Machine Learning

Artificial intelligence

Kristóf Karacs
PPKE-ITK

Recap

- Concepts related to intelligence
- Agent model
- Problem solving by search
- Strategies in games
- Inference in First Order Predicate Logic

Outline

- Supervised vs. unsupervised vs. …
- Logical inference schemes
- Inductive learning
  - ID3 algorithm
  - Version spaces
  - ~~Inductive logic programming~~
- Learning theory

Machine Learning and AI

- Improve task performance through observation, teaching
- Acquire knowledge automatically for use in a task
- Learning is a key component in intelligence

Learning

- Supervised, unsupervised, semi-supervised, reinforcement
- Representation trade-off
  - Efficiency vs. expressive power
- There exist learning methods for several representation schemes

Applications

- Data mining
  - Big data, web mining
- Language/speech
  - Machine translation, text summarization, grammars
- Medical
  - Assessment of illness severity
- Vision
  - Face recognition, digit recognition, outdoor scene recognition
- Security
  - Intrusion detection, network traffic, credit fraud
- Social networks
  - Email traffic

Logical inference schemes

- *Deduction*: formal logical reasoning
  - Premises: 1. All men are mortal. 2. Aristotle is a man.
  - Conclusion: Aristotle is mortal.
- *Induction*: generalization
  - Premise: The sun has risen in the east every morning up until now.
  - Conclusion: The sun will also rise in the east tomorrow.
- *Abduction*: choosing an explanation
  - Premise: 1. Flu causes fever. 2. Peter has fever.
  - Conclusion: Peter has flu.

Approaches

- Inductive learning
  - Discovering general concepts from a limited set of examples (experience)
  - From a formal point of view the obtained knowledge is invalid
  - Supervised
    - Given input data as pairs of $(x_i, f(x_i))$
    - Generate a hypothesis for $f()$
  - Unsupervised
- Analytic or deductive learning
  - Based on explanations

Inductive learning

- Decision trees
- Version-spaces
  - More general, but less efficient
- Inductive logic programming

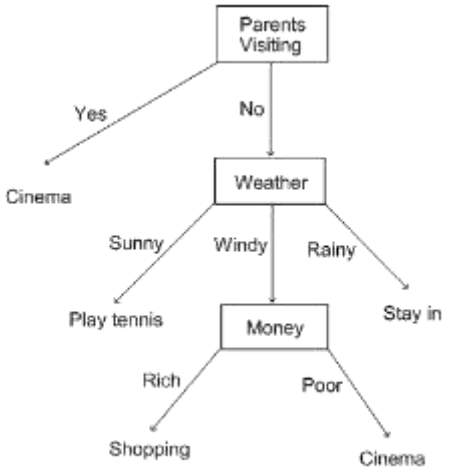
General Approach

- Formulate task
- Prior model (parameters, structure)
- Obtain data
- What representation should be used? (attribute/value pairs)
- Annotate data
- Learn/refine model with data (training)
- Use model for classification or prediction on unseen data (testing)
- Measure accuracy

Decision trees

- Measurements
- Nodes
- Node selection: ID3 algorithm
  - Maximizing information gain
  - Measure for information gain of an attribute: expected value of the information given by it
- Pruning

Sample decision tree



Information theory

- S: set of measurements
- A: an attribute with domain $V = \{v_1, \ldots, v_i, \ldots, v_n\}$
- S_v: set of measurements for which $A = v$
- Entropy:

$$H(S) = \sum_{i=1}^{n} -P(v_i)\log_2 P(v_i)$$

- Gain:

$$G(S, A) = H(S) - \sum_{v \in V} \frac{|S_v|}{|S|} H(S_v)$$

Example

	weather	got HW?	friend comes?	excursion
S_1	nice	Yes	Yes	Yes
S_2	cloudy	No	No	No
S_3	rainy	No	Yes	Yes
S_4	cloudy	Yes	Yes	Yes

- 3 input attributes
- 4 measurements

Information gain for attribute 'weather'

