Business Intelligence and Data Management

academic year 2019-2020

Dr. Ekaterini Ioannou, Department of Management, University of Tilburg

Lecture 6

topic: Naïve Bayes

material: Chapters 8 (book "Data Mining for Business Intelligence")
(two simple example https://www.youtube.com/watch?v=Aly89gurkB4 and https://www.youtube.com/watch?v=EGKeC2S44Rs)

Summary from previous lectures

Data Mining:

- Provides results that are useful for decision making

Process involves multiple steps:

- Business understanding → Data preparation → Model building → Testing & Evaluation → Deployment

Various options for the model:

  - Regression
  - Clustering
  - Decision trees
  - Association rules
  - etc.

Summary of previous lectures / Agenda

Date	Lecture contents		Lecturer	Lab topics	Test
Jan-27	1	Intro. to BI+ Data Management	Caron		
Jan-30				SQL-1	1
Jan-28	2	Data warehousing	Caron		
Feb-06				SQL-2	2
Feb-03	3	OLAP business databases & dashboard	Caron		
Feb-13				SQL-3 & OLAP	3a & 3b
Feb-10	4	Data mining introduction	Ioannou		
	5	Regression models			
Feb-17	6	Naïve Bayes			
	7	k nearest neighbors			
Feb-20				Bayes & neighbors	4
Feb-27	8	Performance measures	Ioannou		
Mar-02	9	Decision trees	Ioannou		
Mar-05				Dec. trees	5
Mar-09	10	Association rules	Ioannou		
Mar-11,12&13				Ass. Rules	6
Mar-16	11	Clustering (+20 mins exam preparation)	Ioannou		
Mar-19				Clustering	7

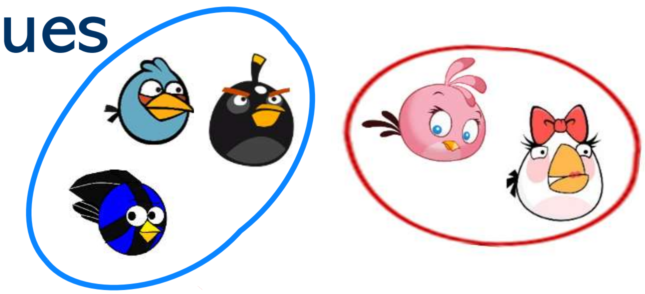
classification techniques

Our focus is currently on Data Mining models

Classification vs. Clustering

Classification vs. Clustering

- Both result in a categorization of records into one or more classes based on their values



