# Lect2: Classification #1
# Basic methods

Tokyo Tech.
Intro. to Comp. & Data
Lecture week2

Basic and well-known "methods" for classification

1. Introduction to classification.
2. Naive Bayes.
3. Decision Tree.
4. On Exercise #2. (Intro. to classification, again)

* Some of the slide materials (in particular, green ones) are from the slides of the authors of the textbook and their group at the University of Waikato.

* In this course, terms that I use for the class (which may not be standard sometimes) are quoted, e.g., "method".

# 1. Introduction to classification

Classification is a task to obtain a model for predicting the class value of given instances.

glossaries

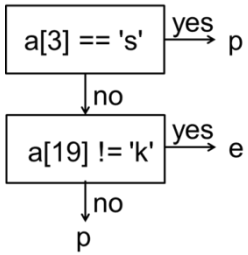
Usually, 1 (yes, positive) or 0 (no, negative), but sometimes {0, 1, 2, 3}, etc.

"method" = model + ML algo.

class or class value
= the target attribute of classification.

model
= a rule for predicting the class of each instance.

E.g., decision list




machine learning algorithm
= a way to compute a model from a given example set.

The mushroom exercise is a typical example of classifications

```
                  1 1 1 1 1 1 1 1 1 1 2 2
  1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
p f s n f s f c n b t s s p w p w o e w v d
p k s e f f f c n b t k k w p p w o e w v p
e f f w f n f w b n t f f w w p w o e k s g
e x s w t l f c b n e s s w w p w o p n n g
e x y u f n f c n p e s f w w p w o f h v d
e x y g t n f c b n t s s p g p w o p k y d
e f f e t n f c b u t s s p w p w o p n v d
p f y y f n f w n y e y y y y p y o e w c l
e b s y t l f c b n e s s w w p w o p k s m
p x s w f c f w n g e s s w w p w o p k v d
e s f g f n f c n k e s s w w p w o p k y u
p x s n f y f c n b t k k p p p w o e w v d
e x f g t n f c b p t s s w g p w o p k v d
                    ⋮
```

UC Irvine ML Repository, 1987

# 1. Introduction to classification

glossaries (cont.)

instance = an example data.
data set (or, sample) = a set of instances.
attribute = a feature of each instance.
nominal (or, categorical) attribute
= an attribute whose value is from a finite set.
numeric atteribute
= an attribute whose value is a (real) number.

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | 85 | 85 | False | No |
| Sunny | 80 | 90 | True | No |
| Overcast | 83 | 86 | False | Yes |
| Rainy | 75 | 80 | False | Yes |
| ... | ... | ... | ... | ... |

A data set of examples of the relation between whether and "play"

# 2. Naive Bayes

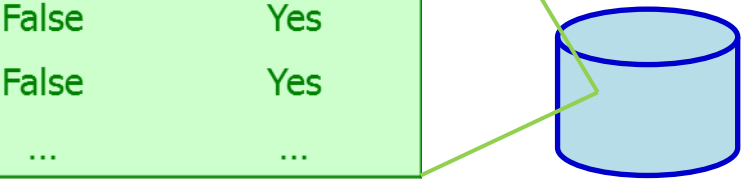
Naive Bayes is a ML algorithm to obtain a probabilistic classification model.

A probabilistic classification model
= probabilities of class values for a given instance.

$$\Pr[\, P = \text{Yes} \mid (\mathit{Weather}, T, H, W) = (\text{sunny}, 80, 90, \text{windy})\,]$$
$$\Pr[\, P = \text{No} \mid (\mathit{Weather}, T, H, W) = (\text{sunny}, 80, 90, \text{windy})\,]$$

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | 85 | 85 | False | No |
| Sunny | 80 | 90 | True | No |
| Overcast | 83 | 86 | False | Yes |
| Rainy | 75 | 80 | False | Yes |
| ... | ... | ... | ... | ... |

class value



A data set of examples of the relation between whether and "play"

# 2. Naive Bayes

## 2.1. Basics on probability (from Probability Theory)

For each *probabilistic event* $E$, we write $\Pr[E]$ to denote the *probability* that $E$ occurs. More concretely, we assume random variables and *probability distributions* on these random variables and specify $E$ in terms of the event that such random variables take some values, e.g., $E \Leftrightarrow X = 5$, $E \Leftrightarrow Y \in [0.5,1.5]$.

The *joint probabilistic* $\Pr[A \wedge B]$ is defined by

$$\Pr[A \wedge B] = \Pr[A|B] \times \Pr[B]$$

where $\Pr[A \mid B]$ is *conditional probability*.