- $H(S) = -3/4 \log_2(3/4) – 1/4 \log_2(1/4) = 0.811$
- $|S_{nice}|/|S| * H(S_{nice}) = 1/4 * (-1/1 \log_2(1/1) - 0/1 \log_2(0/1)) = 1/4 * (1*0 – 0) = 0$
- $|S_{cloudy}|/|S| * H(S_{cloudy}) = 2/4*(-1/2 \log_2(1/2) - 1/2 \log_2(1/2)) = 1/2 * (-1/2 * (-1) – 1/2 * (-1)) = 1/2$
- $|S_{rainy}|/|S| * H(S_{rainy}) = 0$
- $G(S,A) = 0.811 – (0+0.5+0) = 0.311$

The ID3 algorithm

- Given a set of examples, S
  - Described by a set of attributes A_i
  - Categorised into categories c_j

1. Put the attribute the has the highest information gain in the root node (attribute A_1)
2. For each value v_i that A can take
   - Draw a branch and label each with corresponding v_i

The ID3 algorithm

3. For each branch with value v_i
   - If S_{vi} only contains examples in category c, then put that category as a leaf node in the tree
   - If S_{vi} is empty, then put the default category (which contains the most examples from S) as a leaf node in the tree
   - Otherwise construct subtree$_i$ by recursively calling decision tree with S_v, all attributes

Text Classification

- Is text$_i$ a new finance article?

Positive

Negative

Dow Closes Down 38, Nasdaq Gains 3

General Sense of Caution Pushes Stocks Mostly Lower; Dow Closes Down 38, Nasdaq Gains 3

The Associated Press

April 5, 8:59 pm ET — Investors concerned about the deteriorating situation in Iraq looked past solid earnings from General Electric Co. and Yahoo! Inc. Thursday, sending stocks mostly lower and leaving Wall Street with a loss for the holiday-shortened week.

The Dow Jones industrial average was down 38.12, or 0.4 percent, at 10,442.03. The Dow had gained more than 60 points in early trading before falling back. Broader stock indicators were narrowly mixed. The Standard & Poor's 500 index was down 1.21, or 0.1 percent, at 1,139.32, while the Nasdaq composite index gained 2.64, or 0.1 percent, to 2,062.88.

While GE, which posted earnings in line with Wall Street estimates, is seen as a gauge of the overall economy due to the conglomerate's diverse businesses, the threat of increased violence in Iraq and possible consequences from Tuesday's terrorism kept investors from making large bets.

"The market is now in a duel between good economic numbers, good earnings, and the situation in Iraq," said Peter Cardillo, chief strategist and senior vice president at S.W. Bach & Co. "With the market trading at the upper end of its trading range for the year, it induces people to take some money off the table."

Trading was quiet and volume light, with many investors and traders taking time off for the holidays. The stock market was scheduled to close for Good Friday.

For the week, the Dow lost 0.3 percent, while the S&P 500 and Nasdaq both dropped 0.7 percent. The losses followed the month's best week of 2004.

20 attributes

- Investors 2
- Dow 2
- Jones 2
- Industrial 1
- Average 3
- Percent 5
- Gain 6
- Trading 8
- Broader 5
- stock 5
- Indicators 6
- Standard 2
- Rolling 1
- Nasdaq 3
- Early 10
- Rest 12
- More 13
- first 11
- Same 12
- The 30

UConn defense knocks out Georgia Tech for second title

By Jack Carey, USA TODAY

SAN ANTONIO — Dominating inside, outside and especially defensively, Connecticut roared to its second men's basketball national championship in six years Monday night, rolling past Georgia Tech 82-73.

Jim Calhoun, center, and the rest of the UConn Huskies celebrate the school's second national title.

By Sue Ogrocki, AP

With the UConn women playing Tennessee for the national crown Tuesday night in New Orleans, Connecticut has a chance to become the first school to hold both basketball titles in the same year.