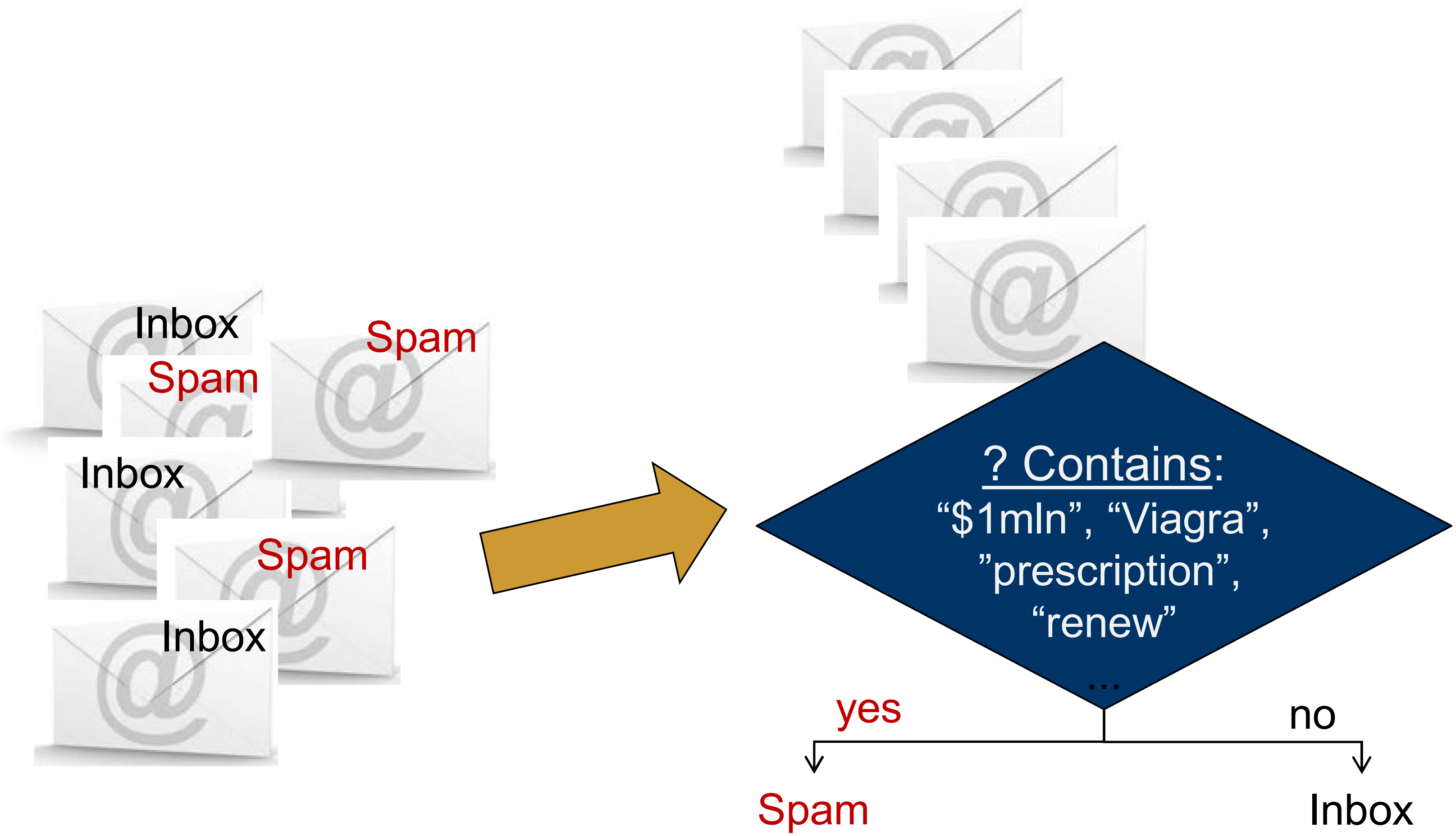
Classification:

- Training a model that allows classifying new records to one of the classes
- Assumes the existence of predefined classes

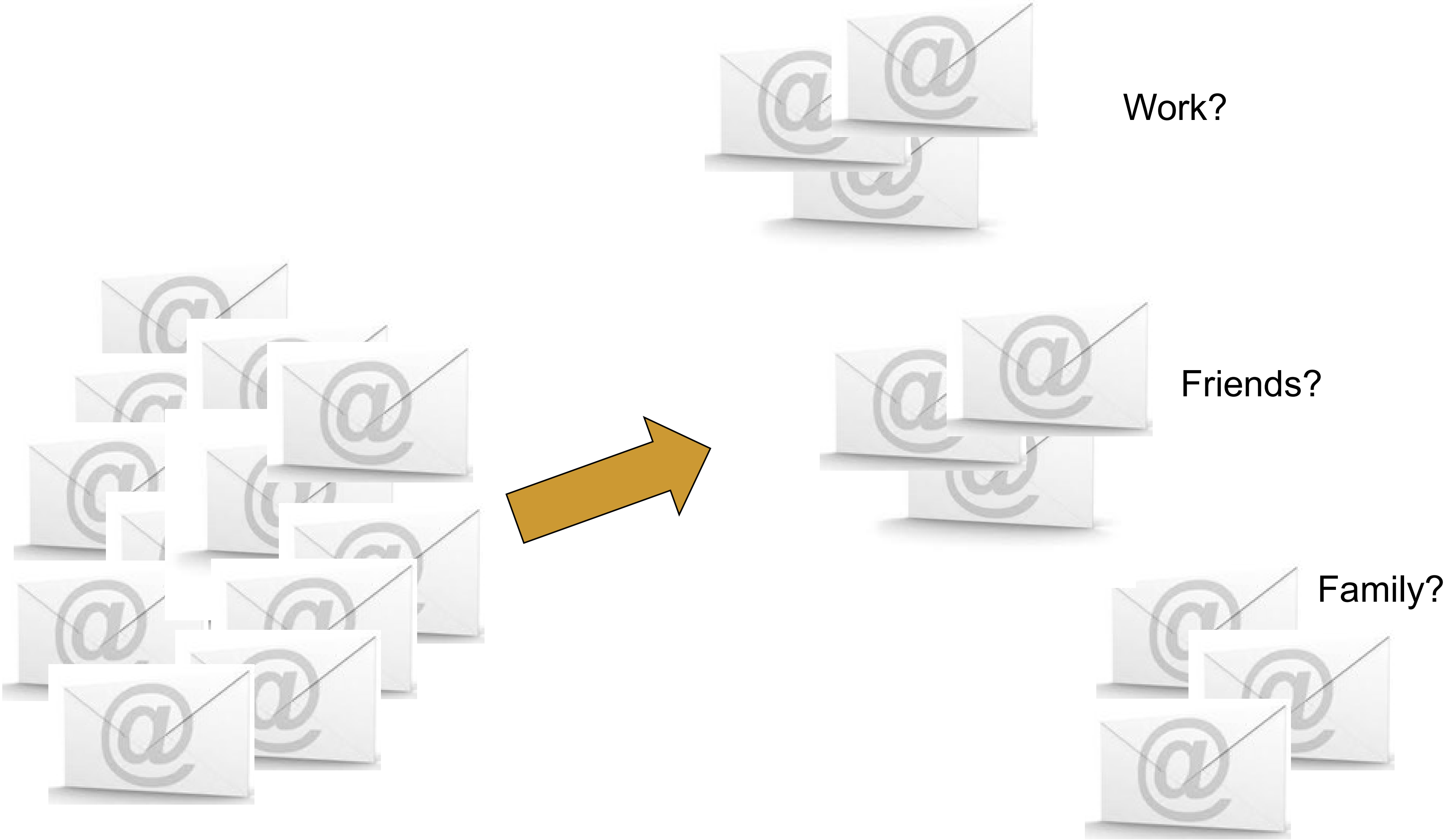
Clustering:

