Master Recherche IAC
TC2: Apprentissage Statistique & Optimisation

Alexandre Allauzen – Anne Auger – Michèle Sebag
LIMSI – LRI

Sept. 22nd, 2014
Where we are

Ast. series

Natural phenomenons

Data / Principles

Maths. Modelling

World

Pierre de Rosette

Human–related phenomenons

Common Sense

You are here
Where we are

Sc. data

Natural phenomena

Maths. Modelling

Data / Principles

World

Human–related phenomena

Common Sense

You are here
Harnessing Big Data

Watson (IBM) defeats human champions at the quiz game Jeopardy (Feb. 11)

- $i$, 1000
- $i$, kilo, mega, giga, tera, peta, exa, zetta, yotta, bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day; Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/seconds
Machine Learning and Optimization

Machine Learning

World $\rightarrow$ instance $\mathbf{x}_i$ $\rightarrow$ Oracle $\downarrow$ $y_i$

Optimization

Smart optimization requires learning about the optimization landscape

ML and Optimization

- ML is an optimization problem: find the best model
- Smart optimization requires learning about the optimization landscape
Types of Machine Learning problems

WORLD – DATA – USER

Observations + Target + Rewards

Understand Code Predict Classification/Regression Decide Action Policy/Strategy

Unsupervised Learning Supervised Learning Reinforcement Learning
2. Support Vector Machines
3. Learning from sequences
4. Unsupervised learning
5. Representation changes
6. Bayesian learning
7. Optimisation
Pointers

- Slides of this module:
  http://tao.lri.fr/tiki-index.php?page=Courses

- Andrew Ng courses
  http://ai.stanford.edu/~ang/courses.html

- PASCAL videos
  http://videolectures.net/pascal/

- Tutorials NIPS Neuro Information Processing Systems
  http://nips.cc/Conferences/2006/Media/

- About ML/DM
  http://hunch.net/
1. Part 1. Generalities
2. Part 2. Decision trees
3. Part 3. Validation
Examples

- Vision
- Control
- Netflix
- Spam
- Playing Go
- Google

http://ai.stanford.edu/~ang/courses.html
Reading cheques

LeCun et al. 1990
MNIST: The drosophila of ML

Fig. 4. Size-normalized examples from the MNIST database.
Detecting faces

Many Uses
- User Interfaces
- Interactive Agents
- Security Systems
- Video Compression
- Image Database Analysis

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001
A. Zisserman, C. Williams, M. Everingham, L. v.d. Gool
The supervised learning setting

Input: set of \((x, y)\)

- An instance \(x\)  
  e.g. set of pixels, \(x \in \mathbb{R}^D\)
- A label \(y\) in \(\{1, -1\}\) or \(\{1, \ldots, K\}\) or \(\mathbb{R}\)
The supervised learning setting

**Input**: set of \((x, y)\)

- An instance \(x\)
  - e.g. set of pixels, \(x \in \mathbb{R}^D\)
- A label \(y\) in \(\{1, -1\}\) or \(\{1, \ldots, K\}\) or \(\mathbb{R}\)

**Pattern recognition**

- **Classification**
  - Does the image contain the target concept?
  - \(h : \{\text{Images}\} \mapsto \{1, -1\}\)

- **Detection**
  - Does the pixel belong to the img of target concept?
  - \(h : \{\text{Pixels in an image}\} \mapsto \{1, -1\}\)

- **Segmentation**
  - Find contours of all instances of target concept in image
The 2005 Darpa Challenge

Thrun, Burgard and Fox 2005

Autonomous vehicle Stanley – Terrains
The Darpa challenge and the AI agenda

What remains to be done

- Reasoning 10%
- Dialogue 60%
- Perception 90%

Thrun 2005
Robots

Ng, Russell, Veloso, Abbeel, Peters, Schaal, ...

