LUND UNIVERSITY

Introduction to Machine Learning

Jacek Malec
Department of Computer Science
Lund University
email: jacek.malec@cs.lth.se

5th March 2009

AI@CS

Department of Computing Science

LUND UNIVERSITY

Introduction to Machine Learning

Topics for today:

1. what is machine learning?
2. what is hypothesis space?
3. what is inductive bias?
4. what are decision trees?
5. how can we learn them?

AI@CS

Department of Computing Science

LUND UNIVERSITY

Learning

"Changes to the content and organization of an agent's knowledge enabling it to improve its performance on a particular task or population of tasks." [Simon]

- Improve over task T,
- with respect to performance measure P,
- based on experience E.

AI@CS

Department of Computing Science

LUND UNIVERSITY

Why Machine Learning

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Interested industry

AI@CS

Department of Computing Science

Three niches for machine learning

- Data mining: using historical data to improve decisions
  - medical records → medical knowledge
- Software applications we can't program by hand
  - autonomous driving
  - speech recognition
- Self-customizing programs
  - Newsreader that learns user interests

Typical Data Mining Task

Data:

- Patient103, time=1, Age: 23 , FirstPregnancy: no , Anemia: no, Diabetes: no, PreviousPrematureBirth: no, Ultrasound: ?, Elective C-Section: ?, Emergency C-Section: ?
- Patient103, time=2, Age: 23, FirstPregnancy: no, Anemia: no, PreviousPrematureBirth: no, Diabetes: yes, Ultrasound: abnormal, Elective C-Section: no, Emergency C-Section: ?
- Patient103, time=5, Age: 23, FirstPregnancy: no, Anemia: no, PreviousPrematureBirth: no, Diabetes: yes, Ultrasound: abnormal, Elective C-Section: no, Emergency C-Section: **Yes**

Typical Data Mining Task 2

Given:

- 9714 patient records, each describing a pregnancy and birth
- Each patient record contains 215 features

Learn to predict:

- Classes of future patients at high risk for Emergency Cesarean Section

Data Mining Result

```
If No previous vaginal delivery, and
   Abnormal 2nd Trimester Ultrasound,
   and Malpresentation at admission
Then Prob. of Emergency C-Section
                              is 0.6

 Over training data: 26/41 = .63,
 Over test data: 12/20 = .60
```

Types of learning

- Inferential basis
  - inductive learning and the acquisition of new knowledge
  - deductive learning and the organization of existing knowledge
- Pedagogical basis
  - supervised learning
  - unsupervised learning

Perceptron, McCulloch-Pitts

$$f_1(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

$$f_2(x) = wx$$

$$f_3(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} w_i x_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Other threshold functions

Gaussian:

$$f(x_1, x_2, \ldots, x_n) = e^{-\frac{1}{2n}\sum_{i=1}^{n}(x_i - c_i)^2}$$

Sigmoid:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Learning as search

$F(x_1, x_2, \ldots, x_n, w_1, w_2, \ldots, w_n)$ continuous and differentiable

$$P(w_1, ..., w_n) = -\frac{1}{|E|} \sum_{(x,y) \in E} (F(x, w_1, ..., w_n) - y)^2$$

P - performance function

Search: how is it done?

$$F(x, w_1, w_2, \ldots, w_n) = \sum_{i=1}^{n} x_i w_i$$

Gradient:

$$\nabla P(w_1, ..., w_n) = \left[\frac{\partial P}{\partial w_1}, \cdots, \frac{\partial P}{\partial w_n}\right]$$

where

$$\frac{\partial P}{\partial w_i} = \frac{2}{|E|} \sum_{(x,y)\in E} (F(x, w_1, ..., w_n) - y)x_i$$

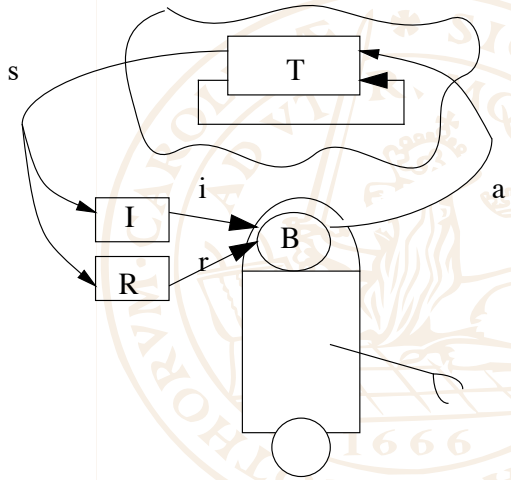
Search: how is it done? (2)

We can now use gradient search to adjust the weights:

$$w_i \leftarrow w_i + \beta \frac{\partial P}{\partial w_i}$$

where β is fixed learning rate

Reinforcement Learning



Reinforcement Learning - notation

R — agent
E — environment
s — state of the environment ($s \in S$)
a — action
I — observation function ($I : S \rightarrow S'$)
S' — observable states
i — information about the current state
R — reinforcement function ($R : S \rightarrow < 0, 1 >$)
r — current reinforcement value

LUND UNIVERSITY

PROBLEM

Learn to behave optimally in each possible situation.

AI@CS

Department of Computing Science

LUND UNIVERSITY

Optimally

- maximizing chances of achieving a goal;
- maximizing chances of survival;
- maximizing some more elaborated criteria
  - robots: energy saving;
  - robots: minimizing time of goal achievement;
  - medical expert system: maximizing survival rate of patients;
  - car estimation damage expert system: minimizing repair costs;
  - etc.