- Divides the records into clusters
- Records with high similarity reside inside a cluster and records of two clusters are dissimilar

Classify e-mails into “Spam” and “Inbox”



Clustering e-mails



Classification Process Example

Classify e-mails to spam vs. inbox

- Given (labeled) records of both types
  - Train a classifier to discriminate between these two
  - During operation, use classifier to select destination folder for new email: *Inbox* or *Spam* folder?

- Many issues to consider along the process
  - Feature extraction/construction,
    - e.g. convert free-form documents into a feature space
  - Feature selection
    - eliminate noise and redundancy while retaining "signal"
  - Classification model/algorithm evaluation and selection

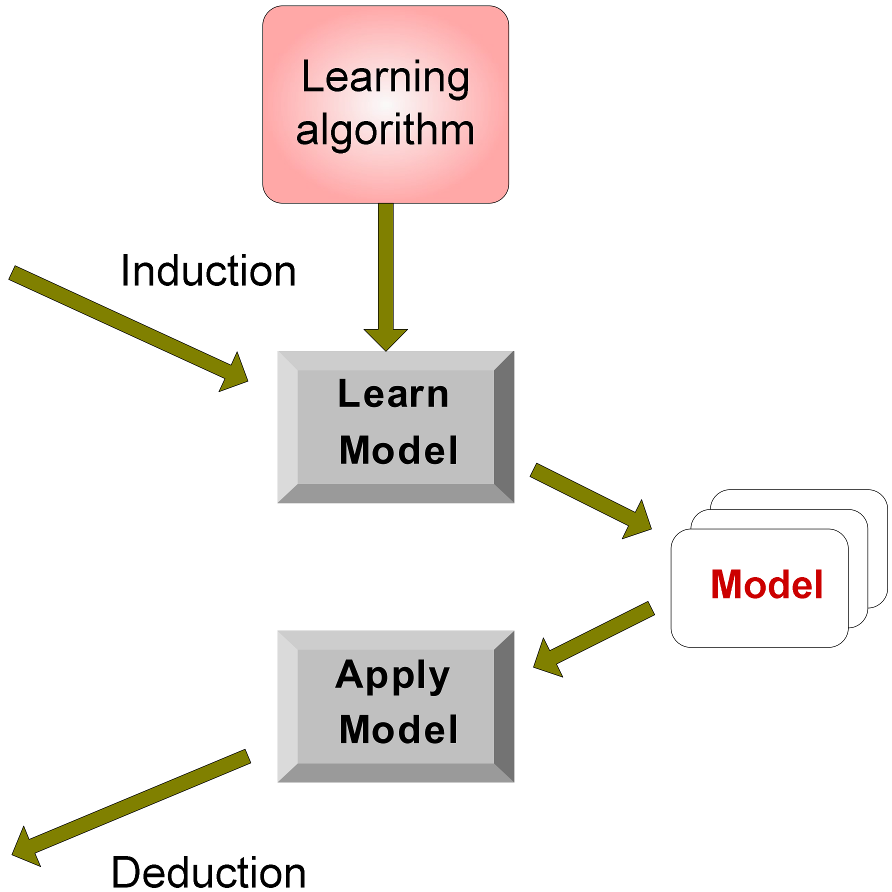
Why do we need Classification?

