Chapter VI: Classification

Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

Chapter VI: Classification

Classification: basic concepts.

Decision-tree induction.

Bayes classification methods.

Rule-based classification.

Model evaluation and selection.

Techniques to improve classification accuracy: ensemble methods.

Summary.

Supervised vs. unsupervised learning

Supervised learning (classification).

- Supervision:
  - The **training data** (observations, measurements, etc.) are accompanied by **labels** indicating the **class** of the observations.
  - New data is classified based on a **model** created from the training data.

Unsupervised learning (clustering).

- The class labels of training data are unknown.
  - Or rather, there are no training data.
- Given a set of measurements, observations, etc.,
  the goal is to find classes or clusters in the data.
  - See next chapter.

Prediction problems: classification vs. numerical prediction

Classification:

- Predicts **categorical class labels** (discrete, nominal).
- Constructs a model based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

Numerical prediction:

- Models **continuous-valued functions**.
- I.e. predicts missing or unknown (future) values.

Typical applications of classification:

- Credit/loan approval: Will it be paid back?
- Medical diagnosis: Is a tumor cancerous or benign?
- Fraud detection: Is a transaction fraudulent or not?
- Web-page categorization: Which category is it?

Classification – a two-step process

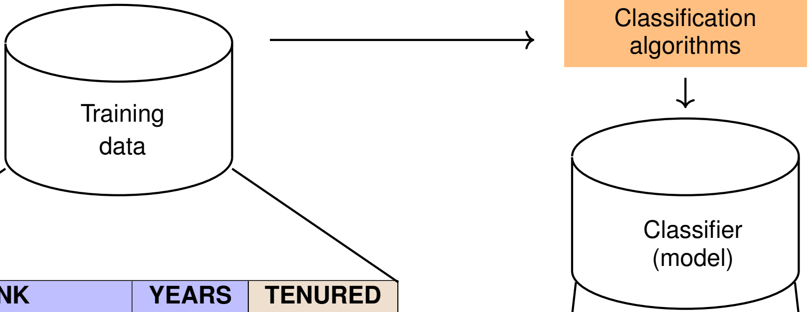
Model construction: describing a set of predetermined classes:

- Each tuple/sample is assumed to belong to a predefined class, as determined by the **class-label attribute**.
- The set of tuples used for model construction is the **training set**.
- The **model** is represented as classification rules, decision trees, or mathematical formulae.

Model usage, for classifying future or unknown objects:

- Estimate **accuracy** of the model:
  - The known label of **test samples** is compared with the result from the model.
  - **Accuracy rate** is the percentage of test-set samples that are correctly classified by the model.
  - Test set is independent of training set (otherwise overfitting).
- If the accuracy is acceptable, **use the model** to classify data tuples whose class labels are not known.

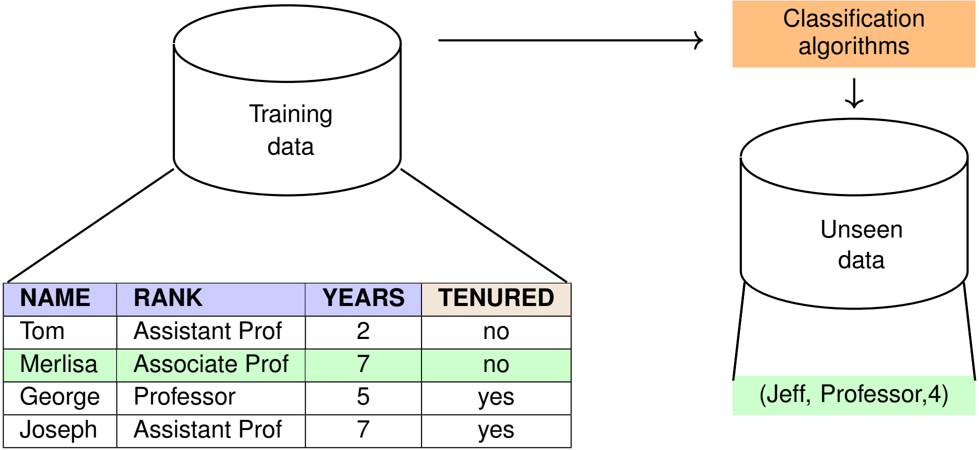
Classification – a two-step process



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

`if` RANK ='Professor'
`or` YEARS >6
`then` TENURED ='yes'

