.
  - In T1 at 13–16 on 15 May 2008.
- Check the exam schedule later, times may still change!
- You must pass the exercise work before participating to the examination (exception: you can participate to the December examination before passing the exercise work; you'll then pass the course if you pass the exercise work).
- You must sign in to the examination at least one week in advance using WebTOPI, `https://webtopi.tkk.fi/`
- The examination will be based on the parts of the Alpaydin's book discussed in the lectures, plus on the PDF chapter to be distributed from the course web site.
- Lectures, problem sessions and doing the exercise work help.

How to Get a Grade

- You need to pass both the exercise work and the examination to pass the course.
- You will get a grade of 1–5 based mainly on the examination. You can increase your grade by...
  - Participating to the problem sessions diligently.
  - Solving the exercise work well.
  - Submitting a good answer by 1 December 2007 to the data analysis challenge of the exercise work.

Literature

- The course follows a subset of the book: *Alpaydin, 2004. Introduction to Machine Learning. The MIT Press.*
- Additionally, there will also be a PDF chapter on algorithmics (complexity of problems, local minima etc.) to be distributed from the course web site.
- The lecture slides are available for download from the course web site. I have also given Edita a permission to print them on request.
- You might also find the material — especially the errata and slides — at the Alpaydin's web site (see the link at the course web site) useful.

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course
2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

Relation to the Old Courses

- The CIS course reform: more weight on the principles of machine learning, less weight to the neural networks beginning Autumn 2007.
- In curriculum and for the purposes of the degree requirements, this course replaces the old course T-61.3030 (and T-61.261) Principles of Neural Computing.
- However, the contents of this course have little overlap with the old course T-61.3030 Principles of Neural Computing.

Relation to the Old Courses

Old course (before Autumn 2007)	**New course**
T-61.3030 Principles of Neural Computing	T-61.3050 Machine Learning: Basic Principles
T-61.5030 Advanced Course in Neural Computing	T-61.5130 Machine Learning and Neural Networks
T-61.5040 Learning Models and Methods	T-61.5140 Machine Learning: Advanced Probabilistic Methods

Table: Correspondences in degree requirements.

Old course (before Autumn 2007)	**New course**
T-61.5040 Learning Models and Methods	T-61.3050 Machine Learning: Basic Principles T-61.5140 Machine Learning: Advanced Probabilistic Methods
T-61.3030 Principles of Neural Computing T-61.5030 Advanced Course in Neural Computing	T-61.5130 Machine Learning and Neural Networks

Table: *Approximate* topical correspondeces.

See `http://www.cis.hut.fi/Opinnot/T-61.3050/oldcourses`

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course
2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

Very Preliminary Plan of the Topics

- Supervised learning, Bayesian decision theory, probability distributions and parametric methods, multivariate methods, clustering (mostly Alpaydin's chapters 1–7 and appendix A)
- Algorithmic issues in machine learning, such as hardness of problems, approximation techniques and their features (such as local minima), time and memory complexity in data analysis (separate PDF chapter to be distributed from the course web site)
- Nonparametric methods (Alpaydin 8.1–8.2), linear discrimination (Alpaydin 10.1–10.8), assessing and comparing classification algorithms (Alpaydin's chapter 14)
- I'll try to keep the Alpaydin's ordering of topics, and emphasize principles rather than to go through all possible algorithms and methods.

What You Should Know After the Course

- After this course, you should...
  - be able to apply the basic methods to real world data;
  - understand the basic principles of the methods; and
  - have necessary prerequisites to understand and apply new concepts and methods that build on the topics covered in the course.
- This course does not include:
  - all possible machine learning methods; or
  - all possible applications of machine learning.

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course
2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

What is Machine Learning?

Definition

Machine learning is programming computers to optimize a performance criterion using example data or past experience. (Alpaydin)

?