- Organizing documents is hard work
  - Route email messages into folders
  - Route help-desk inquiries to correct staff
  - Place documents in predefined categories/topic hierarchy
- Decide about (predefined) user interests/skills/…
  - User modeling
  - Instead of using human-authored expert system, let computer to induce rules or models from log data

Classification: Model Induction vs. Application

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

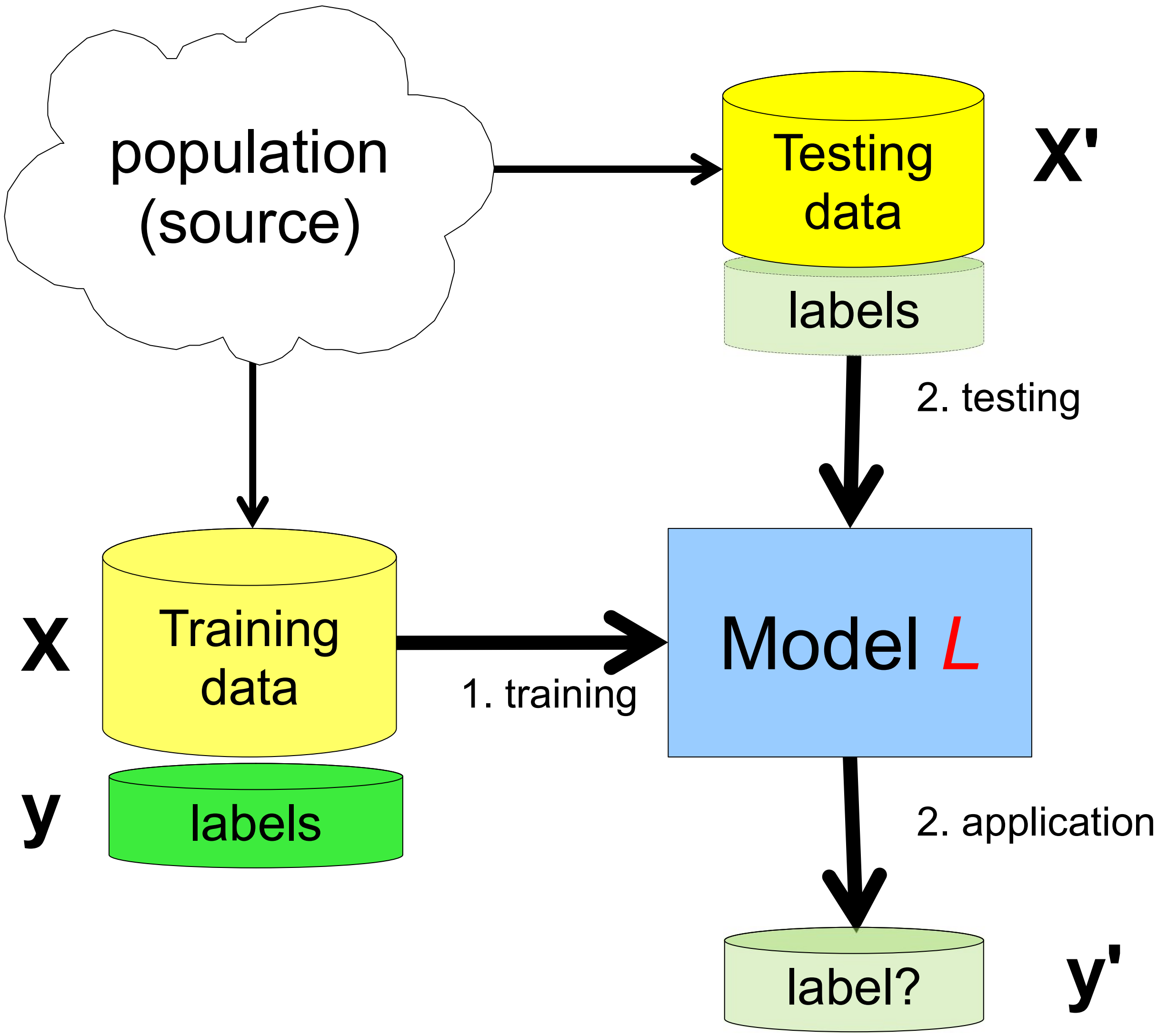
Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Classification