Reinforcement learning

Classification
(a) Factor graph modelling the variable interactions

(b) Behaviour of the 39-DOF Humanoid:
Reaching goal under Balance and Collision constraints

Bayesian Inference for Motion Control and Planning
Go as AI Challenge

Gelly Wang 07; Teytaud et al. 2008-2011

Reinforcement Learning, Monte-Carlo Tree Search
Claim
Many problems can be phrased as optimization in front of the uncertainty.
- Adversarial setting 2 two-player game
- Uniform setting a single player game

Management of energy stocks under uncertainty
States and Decisions

States

- Amount of stock (60 nuclear, 20 hydro.)
- Varying: price, weather alea or archive
- Decision: release water from one reservoir to another
- Assessment: meet the demand, otherwise buy energy
Netflix Challenge 2007-2008

Collaborative Filtering
Collaborative filtering

Input

- A set of users \( n_u, \text{ ca } 500,000 \)
- A set of movies \( n_m, \text{ ca } 18,000 \)
- A \( n_m \times n_u \) matrix: person, movie, rating
  Very sparse matrix: less than 1% filled...

Output

- Filling the matrix!
Collaborative filtering

Input
▶ A set of users \( n_u \), ca 500,000
▶ A set of movies \( n_m \), ca 18,000
▶ A \( n_m \times n_u \) matrix: person, movie, rating
  Very sparse matrix: less than 1% filled...

Output
▶ Filling the matrix!

Criterion
▶ (relative) mean square error
▶ ranking error
Spam – Phishing – Scam

Classification, Outlier detection

Best Buy Viagra Generic Online
Viagra 100mg x 100 Pills $125. Free Pills & Reorder Discount. We accept VISA & E-Check Payments. 90000+ Satisfied Customers!

Top Selling 100% Quality & Satisfaction guaranteed!
The power of big data

- Now-casting
- Public relations >> Advertizing

outbreak of flu
Mc Luhan and Google

We shape our tools and afterwards our tools shape us
Marshall McLuhan, 1964

First time ever a tool is observed to modify human cognition that fast.
Sparrow et al., Science 2011
Types of application

**Domain**

**Physical phenomenons**
- manufacturing, experimental sciences, numerical engineering
- Vision, speech, robotics...

**Social phenomenons**
- Health, Insurance, Banks ...

**Individual phenomenons**
- Consumer Relationship Management, User Modelling
- Social networks, games...

**But : Modelling**

**analysis & control**

**+ privacy**

**+ dynamics**

PASCAL : http://pascallin2.ecs.soton.ac.uk/
Ex: KDD 2009 – Orange

1. Churn
2. Appetency
3. Up-selling

Objectives

1. Ads. efficiency
2. Less fraud
Ex: Risk factors

1. Cardio-vascular diseases
2. Carcinogenic Molecules
3. Obesity genes ...

Objectives

1. Diagnostic
2. Personalized care
3. Identification
Questions

1. Who does what?
2. Good conferences?
3. Hot/emerging topics?
4. Is Mr Q. Lee same as Mr Quoc N. Lee?
Numerical Engineering

- Codes
- Computationally heavy
- Expertise demanding

Fusion based on inertial confinement, ICF
Objectives

- Approximate answer
- .. in tenth of seconds
- Speed up the design cycle
- Optimal design

More is Different
Autonomous robotics

Complexes, monde fermé
Simple, random

Design

[tr. Hod Lipson, 2010]
Autonomous robotics, 2

Reality Gap

- Design in silico (simulator)
- Run the controller on the robot (in vivo)

Active learning Co-evolution

[tr. Hod Lipson, 2010]
Autonomous robotics, 2

Reality Gap
- Design in silico (simulator)
- Run the controller on the robot (in vivo)
- Does not work!

Closing the reality Gap
1. Simulator-based design
2. On-board trials safe environnement
3. Log the data, update the simulator
4. Goto 1

Active learning Co-evolution
[tr. Hod Lipson, 2010]
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation
  Performance indicators
  Estimating an indicator
Types of Machine Learning problems

WORLD – DATA – USER

Observations + Target + Rewards

Understand Code Predict Classification/Regression Decide Policy

Unsupervised LEARNING Supervised LEARNING Reinforcement LEARNING
Data

Example

- row : example/ case
- column : feature/ variable/ attribute
- attribute : class/ label

Instance space $\mathcal{X}$

- Propositionnal : $\mathcal{X} \equiv \mathbb{R}^d$
- Structured : sequential, spatio-temporal, relational.