We say that $A$ and $B$ are independent if $\Pr[A \wedge B] = \Pr[A] \cdot \Pr[B]$ or equivalently, $\Pr[A \mid B] = \Pr[A]$ because

in general, could be < or >

# 2. Naive Bayes

Bayes's rule is a way to compute conditional probability $\Pr[\, H \mid E \,]$ as follows.

$$\Pr[H|E] = \frac{\Pr[E|H] \times \Pr[H]}{\Pr[E]}$$

because...

$$\Pr[H|E] = \frac{\Pr[H \wedge E]}{\Pr[E]} = \frac{\Pr[E|H] \times \Pr[H]}{\Pr[E]}$$

**Thomas Bayes**
**1702－1761 in England**

$\Pr[\, H \,]$ : a priori probability
i.e., the probability of $H$ *before* evidence $E$ is seen.

$\Pr[\, H \mid E \,]$ : a posteriori probability
i.e., the probability of $H$ *after* evidence $E$ is seen.

# 2. Naive Bayes

## 2.2. Probabilistic assumptions for Data Mining

We assume a certain *probability distribution* $\mathcal{D}$ on instances (i.e., tuples of attribute values) of an assumed "domain", and our data set (or, sample) is a good approximation of $\mathcal{D}$.

Sample

(sunny, 80, 90, windy, no play)
(sunny, 70, 50, no wind, play)
⋮

$\hat{\mathcal{D}}$ ≒ $\mathcal{D}$

empirical dist.

event

*random variable*

E.g.,

$\Pr[\ (Weather, T, H, W, P) = (\text{sunny, 80, 90, windy, no})\ ] = 0.12$

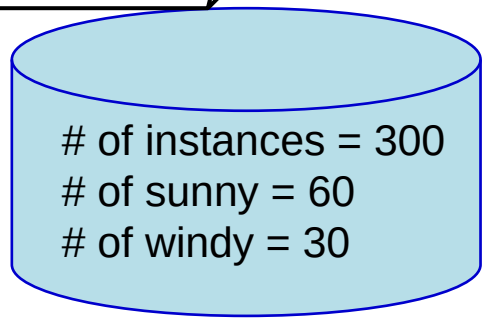
$\Pr[\ \underbrace{P = \text{no}}_{A}\ |\ \underbrace{(Weather, T, H, W) = (\text{sunny, 80, 90, windy})}_{B}\ ] = 0.8$

$\Pr[\ A \mid B\ ]$ or simply, $\Pr[\ \text{no} \mid (\text{sunny, 80, 90, windy})\ ]$

# 2. Naive Bayes

Probabilistic assumption for Naive Bayes

All attributes are independent random variables.
(as you like!)




$\hat{\mathcal{D}}$

$\Pr_{\hat{\mathcal{D}}}[\ \mathit{Weather} = \text{sunny}\ ] = 60/300 = 0.2$
$\fallingdotseq \Pr_{\mathcal{D}}[\ \mathit{Weather} = \text{sunny}\ ]$

$\Pr_{\hat{\mathcal{D}}}[\ W = \text{windy}\ ] = 30/300 = 0.1 \fallingdotseq \Pr_{\mathcal{D}}[\ W = \text{windy}\ ]$

General prob. assumption

---

What a naive assumption *!*

$\Pr_{\mathcal{D}}[\ \mathit{Weather} = \text{sunny} \ \&\ W = \text{windy}\ ] \fallingdotseq 0.2 \times 0.1 = 0.02$

# 2. Naive Bayes

Naive Bayes is a method to obtain a probabilistic classification model.

For example, suppose we want to compute

$$\Pr[\, P = \text{no} \mid (\mathit{Weather}, T, H, W) = (\text{sunny}, 80, 90, \text{windy})\,]$$
$$= \Pr[\, P = \text{no} \mid \mathit{Weather} = \text{sunny} \;\&\; T = 80 \;\&\; H = 90 \;\&\; W = \text{windy}\,]$$

Note that we can compute, e.g.,

$$\Pr_{\mathcal{D}}[\, W = \text{windy} \mid P = \text{no}\,] \fallingdotseq \Pr_{\hat{\mathcal{D}}}[\, W = \text{windy} \mid P = \text{no}\,]$$
$$= \frac{15}{120} = 0.125.$$

# of no = 120
# of windy & no = 15

Thus, our task is to compute

$\Pr[\, P = \text{no} \mid \text{A} \;\&\; \text{B} \;\&\; \text{C} \;\&\; \text{D}\,]$ based on $\Pr[\, P = \text{no} \mid \text{A}\,]$, etc.