In its unexpected run to the final, Georgia Tech (28-10) had five games decided in the closing seconds, but Monday's contest was over early. **(Related item:** Box score)

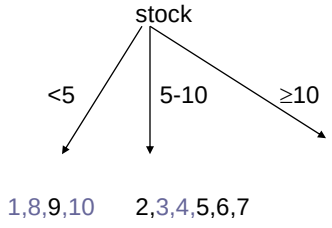
With All-America center Emeka Okafor (24 points, 15 rebounds) virtually unstoppable down low and smooth guard Ben Gordon scoring 21, the Huskies (33-6) were too much for Tech, which couldn't find the range until late against UConn's aggressive defenders.

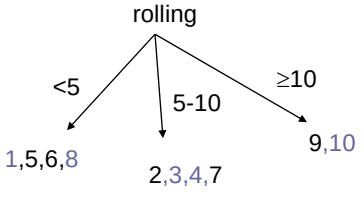
20 attributes

- Men's
- Basketball
- Championship
- UConn Huskies
- Georgia Tech
- Women
- Playing
- Crown
- Titles
- Games
- Rebounds
- All-America
- early
- rolling
- Celebrates
- Rest
- More
- First
- The
- same

Example

	stock	rolling	the	class
1	0	3	40	other
2	6	8	35	finance
3	7	7	25	other
4	5	7	14	other
5	8	2	20	finance
6	9	4	25	finance
7	5	6	20	finance
8	0	2	35	other
9	0	11	25	finance
10	0	15	28	other





Gain(stock) = 1 - [0.4 * H(0.1, 0.3) + 0.6 * H(0.4, 0.2)] =
= 1 - [0.4 ((-0.1 * -3.32) - (0.3 * -1.74)) +
+ 0.6 ((-0.4 * -1.32) - (0.2 * -2.32))] =
= 1-[0.303 + 0.5952] = 0.105

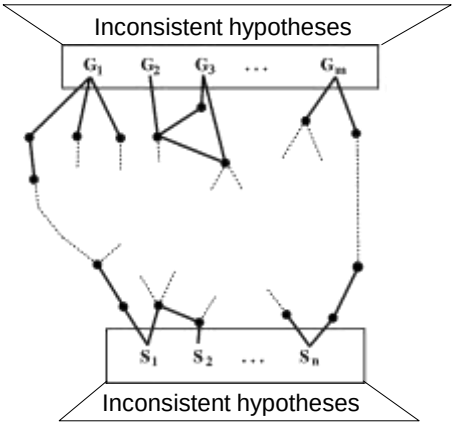
Gain(rolling) = 1 - [0.4 * H(0.5, 0.5) + 0.4 * H(0.5, 0.5) +
+ 0.2 * H(0.5, 0.5)] = 0

Issues

- Representation
  - How to map from a representation in the domain to a representation used for learning?
- Training data
  - How can training data be acquired?
- Amount of training data
  - How well does the algorithm do as we vary the amount of data?
- Which attributes influence learning most?
- Does the learning algorithm provide insight into the generalizations made?

Version space learning

- A technique for learning concepts
- Continuously maintains the set of consistent hypotheses

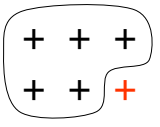


Version space learning

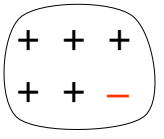
- Least commitment principle
  - G: most general set
  - S: most specific set
- Initialization
  - G: True
  - S: False
- A hypothesis H is **consistent** if H is
  - more specific than some element of G
  - and more general than some element of S

Handling inconsistence

- If S_i is false negative
  - replace S_i by all direct generalizations that classifies e as positive and is more specific than some element of G

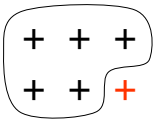


- If S_i is false positive
  - remove S_i from S

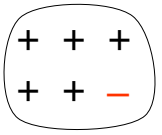


Handling inconsistence

- If G_i is false negative
  - remove G_i from G



- If G_i is false positive
  - replace G_i by all direct specializations that classifies e as negative and is more general than some element of S



Example

- Training set
  - S_1: {Japan, Honda, blue, 1980, economy} +
  - S_2: {Japan, Toyota, green, 1970, sports} –
  - S_3: {Japan, Toyota, blue, 1990, economy} +
  - S_4: {USA, Chrysler, red, 1980, economy} –
  - S_5: {Japan, Honda, white, 1980, economy +

