Introduction to Machine Learning

Davide Bacciu

Computational Intelligence & Machine Learning Group
Dipartimento di Informatica
Università di Pisa
{bacciu}@di.unipi.it

Introduzione all'Intelligenza Artificiale - A.A. 2012/2013

Lecture Outline

What..

...is learning?

- Process by which we acquire new or modify existing knowledge, skills, behaviors or preferences
- Several underlying memory mechanisms
  - Habituation
  - Associative learning
  - Observational Learning
  - ...

...is machine learning?

Definition (T. M. Mitchell, 1997)

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

Why?

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.
(Poggio and Shelton, AI Magazine, 1999)

- An AI methodology
  - Building intelligent/adaptive system
  - Learning allows agents to modify their decision mechanisms to improve their performance
- Statistical learning
  - Building systems for data analysis and prediction

Why? (II)

- A scientific methodology for innovative applications
  - Designing tools for complex problems
- Learning is useful as a system construction method
  - Some tasks cannot be defined well except by example
  - Certain characteristics of the working environment are not known at design time
  - Let the solution emerge from data rather than trying to write down the computational steps
  - Machines that can adapt to a changing environment reduce the need for constant redesign

When?

- Learning is essential for real-world problems
  - Lack of consolidated background knowledge/theory
  - Inefficient to use a mathematical model solving the specific problem
  - Noisy data
  - Too much information available
- It has few major requirements
  - Availability of data that represent the problem well
  - Admissibility of tolerance in the precision of results

Where?

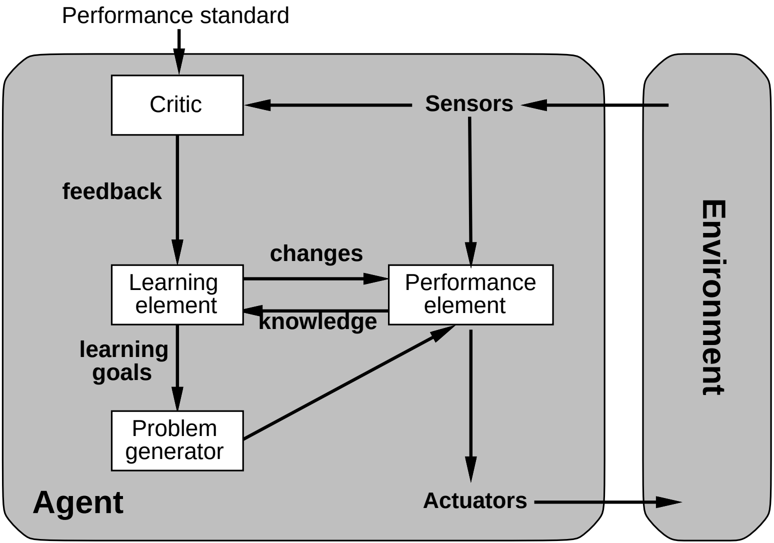
UNITED STATES
POSTAL SERVICE™

- Predicting behaviors or events
- Service personalization
- Recognition from noisy/complex data
- Analysis of large information collections

Who?

- Rigorous foundation in computational science
  - Artificial Intelligence
  - Statistics
  - Computational Intelligence
  - Numerical Analysis and Optimization
- Interdisciplinary applications
  - Pattern Recognition, Computer Vision, Language Processing, Information Retrieval
  - Robotics, Adaptive Systems and Filters
  - Data Mining, Financial forecasting, Analysis of complex data (Medicine, Biology, Chemistry, Web,. . .)
  - Personalized components
- Machine Learning developed with contributions from
  - Mathematical IA foundations
  - Physics and systems theory
  - Cognitive models, Neurobiology ($\Rightarrow$ Artificial Neural Networks)

Learning in Artificial Intelligence



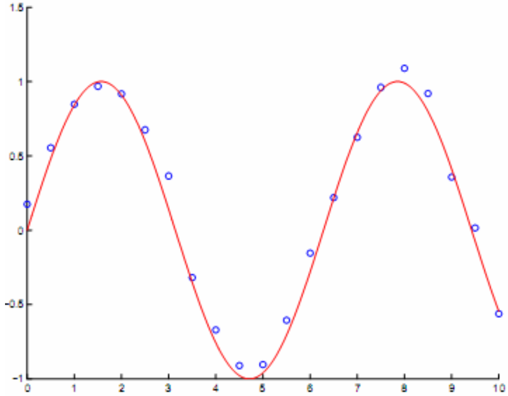
Learning from examples

- Acquisition (inference/induction) from data (examples) of the rules, models or representations which enable the production of a desired behavior
- The goal is not to memorize but to generalize the acquired knowledge
  - More than simply fitting the data
  - Estimating the value of function for unseen examples
- Given a set of N examples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_N, y_N)$$

find a function $f(\cdot)$ such that it is a good predictor of y for a future input x

Which one is the right f?



