Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 1:

Introduction

- What is Machine Learning?
  - A subfield of Artificial Intelligence.
    - Goal: to create computer systems that are intelligent (strong), or fool you into thinking they are (weak).
    - One thing we associate with (natural) intelligent entities: adaptive behaviour, and/or the ability to learn.
    - So machine learning is about creating computer systems that can learn to perform tasks.

- In other words, machine learning software automatically builds software to perform a task.
  - On the basis of…
    - **DATA!**
  - So machine learning is about engineering algorithms that learn from data to solve interesting problems.
  - "…the study of computer algorithms capable of learning to improve their performance on a task on the basis of their own experience."
  - This is sometimes called **inductive learning**.

- Learning is an extremely broad concept, studied by psychologists, neuroscientists and others.
- Machine learning restricts itself to a few well-defined classes of learning problems, that are still VERY general.
- Furthermore, we concentrate mainly on problems that involve numerical data, or at least "structural" data
  - e.g. nothing that involves language comprehension or "high-level" cognitive-type stuff!
  - Our problems might be called sub-symbolic, or low-level…

Is ML Important?

- In an interview with Hal Varian (Chief Economist, Google) from the New York Times (http://freakonomics.blogs.nytimes.com/2008/02/25/hal-varian-answers-your-questions/#more-2345)where he says:
- **Q:** Your job sounds extremely interesting. What jobs would you recommend to a young person with an interest, and maybe a bachelors degree, in economics?
- **A:** If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on.

Why "Learn" ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to "learn" to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

- But why?
  - Clearly, memorization is not the goal – you can use a lookup table for that!
  - If you can build a good model (from the data) of the process that generated the data, then
    - The model is likely to tell you something about the process you didn't know.
    - You can use the model **to produce outputs for new input data, that was not used to produce the model**.
    - This is prediction.

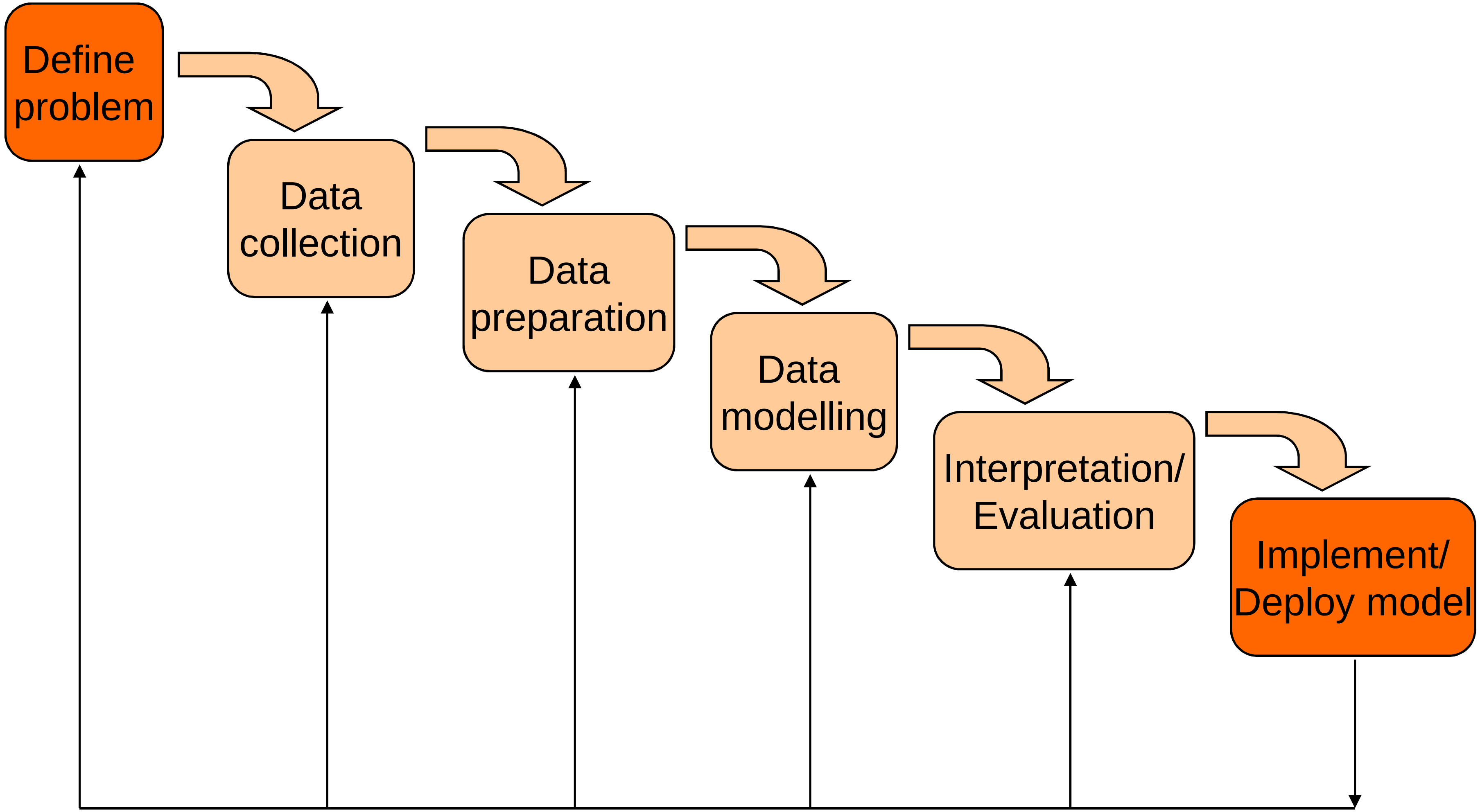
What We Talk About When We Talk About"Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought "Da Vinci Code" also bought "The Five People You Meet in Heaven" (www.amazon.com)
- Build a model that is *a good and useful approximation* to the data.