Training:

y = *L* (X)

Application:
use *L*
for an unseen data

y' = *L* (X')

Classification: Learning & Applying

X y

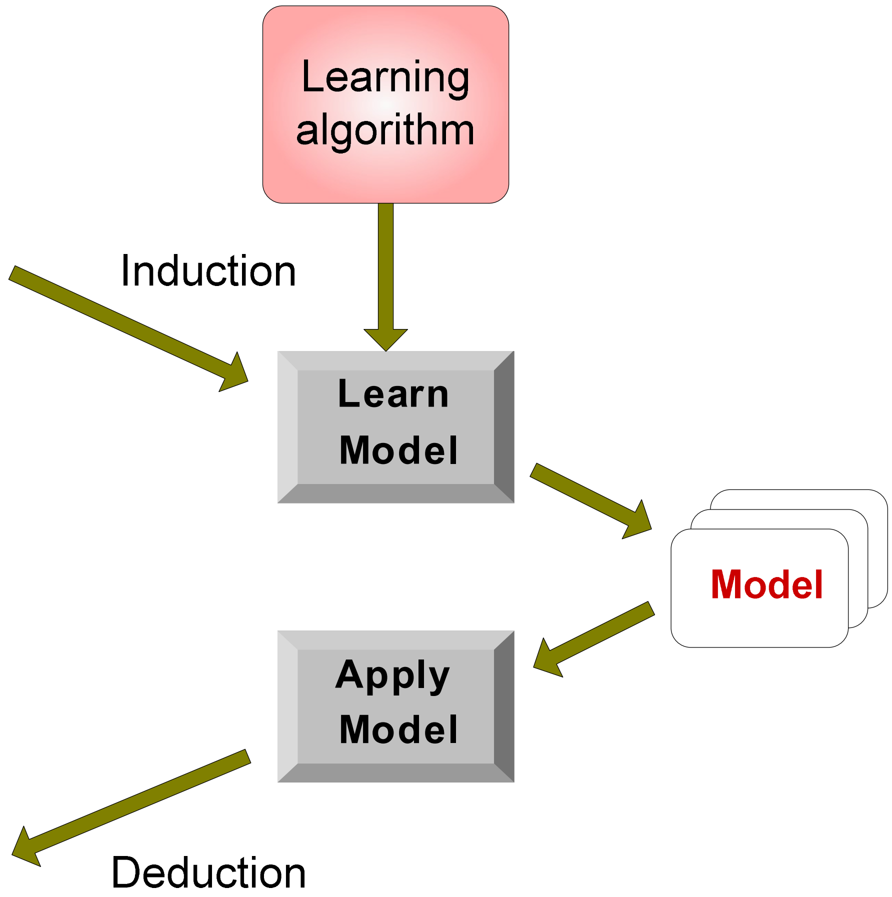
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

X' y'

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classification Techniques

- Naïve Bayes
- Nearest Neighbor
- Decision Trees
- Support Vector Machines
- Logistic regression
- Deep learning
- Ensemble classification
- ….

Naïve Bayes

Motivation

- Weather is currently: Outlook is *Sunny*, Temperature is *Cool*, Humidity is *High*, & Wind is *Strong*
- Given the current weather conditions, decide if we can play tennis
- Consider previous situations, i.e.,
  - each day is a record
  - variable of interest is if we played tennis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Sunny	Cool	High	String	???