Examples of Applications

- Associations (basket analysis)
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
- Reinforcement learning (not in this course)

Association rules
Sales data

- Example: sales data
  - rows: customer transactions (millions)
  - columns: products bought (thousands)
- Question: Can you find something interesting of this?

Association rule

"80% of customers who buy beer and sausage buy also mustard."
Or: $P(\text{mustard} \mid \text{beer}, \text{sausage}) = 0.8$.

- Accuracy (conditional probability): 0.8
- Frequency or support (fraction of clients who bought mustard, beer and sausage): 0.3

Classification
Credit scoring

- Example: data on credit card applicants
- Question: Should a client be granted a credit card?
- Differentiate between low-risk (+) and high-risk (-) customers using their *income* and *savings*.

Discriminant

IF *income*$> \theta_1$ AND *savings*$> \theta_2$ THEN *low-risk* ELSE *high-risk*.

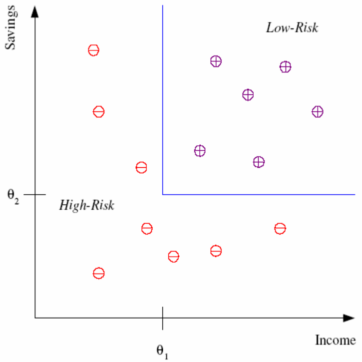


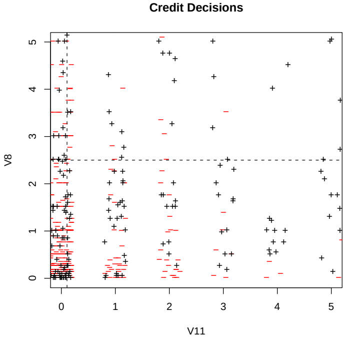
Figure 1.1 of Alpaydin (2004).

Classification
Credit scoring

- Example: data on credit card applicants
- Question: Should a client be granted a credit card?
- Differentiate between low-risk (+) and high-risk (-) customers using their *income* and *savings*.

Discriminant

IF *income*$> \theta_1$ AND *savings*$> \theta_2$ THEN *low-risk* ELSE *high-risk*.



Real CREDIT-SCREENING data from UCI Machine Learning Repository.

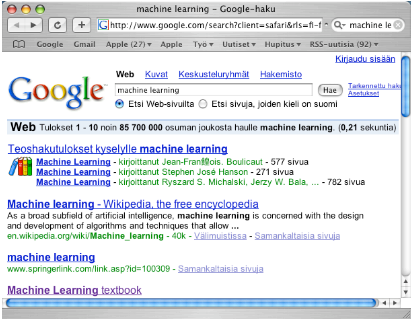
Classification

- Classification: predict something (variate, Y), given something else (covariate, X). Or: try to estimate $P(Y \mid X)$.
- Speech recognition: temporal dependency. Predict words, given the speech signal.
- Character recognition (OCR): different handwriting styles.
- Medical diagnosis: from symptoms to diagnosis.
- Eye movement analysis: is the user interested in the text she is reading?
- ...

Classification

Search engines

- The Internet search engines use machine learning to give the best search results, given a query.
- Fundamental problem in information retrieval: given a query (“machine learning”), list relevant documents (web sites related to “machine learning”).



Classification

Search engines and eye movements

[movie, link]

Classification

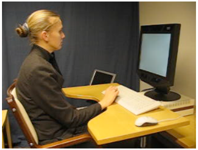
Eye movements

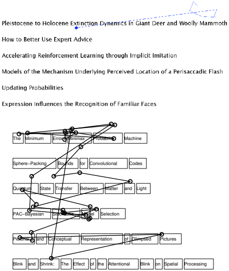
- Example: eye movement measurements during information search (ongoing research by the lecturer and his friends during 2003–2007, see http://www.cis.hut.fi/projects/mi/proact)
- Question 1: Is the user interested in text she is reading?
- Question 2: What is the user interested in?
- This is a classification problem: predict relevance of a viewed document or true interest of the user, given the eye movement trajectory.
- The problem is (was) quite difficult to solve.