Example

S_1: {Japan, Honda, blue, 1980, economy} +
S_2: {Japan, Toyota, green, 1970, sports} –
S_3: {Japan, Toyota, blue, 1990, economy} +
S_4: {USA, Chrysler, red, 1980, economy} –
S_5: {Japan, Honda, white, 1980, economy} +

- S1: G = {G1:(*,*,*,*,*)}
  S = {S1:(Japan, Honda, blue, 1980, economy)}
- S2: G = {G1:(*,Honda,*,*,*), G2:(*,*, blue,*,*),
  G3:(*,*,*, 1980,*), G4:(*,*,*,*, economy)}
  S = {S1:(Japan, Honda, blue, 1980, economy)}
- S3: G = {G2:(*,*,blue,*,*), G4:(*,*,*,*, economy)}
  S = {S1:(Japan,*, blue,*, economy)}
- S4: G = {G2:(*,*, blue,*,*), G4:(Japan,*,*,*, economy)}
  S = {S1:(Japan,*, blue,*, economy)}
- S5: G = {G4:(Japan,*,*,*, economy)}
  S = {S1:(Japan,*,*,*, economy)}

- The concept is “Japanese economy car”

Termination

- G ≠ S, but no more samples
- G = S → a single winner hypothesis
- S or G is empty
  - Inconsistent samples (may be due to noise)
  - Insufficient attributes
  - Chosen language is incapable to express the concept

Problems of VS learning

- Cannot handle noisy data
  - Version spaces collapses
- Unlimited disjunction
  - Wave fronts will not meet

Learning summary

What is the space of hypotheses to be considered when learning?

- If too large: no actual knowledge can be gained (generalization is not possible)
- If too small: it might not include the target function

Learning theory

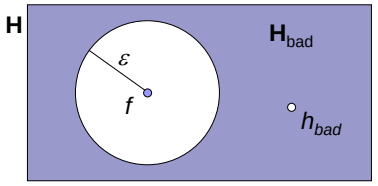
- Correctness of hypothesis *h*() has to be evaluated without knowing *f*(), the function to be learned
- A sufficiently large dataset can ensure an approximately good result with a high probability
- How many samples are needed to evaluate the correctness of a hypothesis?
  - *Probably Approximately Correct* learning (PAC learning)

Probably approximately correct (PAC) learning

- **X**: set of all possible examples
- *D*: distribution of examples
- **H**: set of possible hypotheses
- *m*: # of examples in training set

- Looking for an $h() \in \mathbf{H}$ being close to $f() \in \mathbf{H}$

Hypothesis space



- error(h) = $P(h(x) \neq f(x) \mid x$ is drawn from $D)$
- Hypothesis h is approximately correct if error(h) $\leq \varepsilon$
- P(h_{bad} is consistent with m examples) $\leq (1 - \varepsilon)^m$

PAC-learning

- P($\mathbf{H}_{bad}$ contains a consistent hypothesis) $\leq$ $|\mathbf{H}_{bad}|(1 - \varepsilon)^m \leq |\mathbf{H}|(1 - \varepsilon)^m$
- Let δ be an upper bound for this
  - Sample complexity function of the hypothesis space

$$m(\epsilon, \delta) \geq \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln|\mathbf{H}|\right)$$

  - If the complexity of the hypothesis space is less than exponential, then the function is learnable

PAC-learning

- For Boole functions: $|\mathbf{H}| = 2^{2^n}$
- Sample complexity grows as
  For $\varepsilon = \delta = 10^{-4}$
  $n = 2$ $m \geq 5{,}000$
  $n = 10$ $m \geq 3{,}000{,}000$
- Solution
  - Searching in space of simple solutions
  - Restricting the language of hypotheses

Summary

- Supervised vs. unsupervised vs. …
- Logical inference schemes
- Inductive learning
  - ID3 algorithm
  - Version spaces
  - Inductive logic programming
- Learning theory