Reminder (from previous slides)

- Variables are any measurement on the records
- Dependent variables (denoted as y):
  - The ones we want to predict
  - E.g., PlayTennis
- Independent variables (denoted as X):
  - The variables that explain the dependent ones
  - E.g., Outlook, Temperature, Humidity, and Wind

	X_1	X_2	X_3	X_4	y
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes

Naive Bayes Classifier

- Given n features (independent variables)
  - i.e., $\mathrm{X1}, \mathrm{X2}, \ldots, \mathrm{X}n$
- For all values of y compute the probability:

$$\mathrm{P}(\mathrm{y} \mid \mathrm{X1}, \mathrm{X2}, \ldots, \mathrm{X}n)$$

- Choose value of y that maximizes the probability
- This is denoted as

$$\underset{y}{\operatorname{argmax}}\ \mathrm{P}(\mathrm{y} \mid \mathrm{X1}, \mathrm{X2}, \ldots, \mathrm{X}n)$$

Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
…	…				
D15	Sunny	Cool	High	String	???

- Dependent var., y:
  - The ones we want to predict, e.g., PlayTennis
- Independent var., $X1, X2$, …:
  - The variables that explain the dependent ones, e.g., Outlook, Temperature, Humidity, and Wind
- **Naive Bayes Classifier:**
  - P(PlayTennis=Yes | Outlook=S, Temperature=C, Humidity=H, Wind=S)
  - P(PlayTennis=No | Outlook=S, Temperature=C, Humidity=H, Wind=S)
  - Answer is the one with highest probability, i.e.,

    Yes, if P(PlayTennis=Yes | …) ≥ P(PlayTennis=No | …) or

    No, otherwise

Bayes theorem

- For all values of y compute the probability:

$$P(y \mid X1, X2, \dots, Xn)$$

- Bayes rule is a standard formula for inverting conditional probabilities

$$P(y \mid X1, X2, \dots, Xn) = \frac{P(X1, X2, \dots, Xn \mid Y)\, P(y)}{P(X1, X2, \dots, Xn)}$$

- Denominator can be left out (seen as a constant since it does not depend on y and the values of the features $X1, \dots, Xn$ are given)

$$P(y \mid X1, X2, \dots, Xn) = P(X1, X2, \dots, Xn \mid y)\, P(y)$$

Naïve Bayes Assumption

- Current formula

$$P(y \mid X1, X2, \ldots, Xn) = P(X1, X2, \ldots, Xn \mid y)\, P(y)$$

- Naive conditional independence: assume that all features are independent given the class label y

$$P(X1, X2, \ldots, Xn \mid y)$$
$$= P(X1 \mid y)\, P(X2 \mid y) \ldots P(Xn \mid y)$$
$$= \prod_i P(Xi \mid y)$$

$$\boxed{P(y \mid X1, X2, \ldots, Xn) = P(y) \prod_i P(Xi \mid y)}$$

ITY

Example

- Naive Bayes Classifier:

$$P(\, y \mid X1, X2, \dots, Xn \,) = P(y) \prod_{i} P(\, Xi \mid y \,)$$

- Given the current weather conditions, decide if we can play tennis or not
  - P(PlayTennis=Yes | Outlook=S, Temperature=C, Humidity=H, Wind=S)
  - P(PlayTennis=No | Outlook=S, Temperature=C, Humidity=H, Wind=S)

➔P(PlayTennis=Yes | ...) = P(PT =Yes) .

P(Outlook=S | PT=Y) . P(Outlook=S | PT=Y) .

P(Humidity=H | PT=Y) . P(Wind=S | PT=Y) .

P(Temperature=C | PT=Y)

Example

P(PlayTennis=Yes | …)

= P(PT =Yes) .

P(Outlook=S | PT=Y) .

P(Humidity=H | PT=Y) .

P(Wind=S | PT=Y) .

P(Temp.=C | PT=Y)

Day	Outlook	Temp.	Humid.	Wind	PT
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

$P(\text{PlayTennis} = \text{Yes})$

Day	Outlook	Temp.	Humid.	Wind	PT
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- We have 9 days in which we played tennis
- Out of 14 days

→ P(PlayTennis =Yes) = 9/14

Example

$$P(\text{Outlook} = \text{Sunny} \mid \text{PlayTennis} = \text{Yes})$$

Day	Outlook	PT
~~D1~~	~~Sunny~~	~~No~~
~~D2~~	~~Sunny~~	~~No~~
D3	Overcast	Yes
D4	Rain	Yes
D5	Rain	Yes
~~D6~~	~~Rain~~	~~No~~
D7	Overcast	Yes
~~D8~~	~~Sunny~~	~~No~~
D9	Sunny	Yes
D10	Rain	Yes
D11	Sunny	Yes
D12	Overcast	Yes
D13	Overcast	Yes
~~D14~~	~~Rain~~	~~No~~

- We have 2 days with Sunny outlook
- Out of 9 days in which we played tennis

→ $P(\text{Outlook} = \text{Sunny} \mid \text{PlayTennis} = \text{Yes})$

$= 2/9$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Sunny	Cool	High	String	???