Classification

Is the user interested in the text she is reading?

- Eye movements are measured in a controlled experiment.
- A sentence (title of a scientific article) is partitioned into words.
- Most discriminative word-specific features were used (one or many fixations, total fixation duration, reading behaviour).
- The title relevance was predicted using a discriminative machine learning models.





Classification

What is the user interested of?

Xenotarsosaurus ("strange-ankle lizard") is a little-understood theropod of the late Cretaceous (~83 - 73 mya). It probably weighed 0.7 - 1.0 tons.

The only fossil evidence consists of a small number of vertebrae and leg bones, retrieved from the Bajo Barreal Formation, Chubut, Argentina. From these samples, Martinez, Giminez, Rodriguez and Bochatey named the type species, X. bonapartei, in 1986. It was probably an allosaurid.

The history of film is one of the most rapidly moving of any artistic or communications medium ever, as befits perhaps the first great mass medium of the modern era. Film has gone through a remarkable array of changes and developed a remarkable variety and sophistication in barely more than one hundred years of existence.

A Post Office box is a uniquely-addressable lockable box located on the premises of a Post Office station. Generally, Post Office boxes are rented from the post office either by individuals or by businesses on a basis ranging from monthly to annual, and the cost of rent varies depending on the box size. CBD PO boxes are usually more expensive than a rural PO Box. In the United States, the rental rate used to be uniform across the country. Now, however, a postal facility can be in any of seven fee groups by location; in addition, certain postal patrons qualify for free box rental.

Among the recognizable Olympic symbols :

The Olympic flag: A white flag with the Olympic Rings on it in five colours.

The Olympic Flame: A flame burning day and night for the duration of the Olympic Games.

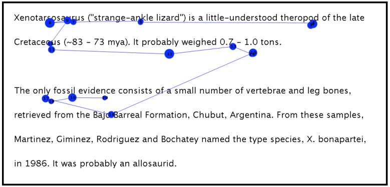
The Olympic Fanfare and Theme: A musical composition by John Williams.

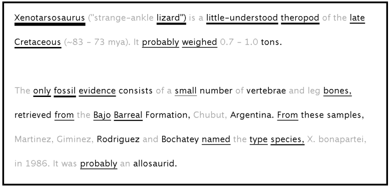
The Kotinos: A crown made from an olive branch, which can seen atop many statues of ancient Olympic victors.

Classification

What is the user interested of?

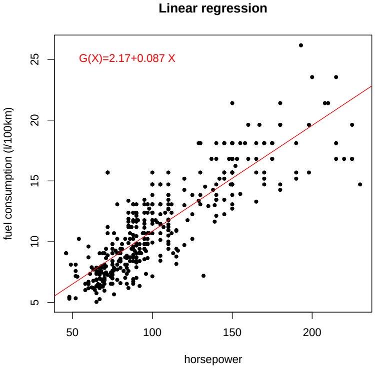
- For this to work, there must be a link between the relevance of a word to a topic of the user's interest and eye movements related to it.
- This link can be learned and used on new topics.





Regression

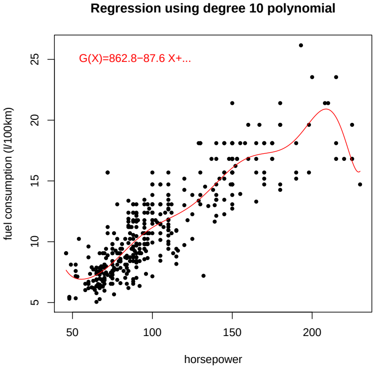
- Regression is classification where the variate Y is a continuous variable.
- The principles in classification and regression are the same, methods differ.
- Example: fuel consumption of cars.
- Y: fuel consumption.
- X: car attributes.
- $Y = G(X \mid \theta)$
  - $G()$: a model.
  - θ: model parameters.