No right answer! You need to make assumptions

Key ingredients of Machine Learning

- Data
- Tasks
- Learning Machinery
  - Computational model - how knowledge is represented
    - Decision Trees
    - Neural Networks
    - Bayesian Models
  - Learning algorithm - how knowledge is adapted to the observations (examples)
    - Backpropagation
    - Expectation-Maximization
- Validation: measures of learning quality and performance

The learning problem (supervised learning setting)

- $\mathcal{X}$ - set of inputs
- $\mathcal{Y}$ - set of outputs (targets)
- $(x, y) \in \mathcal{X} \times \mathcal{Y}$ - an example or sample or observation

Definition (Training set)

A set of samples $\mathcal{D} = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)\}$ independently and identically drawn from $\mathcal{X} \times \mathcal{Y}$ with a given probability distribution

Definition (Hypothesis Space)

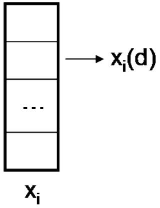
A space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathcal{Y}$

Definition (Learning Algorithm)

A map $L : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ that, using $\mathcal{D}$, selects from $\mathcal{H}$ a function h^* such that $h^*(x) \approx y$ in a predictive way

Sub-symbolic knowledge representation

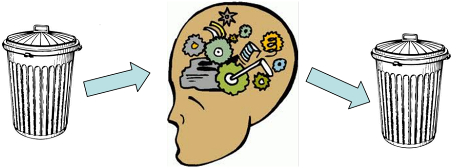
- The i-th input sample x_i is a D-dimensional numerical vector
  - Continuous, categorical or mixed values
  - Describes an individual of our world of interest, e.g. patients in a biomedical application
- The single dimensions d are called features and numerically represent an attribute of the individual
  - E.g. if x_i describes a patient, $x_i(d)$ can be his/her age
- Also output samples y_i are D'-dimensional numerical vectors



Machine learning deals with more than vectorial data, e.g. sequences, graphs, ... (Not in this course!)

Data quality

Garbage-in produces garbage-out, no matter how sophisticated your learning system is



- A machine learning model can only be as good as the data it sees
  - Learning quality increases with dataset size and quality
  - Sufficient coverage of the process that we are willing to model
  - Advanced issues: noise, missing data, balancing,. . .

Preprocessing

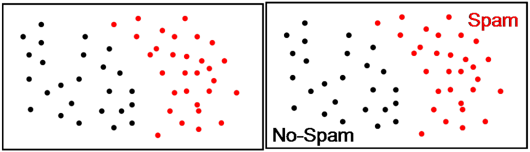
- Preliminary activity of data preparation and filtering
  - Errors correction
  - Missing data
  - Noise reduction
- Finding data representation maximizing the performance of the learning model
  - Scaling and normalization
  - Feature selection and extraction
- ML models themselves can be used to preprocess the data

Learning paradigms

- Algorithms can be differentiated based on the task they address
- Different tasks often require different degrees of feedback (teaching) information from the reality
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

Supervised Learning

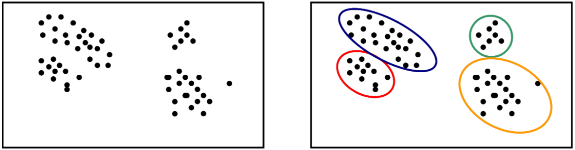
- Learns a function h mapping inputs to desired outputs
  - Classification: assign each input to a discrete class



  - Regression: output is a continuous vector
- Needs supervised information associating the input x_i to the desired target y_i
  - Training set is of the form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
  - Target y_i can be an integer in $\{1, \ldots, C\}$ (classification) or real (regression)
- Want to generalize well to a test-set of unseen data $\mathcal{D}'$

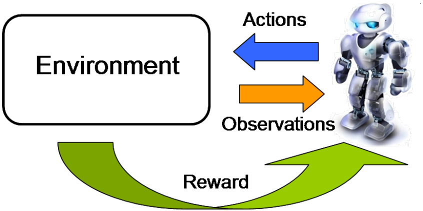
Unsupervised Learning

- Learns a natural grouping of the input data
  - Clustering



  - Finding a compressed representation for the data
  - Density estimation
- Only input pattern x_i is provided (no desired output)
  - Training set is of the form $\mathcal{D} = \{x_1, \ldots, x_N\}$
- The need for generalization remains

Reinforcement Learning



- Learning to chose the best action based on rewards or punishments from the interacting environment
  - Planning
  - Behavior learning
- Data comprises an input pattern x_i describing an observation of the environment and a reward $r_i \in \{-1, +1\}$ returned in response to the predicted action y_i
  - Training set is of the form $\mathcal{D} = \{(x_1, y_1, r_1), \ldots, (x_N, y_N, r_N)\}$
- Learn to choose actions y_i in such a way as to obtain a lot of reward

Inductive Learning Hypothesis

- We are interested in learning algorithms L that select an hypothesis h that generalizes well to unseen data
- What are the conditions ensuring generalization?