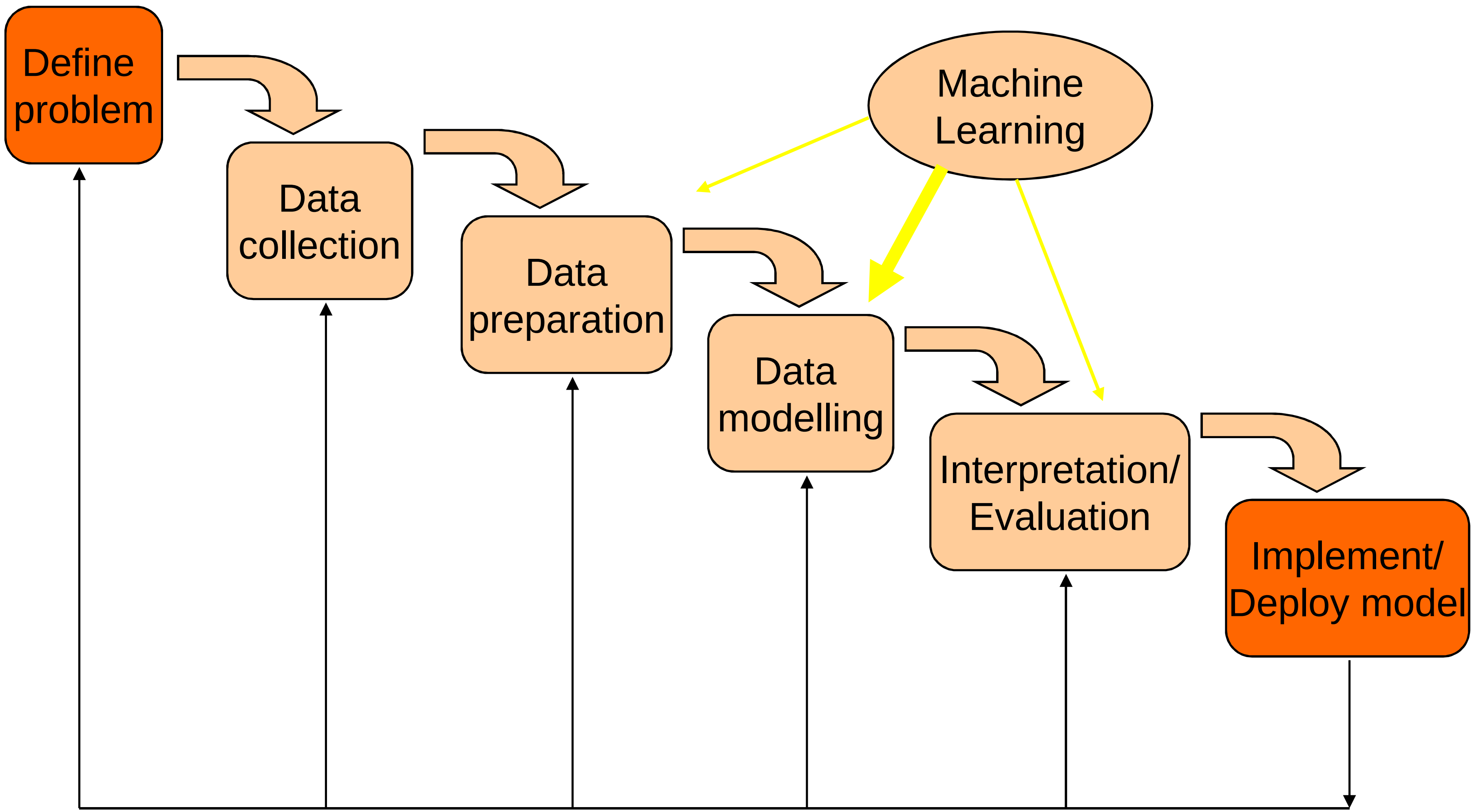
- There is a connection between machine learning and data mining
- Data mining:
  - "…the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner." (Hand et al., 2001)