AUTO-MPG data set from UCI Machine Learning Repository.

Regression

- Regression is classification where the variate Y is a continuous variable.
- The principles in classification and regression are the same, methods differ.
- Example: fuel consumption of cars.
- Y: fuel consumption.
- X: car attributes.
- $Y = G(X \mid \theta)$
  - $G()$: a model.
  - θ: model parameters.



AUTO-MPG data set from UCI Machine Learning Repository.

Uses of Supervised Learning

- Prediction of future cases: Use the rule to predict the output for future inputs.
- Knowledge extraction: The rule is easy to understand.
- Compression: The rule is simpler than the data it explains.
- Outlier detection: Exceptions that are not covered by the rule, for example, fraud.

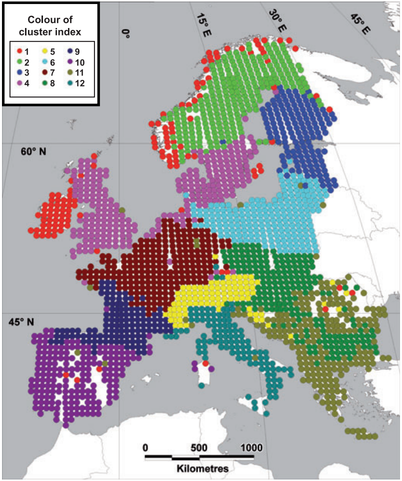
Unsupervised Learning

- In supervised learning, an imaginary "supervisor" tells us in the training phase what is the correct variate (Y), given the covariate (X). We then try to predict $P(Y \mid X)$ without the supervisor.
- Unsupervised learning is like supervised learning, except there is no supervisor telling us the Y. We try to predict $P(X)$. (In supervised learning we really do not care about $P(X)$.)
- Another view: unsupervised learning is like supervised learning, except the covariate Y is fixed, in which case we try to predict $P(Y \mid X) = P(Y)$.
- Again, the principles are the same, but the methods differ.
- Example: clustering (grouping similar instances together)
- Example: probabilistic modeling (find the most likely model to describe the data, given some prior family of models)

Clustering
European land mammals

- Example: European land mammals.
- Question: Can we find ecological communities?
- Question: What explains the communities?
- The 50 $\times$ 50 km map grids were grouped into clusters. Map grids within a cluster should occupy similar mammals.

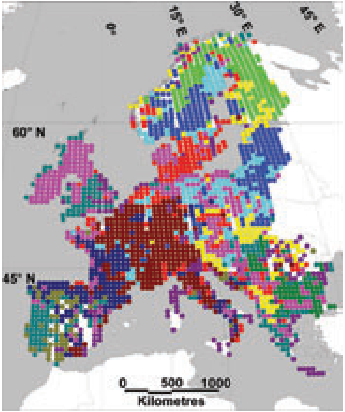
Heikinheimo et al. (2007) Biogeography of European land mammals. . . J Biogeogr.



Clustering
European land mammals

- Endangered species appear to have least spatial coherence.
- The clustering can be explained mostly by temperature and precipitation.
- Somewhat surprisingly the natural factors seem to explain the mammalian metacommunity distributions, despite a long history of intensive human presence.

At risk



Other Applications of Machine Learning

- Bioinformatics
- . . .

Reinforcement Learning

- Learning a policy: A sequence of output.
- No supervised output but delayed reward.
- Credit assignment problem.
- Game playing.
- Robot in a maze.
- Multiple agents, partial observability...
- Example: our search engine is showing an user documents. The user tells us if the shown document is interesting. Tradeoff:
  - *Exploitation:* show the user documents that we think might interest her most (immediate reward).
  - *Exploration:* show the user uninteresting documents with which we would learn more of her interests (delayed reward).
- Not covered in this course.

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course
2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

What is Machine Learning?

