

University of Toronto (Mississauga Campus)

CSC411- Machine Learning and Data Mining

Tutorial 1 – Jan 19th, 2007

Review

Data Mining and Machine Learning

From Wikipedia:

As a broad subfield of artificial intelligence, **machine learning** is concerned with the development of algorithms and techniques that allow computers to "**learn**". Some parts of machine learning are closely related to data mining.



Data mining (DM), also called **Knowledge-Discovery in Databases (KDD)**, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc..



Pictures are from <http://www.aaai.org/AITopics/html/machine.html>
<http://greenbay.usc.edu/csci577/fall2005/projects/team8/miner.bmp>



Data Mining vs. Machine Learning I

From [Frawley et al., 1991]:

Database Management	Machine Learning
Database is an active, evolving entity	Database is just a static collection of data
Records may contain missing or erroneous information	Instances are usually complete and noise-free
A typical field is numeric	A typical feature is binary
A database typically contains millions of records	Data sets typically contain several hundreds instances
AI should get down to reality	"Databases" is a solved problem and is therefore uninteresting



Data Mining vs. Machine Learning II

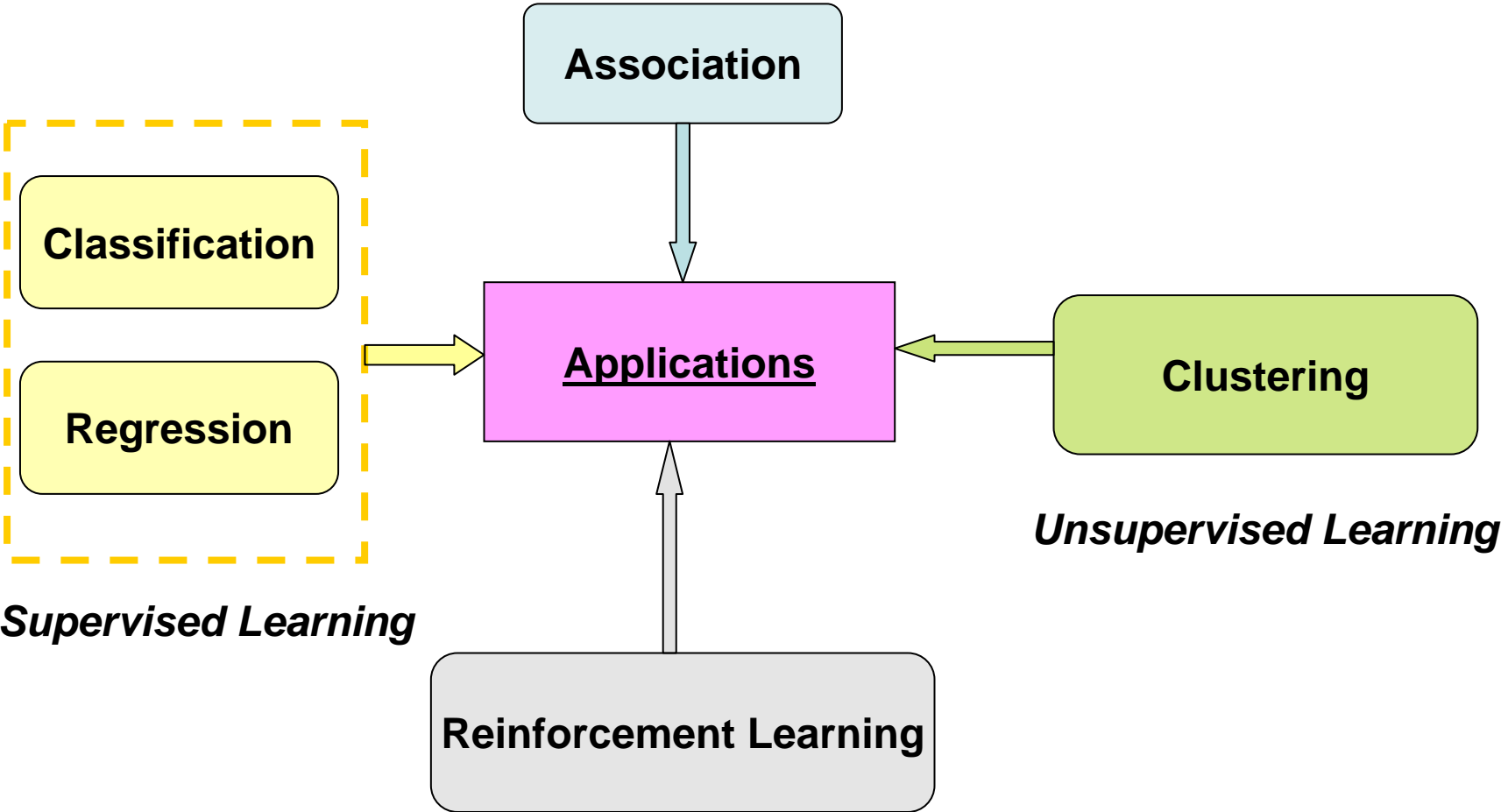


Since then, differences have partly disappeared (e.g. Machine Learning does not expect complete data anymore, data mining does not deal with dynamic aspect of databases very much. . .).