AI@CS

Department of Computing Science

LUND UNIVERSITY

LEARNING

- SUPERVISED
  the agent is given explicit information about input-output pairs (e.g. positive and negative examples)
- UNSUPERVISED
  the learning is based solely on trial and error procedure, the agent learns as it acts

Reinforcement learning is UNSUPERVISED!
Some claim that unsupervised should be divided into

- self-supervised
- (really) unsupervised

Reinforcement is a self-supervised learning method in this case!

AI@CS

Department of Computing Science

LUND UNIVERSITY

Maximize expected reward

- finite horizon model (h-step optimality)

$$E(\sum_{i=0}^{h} r_i)$$

- average reward model

$$\lim_{h\to\infty} E(\frac{1}{h}\sum_{i=0}^{h} r_i)$$

- infinite horizon discounted model

$$E(\sum_{i=0}^{\infty} \gamma^i r_i), 0 < \gamma < 1$$

AI@CS

Department of Computing Science

Measuring learning performance

- eventual convergence to optimality
- speed of convergence
- regret

Unsupervised Learning

clustering

Self-Organising Feature Maps (Kohonen)

Radial-basis neural networks

Questions

- What is a hypothesis space?
- What is inductive bias?
- Decision tree representation
- ID3 learning algorithm
- Entropy, Information gain
- Overfitting

LATER

Now it is time for Russell and Norvig!

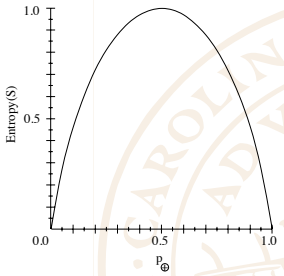
Decision Tree Representation

Learned from medical records of 1000 women

Negative examples are C-sections

```
[833+,167-] .83+ .17-
Fetal_Presentation = 1: [822+,116-] .88+ .12-
| Previous_Csection = 0: [767+,81-] .90+ .10-
| | Primiparous = 0: [399+,13-] .97+ .03-
| | Primiparous = 1: [368+,68-] .84+ .16-
| | | Fetal_Distress = 0: [334+,47-] .88+ .12-
| | | | Birth_Weight < 3349: [201+,10.6-] .95+ .05-
| | | | Birth_Weight >= 3349: [133+,36.4-] .78+ .22-
| | | Fetal_Distress = 1: [34+,21-] .62+ .38-
| Previous_Csection = 1: [55+,35-] .61+ .39-
Fetal_Presentation = 2: [3+,29-] .11+ .89-
Fetal_Presentation = 3: [8+,22-] .27+ .73-
```


Entropy



- S is a sample of training examples
- $p_\oplus$ is the proportion of positive examples in S
- $p_\ominus$ is the proportion of negative examples in S
- Entropy measures the impurity of S

$$Entropy(S) \equiv -p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus$$

Entropy 2

$Entropy(S)$ = expected number of bits needed to encode class ($\oplus$ or $\ominus$) of randomly drawn member of S (under the optimal, shortest-length code)

Information theory: optimal length code assigns $-\log_2 p$ bits to message having probability p.

So, expected number of bits to encode $\oplus$ or $\ominus$ of random member of S:

$$p_\oplus(-\log_2 p_\oplus) + p_\ominus(-\log_2 p_\ominus)$$

$$Entropy(S) \equiv -p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus$$

Information Gain

$Gain(S, A)$ = expected reduction in entropy due to sorting on A

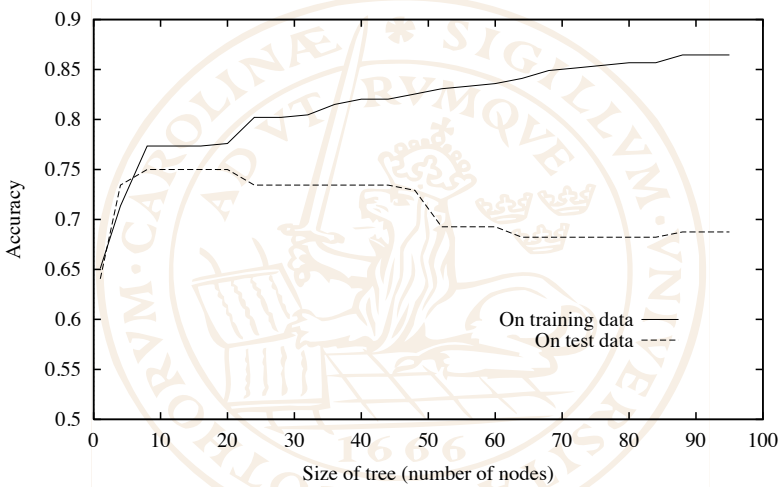
$$Gain(S, A) = Entropy(S) - \\ - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

LUND UNIVERSITY

AI@CS

Department of Computing Science

Overfitting

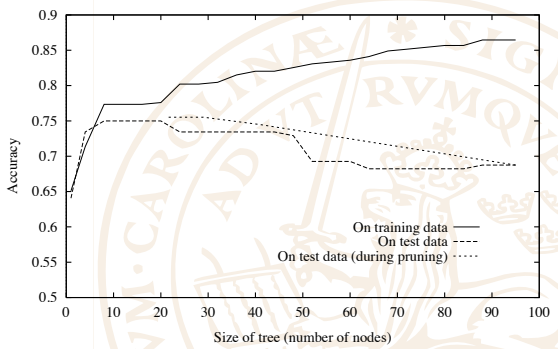


LUND UNIVERSITY

AI@CS

Department of Computing Science

Pruning



rules
C4.5, Quinlan