<table>
<thead>
<tr>
<th>age</th>
<th>employ/degree</th>
<th>education</th>
<th>edur</th>
<th>marital</th>
<th>...</th>
<th>job</th>
<th>relation</th>
<th>race</th>
<th>gender</th>
<th>hour</th>
<th>country</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>HS</td>
<td>Single</td>
<td>...</td>
<td>Adm_cler</td>
<td>Not_in_fam</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>51</td>
<td>Self_emp.</td>
<td>Bachelors</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>43</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>59</td>
<td>Private</td>
<td>HS_grad</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Handlers</td>
<td>Not_in_fam</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>54</td>
<td>Private</td>
<td>11th</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>58</td>
<td>Private</td>
<td>Bachelors</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Wife</td>
<td>Black</td>
<td>Female</td>
<td>40</td>
<td>Cuba</td>
<td>poor</td>
</tr>
<tr>
<td>48</td>
<td>Private</td>
<td>Masters</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Wife</td>
<td>White</td>
<td>Female</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>50</td>
<td>Private</td>
<td>9th</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Other_serv</td>
<td>Not_in_fam</td>
<td>White</td>
<td>Female</td>
<td>40</td>
<td>Jamaica</td>
<td>poor</td>
</tr>
<tr>
<td>52</td>
<td>Self_emp.</td>
<td>HS_grad</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>45</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>31</td>
<td>Private</td>
<td>Masters</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Not_in_fam</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>53</td>
<td>Private</td>
<td>Bachelors</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>37</td>
<td>Private</td>
<td>Some_col</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>80</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>30</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>India</td>
<td>rich</td>
</tr>
<tr>
<td>24</td>
<td>Private</td>
<td>Bachelors</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Adm_cler</td>
<td>Own_child</td>
<td>White</td>
<td>Female</td>
<td>30</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>Assoc_acc</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Sales</td>
<td>Not_in_fam</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Assoc_acc</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Craft_repair</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>Missing</td>
<td>rich</td>
</tr>
<tr>
<td>34</td>
<td>Private</td>
<td>7th_8th</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Transport_Husband</td>
<td>Amer_Ind</td>
<td>Male</td>
<td>45</td>
<td>Mexico</td>
<td>poor</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Self_emp.</td>
<td>HS_grad</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Farming</td>
<td>Own_child</td>
<td>White</td>
<td>Male</td>
<td>35</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>HS_grad</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Machine</td>
<td>Unmarried</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>38</td>
<td>Private</td>
<td>11th</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Sales</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>44</td>
<td>Self_emp.</td>
<td>Masters</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Exec_man</td>
<td>Unmarried</td>
<td>White</td>
<td>Female</td>
<td>50</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Doctorate</td>
<td>HS</td>
<td>Married</td>
<td>...</td>
<td>Prof_spec</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>60</td>
<td>United_States</td>
<td>rich</td>
</tr>
</tbody>
</table>
- Propositionnal data
- Spatio-temporal data
- Relationnal data
- Semi-structured data
- Multi-media

Data / Applications

- 80% des applis.
- alarms, mines, accidents
- chemistry, biology
- text, Web
- images, music, movies,..
Difficulty factors

Quality of data / of representation
- Noise; missing data
+ Relevant attributes
- Structured data: spatio-temporal, relational, text, videos,..

Data distribution
+ Independants, identically distributed examples
- Other: robotics; data streams; heterogeneous data

Prior knowledge
+ Goals, interestingness criteria
+ Constraints on target hypotheses
**Learning criterion**

- Convex optimization problem
- Complexity: \( n, n \log n, \ n^2 \)
- Combinatorial optimization

**Scalability**

H. Simon, 1958:

*In complex real-world situations, optimization becomes approximate optimization since the description of the real-world is radically simplified until reduced to a degree of complication that the decision maker can handle.*

*Satisficing seeks simplification in a somewhat different direction, retaining more of the detail of the real-world situation, but settling for a satisfactory, rather than approximate-best, decision.*
The user’s criteria