Some remaining discernible differences:

Data Mining	Machine Learning
<p>Descriptive models and patterns for existing data</p> <p>Emphasis on handling large datasets</p> <p>"Private Topics":</p> <ul style="list-style-type: none">• Data Visualization• Association rules• ...	<p>Predictive models and rules for future instances</p> <p>Emphasis on performance of models</p> <p>"Private Topics":</p> <ul style="list-style-type: none">• Learning behaviors (e.g. reinforcement learning)• Computational learning theory (what can be learned?)• ...

Applications



Classification – Bayesian Methods

Probabilistic approach, notation

Prior (unconditional) probability: $P(\text{Roll} = 3) = 1/6$

Joint probability: probability of combination of values of random variables:

$$P(R_1 = 6, R_2 = 6) = P(R_1 = 6 \wedge R_2 = 6) = 1/36$$

conditional probability $P(A|B)$: basic expressions in Bayesian formalism for probabilities
“probability of A given that all we know is B ”

posterior – conditioned on evidence - given that B is known with certainty

Bayes Rule:

$$\text{posterior} \longrightarrow P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

likelihood **prior**

evidence or normalization

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

$$P(A|B)P(B) = P(A, B) = P(B|A)P(A)$$

Bayesian classification

Bayes Rule:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Apply Bayes Rule: c is the class, $\{v\}$ observed attribute values:

$$P(c|\{v\}) = \frac{P(\{v\}|c)P(c)}{P(\{v\})}$$

$$P(c_k|\{v\}) = \frac{P(c_k)P(\{v\}|c_k)}{P(\{v\})}, \text{ where } k = 1, 2, \dots, n, \text{ and } P(\{v\}) = \sum_k P(c_k)P(\{v\}|c_k)$$

1. ML (Maximum Likelihood) Hypothesis

assume all hypotheses equiprobable a priori – simply maximize *data likelihood*:

$$c_{ML} = \arg \max_{c \in C} P(\{v\}|c)$$

2. MAP (Maximum A Posteriori) Class Hypothesis

$$c_{MAP} = \arg \max_{c \in C} P(c|\{v\}) = \arg \max_{c \in C} \frac{P(\{v\}|c)P(c)}{P(\{v\})}$$

Bayes optimal classifier:

$$c_{MAP} = \arg \max_{c_k \in C} P(v_1, v_2, \dots, v_N | c_k) P(c_k)$$

Naive Bayes Classifier

Bayes optimal classifier:

$$c_{MAP} = \arg \max_{c_k \in C} P(v_1, v_2, \dots, v_N | c_k) P(c_k)$$

Key simplifying assumption: *conditional independence*

$$P(v_1, v_2, \dots, v_N | c_k) = \prod_j P(v_j | c_k)$$

Naive Bayes classifier:

$$c_{NB} = \arg \max_{c_k \in C} P(c_k) \prod_j P(v_j | c_k)$$

Naïve Bayes Classifier Example

A researcher did a survey to study whether smoking leads to the cancer. You are asked to use Naïve Bayes Classifier to find out whether a female who is younger than 60 with smoking history will have cancer based on this survey.

Data table:

Example No.	Sex	Age	Smoking history	Cancer
1	Female	>60	Yes	Yes
2	Female	>60	Yes	No
3	Female	>60	Yes	Yes
4	Male	>60	Yes	No
5	Male	>60	No	Yes
6	Male	<=60	No	No
7	Male	<=60	No	Yes
8	Male	<=60	Yes	No
9	Female	<=60	No	No
10	Female	>60	No	Yes

Naive Bayes classifier:

$$c_{NB} = \arg \max_{c_k \in C} P(c_k) \prod_j P(v_j | c_k)$$

$P(\text{Cancer} = \text{Yes} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes})$
 $= P(\text{Cancer} = \text{Yes}) * P(\text{Sex} = \text{Female} \mid \text{Cancer} = \text{Yes}) * P(\text{Age} \leq 60 \mid \text{Cancer} = \text{Yes}) * P(\text{Smoking history} = \text{Yes} \mid \text{Cancer} = \text{Yes})$

$P(\text{Cancer} = \text{No} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes})$
 $= P(\text{Cancer} = \text{No}) * P(\text{Sex} = \text{Female} \mid \text{Cancer} = \text{No}) * P(\text{Age} \leq 60 \mid \text{Cancer} = \text{No}) * P(\text{Smoking history} = \text{Yes} \mid \text{Cancer} = \text{No})$

	Cancer = Yes	Cancer = No
Sex = Female	3 out of 5	2 out of 5
Age ≤ 60	1 out of 4	3 out of 4
Smoking History = Yes	2 out of 5	3 out of 5