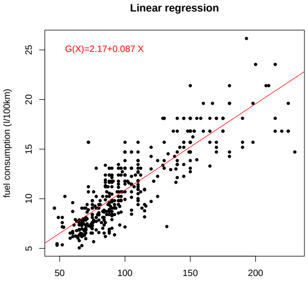
Definition

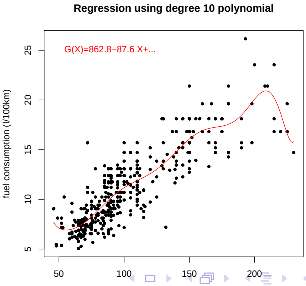
Machine learning is programming computers to optimize a performance criterion using example data or past experience. (Alpaydin)

- Machine learning is using computers to analyze data.
- The data is noisy, there are measurement errors etc.
- We usually do not observe all factors that would be needed for certainty: we must resort to statistics.
- What is "learning"? Often, we do not want just to describe the data we have, but be able to predict of (yet) unseen data.

About Generalization

- Often, it would be quite easy to make a model that would describe already known data.
- It is more difficult to. . .
  - Say something (predict) of yet unseen data (generalization).
  - Make a good (not too complex and not too simple) description of known data.
- Prior knowledge is important.





What is Machine Learning?
Related areas

- How does machine learning relate to data mining?
- How does machine learning relate to statistics?
- How does machine learning relate to algorithms?
- How does machine learning relate to artificial intelligence, neural networks, . . . ?

Machine Learning and Data Mining

- Machine learning has (depending on the speaker) a strong overlap with data mining.
- Machine learning emphasizes statistical principles and methods.
- Data mining emphasizes algorithms which also work on large data volumes.
- Data miners may also have a modest goal of helping user to find something interesting of the data, not attempting to make a model of the world.

Machine Learning and Statistics

- Modern statistics forms (with algorithms) the theoretical foundations of machine learning.
- In “traditional” statistics one typically tests single hypothesis of the data. Example: patients with a new treatment had 80% recovery rate, while patients with the old treatment had 60% recovery rate. Is the new treatment more effective than the old one?

Machine Learning and Algorithms

- Algorithms are needed to solve machine learning problems.
- In machine learning the algorithmic aspects (convergence, running times etc.) have not been emphasized. This is however changing.
- Summary: there are lots of connections between machine learning and various disciplines. The exact connections vary depending on whom you ask. The field is still developing.

Outline

1. Course Bureaucracy
   - General Information
   - Relation to Old Courses
   - Contents of the Course

2. Chapter 1: Introduction
   - Examples of Machine Learning Applications
   - What is Machine Learning?
   - Resources

Software

- There is lots of good software available. You will need some software to pass this course (for example, exercise work). Some examples follow.
- R. An open source software for statistical computing and publication quality graphics. An usable functional programming language. (Lecturer's favourite.)
- Matlab. Matlab is a commercial software that is especially popular in signal processing. It is too matrix-oriented for the lecturer's taste. Quite a few people use it (including Alpaydin), though. Matlab has an open source variant, GNU Octave.
- Weka. Open source Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. (Assistant seems to like it.)

Datasets

- Often, finding a good data set one of the most difficult tasks in developing machine learning methods.
- UCI Repository: `http://www.ics.uci.edu/~mlearn/MLRepository.html`
- UCI KDD Archive: `http://kdd.ics.uci.edu/summary.data.application.html`
- Statlib: `http://lib.stat.cmu.edu/`
- Delve: `http://www.cs.utoronto.ca/~delve/`

Journals

- Journal of Machine Learning Research
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Joint Conference on Artificial Intelligence (IJCAI)
- International Conference on Neural Networks (Europe)
- ...

Questions?

Next lecture

- Next Tuesday: Chapter 2 of Alpaydin (2004), "Supervised Learning".
- Remember the problem session next Friday at 10 o'clock.