- Relevance, causality,
- INTELLIGIBILITY
- Simplicity
- Stability
- Interactive processing, visualisation
- ... Preference learning
Difficulty factors, 3

Crossing the chasm

▶ No *killer algorithm*
▶ Little expertise about algorithm selection

How to assess an algorithm

▶ Consistency

When number $n$ of examples goes to infinity and target concept $h^*$ is in $\mathcal{H}$

$h^*$ is found:

$$\lim_{n \to \infty} h_n = h^*$$

▶ Speed of convergence

$$\|h^* - h_n\| = \mathcal{O}(1/n), \mathcal{O}(1/\sqrt{n}), \mathcal{O}(1/\ln n)$$
Context

Disciplines et critères

- Data bases, Data Mining
- Statistics, data analysis
- Machine learning
- Optimisation
- Computer Human Interaction
- High performance computing

- Scalability
- Predefined models
- Prior knowledge; complex data/hypotheses
- Well / ill posed problems
- No final solution: a process
- Distributed processing; safety
Supervised Machine Learning

**Context**

World $\rightarrow$ instance $x_i$ $\rightarrow$ Oracle $\downarrow$ $y_i$

**Input**

Training set $E = \{(x_i, y_i), \ i = 1 \ldots n, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

**Milestones**

- Select hypothesis space $\mathcal{H}$
- Assess hypothesis $h \in \mathcal{H}$ $\quad \text{score}(h)$
- Find best hypothesis $h^*$
iid: Independent identically distributed.

**Independent**

\[(x_i, y_i)\] does not depend on \[(x_j, y_j)\]

Counter-example:
- \(x_i\) is the vector of sensor values of the robot at time \(i\)

**Identically distributed**

\(x_i\) are drawn after the same distribution

Counter-example:
- \(x_i\) is the length travelled for fixed actuator values; the distribution changes as the robot goes on different types of ground.
What is the goal?

Underfitting

Overfitting

The goal is not to be perfect on the training set
What is the goal?

- Underfitting
- Overfitting

The goal is not to be perfect on the training set.

**The villain: overfitting**

![Graph](image)

- Test error
- Training error

Complexity of Hypotheses
What is the goal?

**Prediction good** on future instances

**Necessary condition:**
Future instances must be similar to training instances
“identically distributed”

**Minimize (cost of) errors**
not all mistakes are equal.

\[ \ell(y, h(x)) \geq 0 \]
Minimize expectation of error cost

Minimize $E[\ell(y, h(x))] = \int_{X \times Y} \ell(y, h(x)) p(x, y) dx dy$

Generalization error
Minimize expectation of error cost

Minimize \( E[\ell(y, h(x))] = \int_{X \times Y} \ell(y, h(x))p(x, y)\,dx\,dy \)

Define Empirical Error

\[ Err_e(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) \]

Principle

Si function \( F \) “is well-behaved“ on space \( \mathcal{X} \) and \( \mathcal{E} \) is a ”sufficient” sample of \( \mathcal{X} \), then integral of \( F \) on \( \mathcal{X} \) is close to its empirical average on \( \mathcal{E} \).

\[ E[F] \leq \frac{\sum_{i=1}^{n} F(x_i)}{n} + c(F, n) \]
Classification, criteria

Generalisation error

\[ Err(h) = E[\ell(y, h(x))] = \int \ell(y, h(x)) dP(x, y) \]

Empirical error

\[ Err_e(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) \]

Bound risk minimization

\[ Err(h) < Err_e(h) + \mathcal{F}(n, d(\mathcal{H})) \]

\[ d(\mathcal{H}) = \text{VC-dimension of } \mathcal{H} \]
**Bias**

Bias ($\mathcal{H}$): error of the best hypothesis $h^*$ in $\mathcal{H}$

**Variance**

Variance of $h_n$ depending on $\mathcal{E}$

---

**The Bias-Variance trade-off**

As hypothesis space increases, bias decreases; but variance increases.
Classification: Ingredients of error

**Bias**
Bias ($\mathcal{H}$): error of the best hypothesis $h^*$ in $\mathcal{H}$

**Variance**
Variance of $h_n$ depending on $\mathcal{E}$

Optimization
negligible in small scale
takes over in large scale

(Google)
Classification, Problem posed

INPUT

\[ \mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \{0, 1\}, i = 1 \ldots n\} \]

HYPOTHESIS SPACE

\[ \mathcal{H} : \mathcal{X} \mapsto \{0, 1\} \]

SEARCH SPACE

LOSS FUNCTION

\[ \ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R} \]

OUTPUT

\[ h^* = \arg \max \{\text{score}(h), h \in \mathcal{H}\} \]
Key notions

- The main issue regarding supervised learning is overfitting.