Data Mining



Define problem
Data collection
Data preparation
Data modelling
Interpretation/ Evaluation
Implement/ Deploy model

Data Mining



Define problem
Data collection
Data preparation
Machine Learning
Data modelling
Interpretation/ Evaluation
Implement/ Deploy model

Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Optimization, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Quality of service optimization
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- ...

- About the data (see Hand et al., 2001, chap. 2)
  - "Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity. The relationships between objects are represented by numerical relationships between variables. These numerical relationships, that data items, are stored in the data set; it is these items that are the subjects of our data mining activities".

- Main important issues:
  - The measurement process
  - Distance metrics
  - Prior knowledge of the nature of the data (and the collection process)
  - Errors and corruption of data.
- Numerical data can be real-valued, integer, binary.
  - Ranking, ordering
- Nominal data (e.g. colours, categories)

- Data transformations

- Forms (schema) of data
  - Matrix/table
  - Multi-relational
  - Strings/trees/graphs
  - Hierarchical
  - And more!

What is Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

Applications

- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

Learning Associations

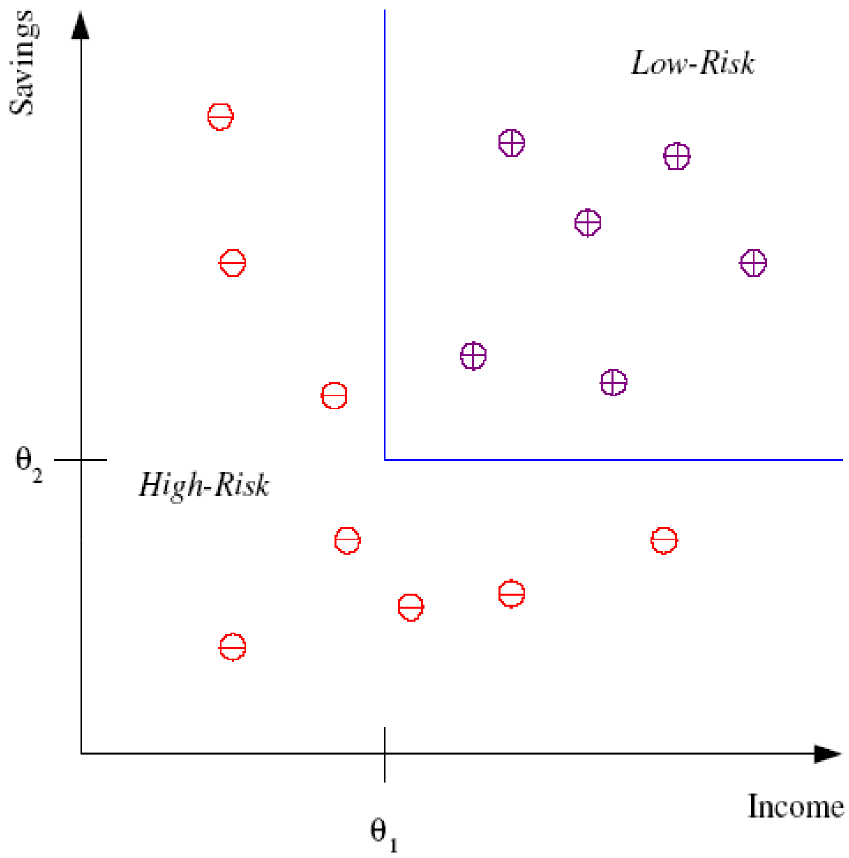
- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: P (chips | beer) = 0.7

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

Classification: Applications

- Aka Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- ...

Face Recognition

Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Regression

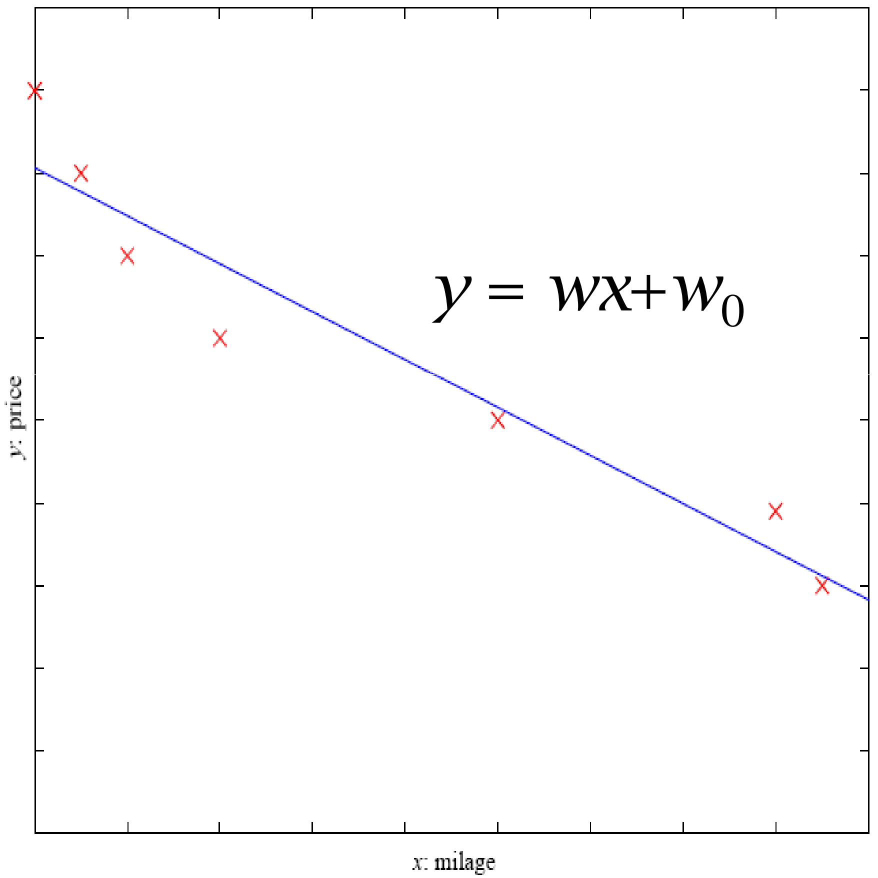
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g(x \mid \theta)$$