	Cancer = Yes	Cancer = No
Sex = Female	3 out of 5	2 out of 5
Age<=60	1 out of 4	3 out of 4
Smoking History = Yes	2 out of 5	3 out of 5

	Cancer = Yes	Cancer=No
Sex = Female	0.6	0.4
Age<=60	0.25	0.75
Smoking History = Yes	0.4	0.6

Apply m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}, \text{ where}$$

$P = 0.5$ (binary classification), $m = 3$ (random)

- $n =$ the number of training examples for which $v = v_j$
- $n_c =$ number of examples for which $v = v_j$ and $a = a_i$
- $p =$ a priori estimate for $P(a_i|v_j)$
- $m =$ the equivalent sample size

	Cancer = Yes	Cancer=No
Sex = Female	0.56	0.43
Age<=60	0.31	0.56
Smoking History = Yes	0.43	0.56

	Cancer = Yes	Cancer=No
Sex = Female	0.56	0.43
Age<=60	0.31	0.56
Smoking History = Yes	0.43	0.56

$$\begin{aligned}
& P(\text{Cancer} = \text{Yes} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes}) \\
&= P(\text{Cancer}=\text{Yes}) * P(\text{Sex} = \text{Female} \mid \text{Cancer} = \text{Yes}) * P(\text{Age} \leq 60 \mid \text{Cancer} = \text{Yes}) * P(\text{Smoking history} = \text{Yes} \mid \text{Cancer} = \text{Yes}) \\
&= 0.5 * 0.56 * 0.31 * 0.43 = 0.037
\end{aligned}$$

$$\begin{aligned}
& P(\text{Cancer} = \text{No} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes}) \\
&= P(\text{Cancer}=\text{No}) * P(\text{Sex} = \text{Female} \mid \text{Cancer} = \text{No}) * P(\text{Age} \leq 60 \mid \text{Cancer} = \text{No}) * P(\text{Smoking history} = \text{Yes} \mid \text{Cancer} = \text{No}) \\
&= 0.5 * 0.43 * 0.56 * 0.56 = 0.069
\end{aligned}$$

Since $P(\text{Cancer} = \text{No} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes}) > P(\text{Cancer} = \text{Yes} \mid \text{Sex} = \text{Female}, \text{Age} \leq 60, \text{Smoking history} = \text{Yes})$, our answer is 'No'

Classification – K Nearest Neighbors Algorithm

algorithm: find example $\langle \mathbf{x}^*, c(\mathbf{x}^*) \rangle$ closest to test instance \mathbf{x}_q

output: $\hat{c}(\mathbf{x}_q) = c(\mathbf{x}^*)$

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{j=1}^d (a_j - b_j)^2}$$

Steps:

- 1) Determine parameter K = number of nearest neighbors
- 2) Calculate the distance between the query-instance and all the training samples
- 3) Select the K th nearest neighbors based on the K -th minimum distance
- 4) Assign the category Y to the selected instances
- 5) Use the majority of the category of nearest neighbors as the prediction value of the query instance

K Nearest Neighbor Example

A supermarket manager collects the in-store customers shopping history to improve the store sales. After the study, he came up the following table to assign the promotion code for the selected valuable customers.

Customer No.	Average Number of Items Purchased Per Week	Average Times of Visits Per Week	Should this customer be assigned promotion code?
1	10	1	Yes
2	12	2	Yes
3	6	1	No
4	20	1	Yes
5	7	3	Yes
6	5	2	No
7	7	2	No
8	4	1	No

Question: Can you help the manager to guess whether he should assign the promotion code to a new customer # 9, whose average weekly number of items purchased is 8 and visits number is 2? (Suppose we use $K = 3$)

CSC411 Machine Learning and Data Mining
Tutorial 1 – Jan 19th, 2007

Naïve Bayes Cancer Study Example

A researcher did a survey to study whether smoking leads to the cancer. You are asked to use Naïve Bayes Classifier to find out whether a female who is younger than 60 with smoking history will have cancer based on this survey.

Data table:

Example No.	Sex	Age	Smoking history	Cancer
1	Female	>60	Yes	Yes
2	Female	>60	Yes	No
3	Female	>60	Yes	Yes
4	Male	>60	Yes	No
5	Male	>60	No	Yes
6	Male	<=60	No	No
7	Male	<=60	No	Yes
8	Male	<=60	Yes	No
9	Female	<=60	No	No
10	Female	>60	No	Yes

K Nearest Neighbor Example

A supermarket manager collects the in-store customers shopping history to improve the store sales. After the study, he came up the following table to assign the promotion code for the selected valuable customers.

Customer No.	Average Number of Items Purchased Per Week	Average Times of Visits Per Week	Should this customer be assigned promotion code?
1	10	1	Yes
2	12	2	Yes
3	6	1	No
4	20	1	Yes
5	7	3	Yes
6	5	2	No
7	7	2	No
8	4	1	No

Question: Can you help the manager to guess whether he should assign the promotion code to a new customer # 9, whose average weekly number of items purchased is 8 and visits number is 2? (Suppose we use $K = 3$)