Process (II): using the model in prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

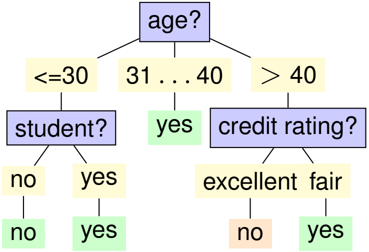
Chapter VI: Classification

Decision-tree induction: an example

Training dataset: buys_computer.

The dataset follows an example of Quinlan's ID3 (playing tennis).

Resulting tree:



age	income	student	credit_rating	buys_coputer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
$31 \ldots 40$	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
$31 \ldots 40$	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	no	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
$31 \ldots 40$	medium	no	excellent	yes
$31 \ldots 40$	high	yes	fair	yes
> 40	medium	no	excellent	no

Algorithm for decision-tree induction

Basic algorithm (a greedy algorithm):

- Tree is constructed in a **top-down recursive divide-and-conquer manner.**
- Attributes are categorical.
  - If not: discretize in advance.
- At start, all the training examples are at the root.
- Examples are **partitioned recursively** based on selected attributes.
- Test attributes are selected on the basis of a heuristic or statistical measure.
  - E.g. information gain – see on the next slide.

Conditions for stopping partitioning:

- All samples for a given node belong to the same class.
- There are no remaining attributes for further partitioning.
  - Majority voting is employed for classifying the leaf.
- There are no samples left (i.e. partition for particular value is empty).

Attribute-selection measure: information gain (ID3/C4.5)

Attribute selection: information gain

Class P: buys_computer = "yes"

Class N: buys_computer = "no"

$$\text{Info}(D) = I(9,5) = -\tfrac{9}{14}\log_2(\tfrac{9}{14}) - \tfrac{5}{14}\log_2(\tfrac{5}{14}) = 0.94$$

age	p	n	$I(p,n)$
≤ 30	2	3	0.971
$31\ldots40$	4	0	0
> 40	3	2	0.971

Similarly,

$$\text{Gain}(\texttt{income}) = 0.029,$$
$$\text{Gain}(\texttt{student}) = 0.151,$$
$$\text{Gain}(\texttt{credit\_rating}) = 0.048.$$

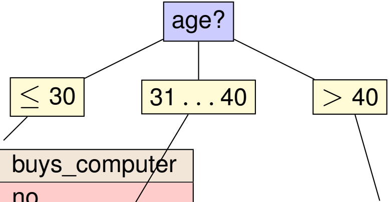
$$\text{Info}_{\text{age}}(D) = \tfrac{5}{14}I(2,3) + \tfrac{4}{14}I(4,0) + \tfrac{5}{14}I(3,2) = 0.694.$$

$\frac{5}{14}I(2,3)$ means "age ≤ 30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence,

$$\text{Gain}(\text{age}) = \text{Info}(D) - \text{Info}_{\text{age}}(D) = 0.246.$$

age	income	student	credit_rating	buys_computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
$31\ldots40$	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
$31\ldots40$	low	yes	excellent	yes
≤ 30	medium	no	fair	no
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
$31\ldots40$	medium	no	fair	yes
$31\ldots40$	high	yes	fair	yes
> 40	medium	no	excellent	no

Partitioning in the example



income	student	credit_rating	buys_computer
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	
high	no	fair	[illegible]
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	student	credit_rating	buys_computer
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Thank you for your attention.
Any questions about the sixt chapter?

Ask them now, or again, drop me a line:
luciano.melodia@fau.de.