- How to tackle overfitting:
  - Before learning: use a sound criterion regularization
  - After learning: cross-validation

Summary

- Learning is a search problem
- What is the space? What are the navigation operators?
Hypothesis Spaces

Logical Spaces

Concept ← $\bigvee \bigwedge$ Literal,Condition

- Conditions = [color = blue]; [age < 18]
- Condition $f : X \mapsto \{ True, False \}$
- Find: disjunction of conjunctions of conditions

- Ex: (unions of) rectangles of the 2D-plane $X$. 
Hypothesis Spaces

Numerical Spaces

Concept = \((h()) > 0\)

- \(h(x) = \) polynomial, neural network, ...
- \(h : X \mapsto \mathbb{R}\)
- Find: (structure and) parameters of \(h\)
Hypothesis Space $\mathcal{H}$

**Logical Space**

- $h$ covers one example $x$ iff $h(x) = True$.
- $\mathcal{H}$ is structured by a partial order relation

\[
h \prec h' \text{ iff } \forall x, h(x) \rightarrow h'(x)
\]

**Numerical Space $\mathcal{H}$**

- $h(x)$ is a real value (more or less far from 0)
- we can define $\ell(h(x), y)$
- $\mathcal{H}$ is structured by a partial order relation

\[
h \prec h' \text{ iff } E[\ell(h(x), y)] < E[\ell(h'(x), y)]
\]
## Hypothesis Space $\mathcal{H}$ / Navigation

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathcal{H}$ navigation operators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version Space</td>
<td>Logical spec / gen</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>Logical specialisation</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>Numerical gradient</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>Numerical quadratic opt.</td>
</tr>
<tr>
<td>Ensemble Methods</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>adaptation $\mathcal{E}$</td>
</tr>
</tbody>
</table>
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation
  Performance indicators
  Estimating an indicator
C4.5 (Quinlan 86)

- Among the most widely used algorithms
- Easy
  - to understand
  - to implement
  - to use
  - and cheap in CPU time
- J48, Weka, SciKit
Decision Trees

$$(x_i, y_i)$$

Diagram showing a decision tree with nodes labeled S and L, with color-coded bars.
Procedure DecisionTree(\(\mathcal{E}\))

1. Assume \(\mathcal{E} = \{(x_i, y_i)_{i=1}^n, x_i \in \mathbb{R}^D, y_i \in \{0, 1\}\}
   
   - If \(\mathcal{E}\) single-class (i.e., \(\forall i, j \in [1, n]; y_i = y_j\)), return
   - If \(n\) too small (i.e., \(<\) threshold), return
   - Else, find the most informative attribute \(att\)

2. For all value \(val\) of \(att\)
   
   - Set \(\mathcal{E}_{val} = \mathcal{E} \cap [att = val]\).
   - Call DecisionTree(\(\mathcal{E}_{val}\))

Criterion: information gain

\[
p = Pr(Class = 1|att = val)
\]

\[
I([att = val]) = -p \log p - (1 - p) \log (1 - p)
\]

\[
I(att) = \sum_i Pr(att = val_i).I([att = val_i])
\]
Contingency Table

<table>
<thead>
<tr>
<th>agegroup</th>
<th>10s</th>
<th>20s</th>
<th>30s</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
<th>70s</th>
<th>80s</th>
<th>90s</th>
</tr>
</thead>
<tbody>
<tr>
<td>wealth values: poor</td>
<td>2507</td>
<td>11262</td>
<td>9468</td>
<td>6738</td>
<td>4110</td>
<td>2245</td>
<td>668</td>
<td>115</td>
<td>42</td>
</tr>
<tr>
<td>rich</td>
<td>3</td>
<td>743</td>
<td>3461</td>
<td>3986</td>
<td>2509</td>
<td>809</td>
<td>147</td>
<td>16</td>
<td>13</td>
</tr>
</tbody>
</table>