Definition (Inductive Learning Hypothesis)

Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples

We need a means for measuring how well an hypothesis approximates the target function

Empirical Error

Suppose we have a finite set $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ providing the target values y_i over N samples, we have

Definition (Empirical Error)

The empirical (sample) error of hypothesis h with respect to the sample $\mathcal{D}$ is

$$Err_{\mathcal{D}}(h) = \frac{1}{N} \sum_{(x_i, y_i) \in \mathcal{D}} L(h(x_i), y_i)$$

where $L(h(x_i), y_i)$ is the loss, i.e. a function measuring the discrepancy between the predicted $h(x_i)$ and the target value y_i

E.g. in classification $L(h(x_i), y_i) = 0$ if x_i is predicted to be in class y_i and is 1 otherwise

Expected Error

Define $z = (x, y)$ and given the joint distribution $\mu(z) = \mu(x, y)$, we have

Definition (Expected Error)

The expected (true) error of hypothesis h under distribution μ is

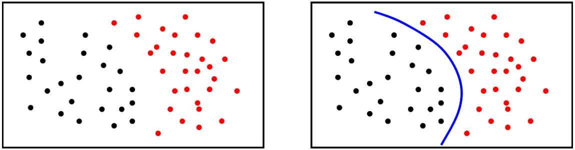
$$Err_{\mu}(h) = \int \mu(z)L(h(x), y)dz$$

By the Inductive Learning Hypothesis, we expect the empirical error to converge to the true error for a sufficiently large training set $\mathcal{D}$

$$\forall \mu \quad \lim_{N=|\mathcal{D}|\to\infty} P(|Err_{\mathcal{D}}(h) - Err_{\mu}(h)| > \epsilon) = 0$$

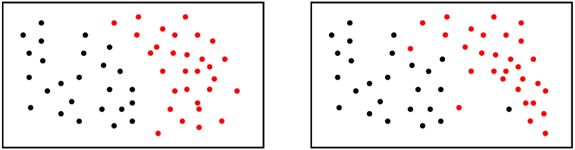
Choosing the Best Hypothesis

- Choosing an appropriate hypothesis space H can guarantee generalization
  - E.g. a compact set of continuous functions
- Given an appropriate H
  - We typically do not know the true error $Err_\mu(h)$
  - We use the empirical error $Err_{\mathcal{D}}(h)$ to find the hypothesis h that makes less errors on a large-enough sample $\mathcal{D}$
- Find a function of the point coordinates (hypothesis) having one output for red points and a different output for the black ones (classification). What is the right hypothesis?

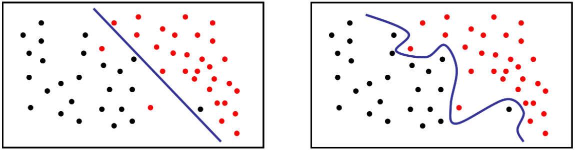


Hypothesis Complexity

What happens if we change the data slightly?



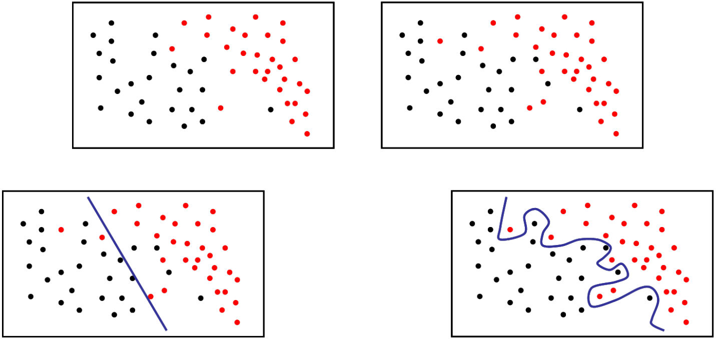
What is the best hypothesis now?



A line separates worse (more errors) but the spline is more complex since it has more parameters

What is the problem with complexity?