Day	Outlook	PT
D1	Sunny	No
D2	Sunny	No
~~D3~~	~~Overcast~~	~~Yes~~
~~D4~~	~~Rain~~	~~Yes~~
~~D5~~	~~Rain~~	~~Yes~~
D6	Rain	No
~~D7~~	~~Overcast~~	~~Yes~~
D8	Sunny	No
~~D9~~	~~Sunny~~	~~Yes~~
~~D10~~	~~Rain~~	~~Yes~~
~~D11~~	~~Sunny~~	~~Yes~~
~~D12~~	~~Overcast~~	~~Yes~~
~~D13~~	~~Overcast~~	~~Yes~~
D14	Rain	No

What is the following probability:

P (Outlook = Sunny | PlayTennis = No) ?

a) 3/5

b) 3/6

c) Unknow (since D15 has no value for PT)

d) 3/14

Example

PlayTennis=Yes	PlayTennis=No
9/14	5/14

Outlook	PlayTennis=Yes	PlayTennis=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temp.	PlayTennis=Yes	PlayTennis=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	PlayTennis=Yes	PlayTennis=No
High	3/9	4/5
Normal	6/9	1/5

Wind	PlayTennis=Yes	PlayTennis=No
Strong	3/9	3/5
Weak	6/9	2/5

- P(PlayTennis=Yes | Outlook=S, Hum.=H, Wind=S, Temp.=C)
  = P(PT=Y) . P (Outlook=S| PT=Y) . P (Hum.=H | PT=Y) .
  P(Wind=S | PT=Y) . P (Temp.=C | PT=Y)
  = 9/14 . 2/9 . 3/9 . 3/9 . 3/9 = 0.0053
- P(PlayTennis=No | Outlook=S, Hum.=H, Wind=S, Temp.=C)
  = P(PT=No) . P (Outlook=S| PT=No) . P (Hum.=H | PT=No) .
  P(Wind=S | PT=No) . P (Temp.=C | PT=No)
  = 5/14 . 3/5 . 4/5 . 3/5 . 1/5 = 0.0206
- P(PlayTennis=No | …) > P(PlayTennis=Yes | …) → No

Additional Concerns

Laplace Smoothing

- From previous slides, the probability of α given classs c:

$$P(\alpha \mid c) = \frac{\text{count}(\alpha, c)}{\text{count}(c)}$$

- P(Outlook="Sunny" | PlayTennis="Yes") = 0

Day	Outlook	PT
D1	Sunny	No
D2	Sunny	No
D3	Overcast	Yes
D4	Rain	Yes
D5	Rain	Yes
D6	Rain	No
D7	Overcast	Yes
D8	Sunny	No
D9	Rain	Yes
D10	Rain	Yes
D11	Overcast	Yes
D12	Overcast	Yes
D13	Overcast	Yes
D14	Rain	No

Problem:

- An attribute value doesn't occur with every class
- Probability of α given class c becomes 0

Laplace Smoothing

- Having a probability zero is problematic, because it wipes out all information in other probabilities
- We need to find a solution for this!

Laplace Smoothing, or Correction, or Estimator

- Incorporates a small-sample correction in every probability computation
- Increase the numerator/denominator
- Thus, no probability will be zero

$$P(\alpha \mid c) = \frac{\text{count}(\alpha, c) + 1}{\text{count}(c) + \text{number}-\text{of}-\text{values}-\text{in}-\text{class}}$$

Numerical Attribute Values

- Attributes can also be is numeric
- E.g., salary, temperature
- It is not likely that you find the value of a new record in the training set
- Estimate the distribution of the attribute in each class (then compute as explained)
- Assume normal distribution, for simplicity

- Details not covered in the course

Summary

Naive Bayes is Not So Naïve:

- Its beauty is in its simplicity
- Ability to handle categorical variables directly
- Computational efficient
- Good classification performance, especially when the number of predictors is very large

Negative aspects:

- Requires a very large number of records to obtain good results
- Independence assumption may not hold for some attributes