Quantity of Information (QI)

<table>
<thead>
<tr>
<th>p</th>
<th>0.10</th>
<th>0.30</th>
<th>0.50</th>
<th>0.70</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>QI</td>
<td>0.00924</td>
<td>0.00474</td>
<td>0.232</td>
<td>0.0570323</td>
<td>0.153715</td>
</tr>
</tbody>
</table>

Computation

<table>
<thead>
<tr>
<th>value</th>
<th>p(value)</th>
<th>p poor</th>
<th>QI (value)</th>
<th>p(value) * QI (value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,10]</td>
<td>0.051</td>
<td>0.999</td>
<td>0.00924</td>
<td>0.000474</td>
</tr>
<tr>
<td>[10,20]</td>
<td>0.25</td>
<td>0.938</td>
<td>0.232</td>
<td>0.0570323</td>
</tr>
<tr>
<td>[20,30]</td>
<td>0.26</td>
<td>0.732</td>
<td>0.581</td>
<td>0.153715</td>
</tr>
</tbody>
</table>
Limitations

- XOR-like attributes
- Attributes with many values
- Numerical attributes
- Overfitting
Limitations

Numerical Attributes

- Order the values $val_1 < \ldots < val_t$
- Compute $QI([att < val_i])$
- $QI(att) = \max_i QI([att < val_i])$

The XOR case
Bias the distribution of the examples
Complexity

Quantity of information of an attribute

\[ n \ln n \]

Adding a node

\[ D \times n \ln n \]
Tackling Overfitting

Penalize the selection of an already used variable

▶ Limits the tree depth.

Do not split subsets below a given minimal size

▶ Limits the tree depth.

Pruning

▶ Each leaf, one conjunction;
▶ Generalization by pruning literals;
▶ Greedy optimization, QI criterion.
Decision Trees, Summary

Still around after all these years

- Robust against noise and irrelevant attributes
- Good results, both in quality and complexity

Random Forests

Breiman 00
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation
  Performance indicators
  Estimating an indicator
Validation issues

1. What is the result?

2. My results look good. Are they?

3. Does my system outperform yours?

4. How to set up my system?
Validation: Three questions

Define a good indicator of quality
- Misclassification cost
- Area under the ROC curve

Computing an estimate thereof
- Validation set
- Cross-Validation
- Leave one out
- Bootstrap

Compare estimates: Tests and confidence levels
Settings

- Large/few data

Data distribution

- Dependent/independent examples
- balanced/imbalanced classes
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation
  Performance indicators
  Estimating an indicator
Performance indicators

Binary class
- $h^*$ the truth
- $\hat{h}$ the learned hypothesis

Confusion matrix

<table>
<thead>
<tr>
<th>$\hat{h} / h^*$</th>
<th>1</th>
<th>0</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
<td></td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance indicators, 2

\[ \hat{h} / h^* \]

<table>
<thead>
<tr>
<th>( \hat{h} / h^* )</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

- Misclassification rate \( \frac{b+c}{a+b+c+d} \)
- Sensitivity (recall), True positive rate (TP) \( \frac{a}{a+c} \)
- Specificity, False negative rate (FN) \( \frac{b}{b+d} \)
- Precision \( \frac{a}{a+b} \)

**Note:** always compare to random guessing / baseline alg.
The Area under the ROC curve

- ROC: Receiver Operating Characteristics
- Origin: Signal Processing, Medicine

Principle

\[ h : X \mapsto \mathbb{R} \quad h(x) \text{ measures the risk of patient } x \]

\( h \) leads to order the examples:

\[
\begin{aligned}
+ & + + - + - + + + + - - - + - - - + - - - - - - - - - - - - \\
\end{aligned}
\]
The Area under the ROC curve