Lets add some more samples

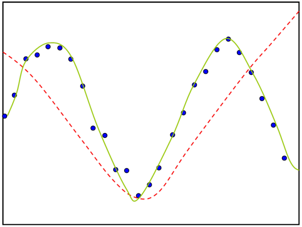


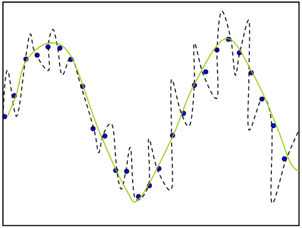
The line hypothesis does not need much adaptation to accommodate new data

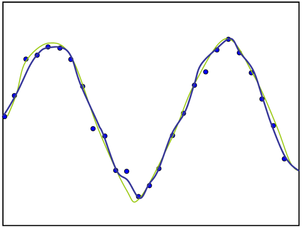
The spline changes radically

Bias-Variance Dilemma

Complexity and Generalization







Testing and Validation

How well does an hypothesis performs, in practice?

- Interest in how h^* will perform on new data

In general, measuring $L(h^*(x), y)$ on training data is not indicative of h^* performance on new data

- Maintain an external test set not used for training
- Reasonable estimate of performance on new data

Fundamental issues

- Separate training and model selection from testing (generalization assessment)
- Sophisticated statistical methods can be used to asses model performance in case of small data sets (bootstrapping, cross-validation)

Machine learning models - In brief

- Acquired knowledge is stored into the model parameters $W = \{w_1, \dots, w_P\}$
- Two operational modes
  - Learning phase (training, fitting)
    - Building the model from known data
    - Estimate the model parameters from the training data $\mathcal{D}_{train}$
  - Predictive phase (test)
    - Running the model with new samples $\mathcal{D}_{test}$
    - Feed new data $x \in \mathcal{D}_{test}$ in input to predict an output $out(x)$
- A loss function $L(\mathcal{D}, W)$ is used to estimate the quality of learned parameters W against data $\mathcal{D}$

The training phase - In brief

An iterative process that

1. Determines new values for the model parameters W' based on the training data $\mathcal{D}_{train}$
2. Evaluates the newly obtained model based on the loss $L(\mathcal{D}_{eval}, W')$ where $\mathcal{D}_{eval}$ is either
   - The training set $\mathcal{D}_{train}$
   - An external validation set $\mathcal{D}_{valid}$
3. If $L(\mathcal{D}_{eval}, W')$ is sufficiently small it stops, otherwise it iterates the two steps above

$$\mathcal{D}_{valid} \neq \mathcal{D}_{test}$$

Examples of loss functions

- Classification tasks

$$\text{accuracy} = \frac{\#\ \text{correctly predicted samples}}{\text{total number of samples}}$$

- Regression tasks

$$RMS = \sqrt{\sum_{i=1}^{N} (y_i - out(x_i))^2}$$

i.e. the Root Mean Squared error

Take Home Messages

- Learning is essential
  - For unknown or changing environments
  - To let the solution emerge from the data
- The key ingredients
  - Data - Garbage-in/Garbage-out
  - Tasks - Supervised, unsupervised and reinforcement learning
  - Learning machinery - How knowledge is represented and adapted to the data
  - Measures of learning performance
- Learning performance needs to measure prediction accuracy on unseen data
  - Generalization
  - Test set

Outline of the Module

1. Introduction to machine Learning
2. Inductive Learning (Simi)
3. Decision Trees (Simi)
4. Exploratory Analysis: Feature Extraction
5. Exploratory Analysis: Feature Selection
6. Exploratory Analysis: Clustering
7. Bayesian Learning
8. Reinforcement Learning
9. Machine Learning Applications
10. Advanced Machine Learning Models and the Computational Learning Theory (Micheli)

Course Information

Few course prerequisites

- Mathematical Analysis: functions, differential calculus
- Algorithms
- Matrix algebra
- Foundations of probability theory and statistics

Reference Webpage:

`http://www.di.unipi.it/~bacciu/IASpring13.html`

Here you can find

- Lecture slides
- Articles and course materials

Introductory readings to machine learning:

`http://www.di.unipi.it/~micheli/DID/`

Bibliography and Contacts

Bibliography

- Russell, S. and Norvig, N. *Artificial Intelligence: A Modern Approach*, 3rd Edition, Pearson Education, 2010
- Mitchell, T. *Machine Learning*, McGraw Hill.1997.

Contacts

Davide Bacciu - `bacciu@di.unipi.it`

Alessio Micheli - `micheli@di.unipi.it`

My Office Hours

Where? Dipartimento di Informatica - 2nd Floor - Room 301 (Neurolab)

When? Monday 14.30-16 / Friday 14.30-16

When? Basically anytime, if you send me an email beforehand.