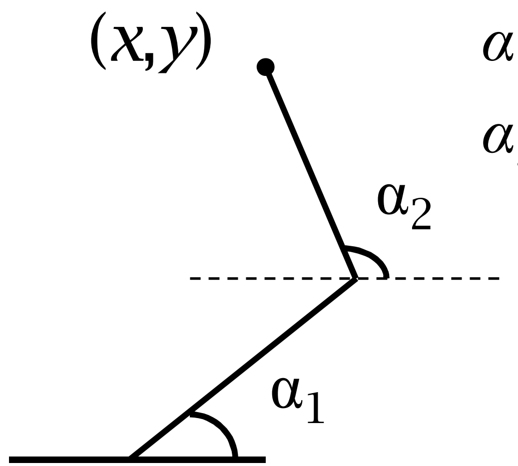
  $g(\ )$ model,

  θ parameters

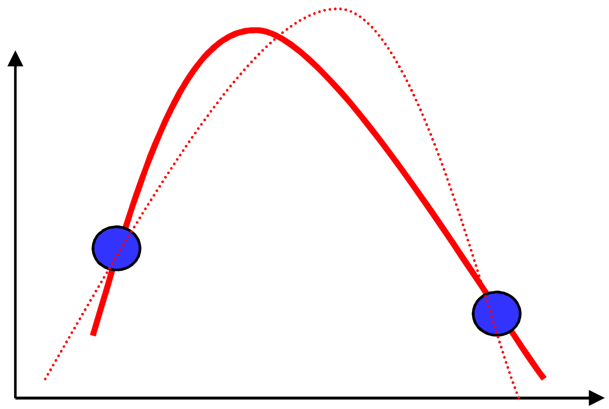


Regression Applications

- Navigating a car: Angle of the steering wheel (CMU NavLab)
- Kinematics of a robot arm



$$\alpha_1 = g_1(x,y)$$
$$\alpha_2 = g_2(x,y)$$



- Response surface design

Supervised Learning: Uses

- 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, e.g., fraud

Unsupervised Learning

- Learning "what normally happens"
- No output
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Unsupervised Learning

  - Given: a set (sample) of data (observations).
  - Task: build a model of the process that generated the data.
  - This time however, we’re not trying to learn about the relationship between inputs and outputs, we just want to find (and/or take advantage of) **structure** in the data.
  - Sometimes called descriptive (rather than predictive) modelling.
- Problems that can be framed as unsupervised learning: dimensionality reduction, compression, probability density estimation.

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Resources: Datasets

- 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/

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- 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
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
  - ICML05: http://icml.ais.fraunhofer.de/
- European Conference on Machine Learning (ECML)
  - ECML05: http://ecmlpkdd05.liacc.up.pt/
- Neural Information Processing Systems (NIPS)
  - NIPS05: http://nips.cc/
- Uncertainty in Artificial Intelligence (UAI)
  - UAI05: http://www.cs.toronto.edu/uai2005/
- Computational Learning Theory (COLT)
  - COLT05: http://learningtheory.org/colt2005/
- International Joint Conference on Artificial Intelligence (IJCAI)
  - IJCAI05: http://ijcai05.csd.abdn.ac.uk/
- International Conference on Neural Networks (Europe)
  - ICANN05: http://www.ibspan.waw.pl/ICANN-2005/
- ...