- ROC: Receiver Operating Characteristics
- Origin: Signal Processing, Medicine

Principle

\[ h : X \mapsto \mathbb{R} \quad h(x) \text{ measures the risk of patient } x \]

\[ h \text{ leads to order the examples:} \]

\[ + + + - - + + + + - - - + - - - + - - - - - - - - - - - - - - - \]

Given a threshold \( \theta \), \( h \) yields a classifier: Yes iff \( h(x) > \theta \).

\[ + + + - + - + + + + \quad | \quad - - - + - - - + - - - - - - - - - - - - - - - \]

Here, \( TP(\theta) = 0.8 \); \( FN(\theta) = 0.1 \)
ROC
Ideal classifier: (0 False negative, 1 True positive)
Diagonal (True Positive = False negative) ≡ nothing learned.
Properties

ROC depicts the trade-off True Positive / False Negative.

Standard: misclassification cost (Domingos, KDD 99)

\[
\text{Error} = \# \text{ false positive} + c \times \# \text{ false negative}
\]

In a multi-objective perspective, ROC = Pareto front.

Best solution: intersection of Pareto front with \( \Delta(-c, -1) \)
ROC Curve, Properties, foll’d

Bradley 97

Used to compare learners
multi-objective-like
insensitive to imbalanced distributions
shows sensitivity to error cost.
Area Under the ROC Curve

Often used to select a learner
Don’t ever do this!  
Hand, 09

Sometimes used as learning criterion
Mann Whitney, Wilcoxon

\[ AUC = \Pr(h(x) > h(x')|y > y') \]

WHY
- More stable \( \mathcal{O}(n^2) \) vs \( \mathcal{O}(n) \)
- With a probabilistic interpretation  
  Clemençon et al. 08

HOW
- SVM-Ranking  
  Joachims 05; Usunier et al. 08, 09
- Stochastic optimization
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

Performance indicators

Estimating an indicator
Validation, principle

Desired: performance on further instances

Assumption: Dataset is to World, like Training set is to Dataset.
Validation, 2

Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Validation, 2

Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation
  Performance indicators
  Estimating an indicator
Confidence intervals

Definition
Given a random variable $X$ on $\mathbb{R}$, a p\%-confidence interval is $I \subset \mathbb{R}$ such that
$$Pr(X \in I) > p$$

Binary variable with probability $\epsilon$
Probability of $r$ events out of $n$ trials:
$$P_n(r) = \frac{n!}{r!(n-r)!} \epsilon^r (1 - \epsilon)^{n-r}$$

- Mean: $n\epsilon$
- Variance: $\sigma^2 = n\epsilon(1 - \epsilon)$

Gaussian approximation
$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{x-\mu}{\sigma}^2\right)$$
Bounds on (true value, empirical value) for $n$ trials, $n > 30$

$$Pr(|\hat{x}_n - x^*| > 1.96 \sqrt{\frac{\hat{x}_n(1-\hat{x}_n)}{n}}) < .05$$

Table

<table>
<thead>
<tr>
<th>$z$</th>
<th>.67</th>
<th>1.</th>
<th>1.28</th>
<th>1.64</th>
<th>1.96</th>
<th>2.33</th>
<th>2.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$</td>
<td>50</td>
<td>32</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Empirical estimates

When data abound

(MNIST)

Training

Test

Validation

Cross validation

N-fold Cross Validation

Error = Average (error on N-fold Cross Validation of h learned from )
Empirical estimates, foll’d

Cross validation → Leave one out

Same as N-fold CV, with $N = \text{number of examples}$.

**Properties**

- Low bias; high variance; underestimate error if data not independent
Empirical estimates, foll’d

Dataset

Bootstrap

uniform sampling with replacement

Training set

Test set.

rest of examples

Average indicator over all (Training set, Test set) samplings.
Beware

Multiple hypothesis testing

- If you test many hypotheses on the same dataset
- one of them will appear confidently true...

More

- Video and slides (soon): ICML 2012, Videolectures, Tutorial Japkowicz & Shah
  http://www.mohakshah.com/tutorials/icml2012/
Validation, summary

What is the performance criterion
- Cost function
- Account for class imbalance
- Account for data correlations

Assessing a result
- Compute confidence intervals
- Consider baselines
- Use a validation set

If the result looks too good, don’t believe it