# 2. Naive Bayes

$$\Pr[H|E] = \frac{\Pr[E|H] \times \Pr[H]}{\Pr[E]}$$

N

$$\Pr[\,P = \text{no} \mid A \,\&\, B \,\&\, C\,]$$

$$= \frac{\Pr[\,N \mid A \,\&\, B \,\&\, C\,] \times \Pr[\,N\,]}{\Pr[\,A \,\&\, B \,\&\, C\,]} = \frac{\Pr[\,N \,\&\, A \,\&\, B \,\&\, C\,]}{\Pr[\,A \,\&\, B \,\&\, C\,]}$$

$$\Pr[\,A \mid B \,\&\, C \,\&\, N\,] \times \Pr[\,B \,\&\, C \,\&\, N\,]$$

$$= \Pr[\,A \mid B \,\&\, C \,\&\, N\,] \times \Pr[\,B \mid C \,\&\, N\,] \times \Pr[\,C \,\&\, N\,]$$

$$= \Pr[\,A \mid B \,\&\, C \,\&\, N\,] \times \Pr[\,B \mid C \,\&\, N\,] \times \Pr[\,C \mid N\,] \times \Pr[\,N\,]$$

by the assumption of independence

$$= \Pr[\,A \mid N\,] \times \Pr[\,B \mid N\,] \times \Pr[\,C \mid N\,] \times \Pr[\,N\,]$$

$$\Pr[\,B \mid C \,\&\, N] \times \Pr[\,N\,] = \Pr[\,B \,\&\, N \mid C\,] = \Pr[\,B \,\&\, N\,] = \Pr[\,B \mid N\,] \times \Pr[\,N\,]$$

# 2. Naive Bayes

Thus, we can compute

$$\Pr[\, P = \text{no} \mid (\mathit{Weather}, T, H, W) = (\text{sunny}, 80, 90, \text{windy})\,]$$

$$= \Pr[\, P = \text{no} \mid \mathit{Weather} = \text{sunny} \;\&\; T = 80 \;\&\; H = 90 \;\&\; W = \text{windy}\,]$$

$$\doteqdot \frac{\Pr[\mathit{Wth}=\text{sny} \mid P=\text{n}] \cdot \Pr[T=80 \mid P=\text{n}] \cdot \Pr[H=90 \mid P=\text{n}] \cdot \Pr[W=\text{wind} \mid P=\text{n}] \cdot \Pr[P=\text{n}]}{\Pr[\mathit{Wth}=\text{sny} \;\&\; T=80 \;\&\; H=90 \;\&\; W=\text{wind}\,]}$$

Also, we have

$$\Pr[\, P = \text{yes} \mid (\mathit{Weather}, T, H, W) = (\text{sunny}, 80, 90, \text{windy})\,]$$

$$\doteqdot \frac{\Pr[\mathit{Wth}=\text{sny} \mid P=\text{y}] \cdot \Pr[T=80 \mid P=\text{y}] \cdot \Pr[H=90 \mid P=\text{y}] \cdot \Pr[W=\text{wind} \mid P=\text{y}] \cdot \Pr[P=\text{y}]}{\Pr[\mathit{Wth}=\text{sny} \;\&\; T=80 \;\&\; H=90 \;\&\; W=\text{wind}\,]}$$

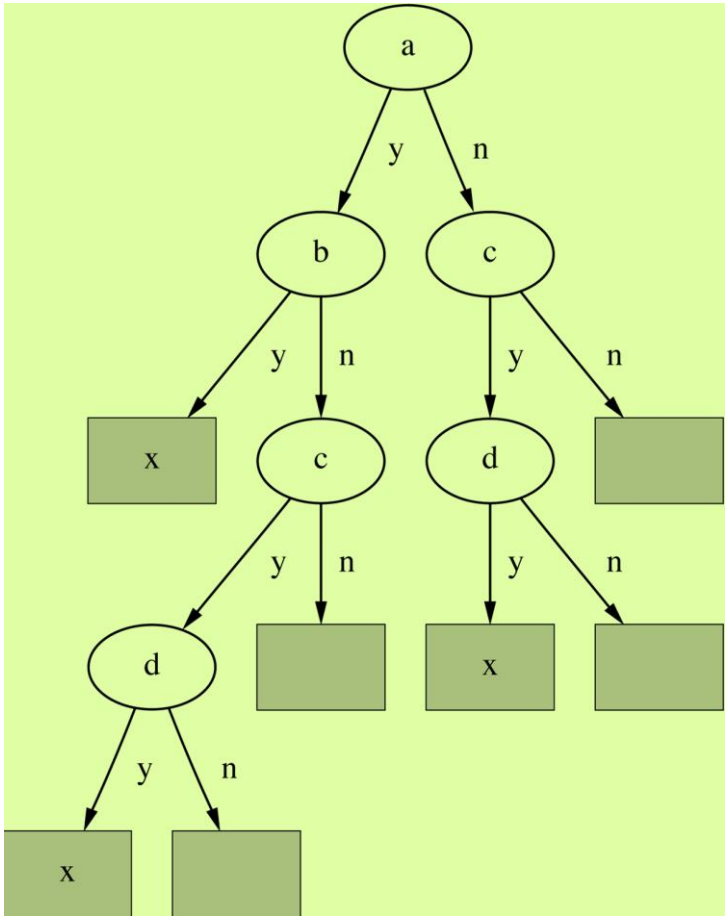
Note that we do not have to compute

$$\Pr[\mathit{Wth}=\text{sny} \;\&\; T=80 \;\&\; H=90 \;\&\; W=\text{wind}\,]$$

for predicting the class value.

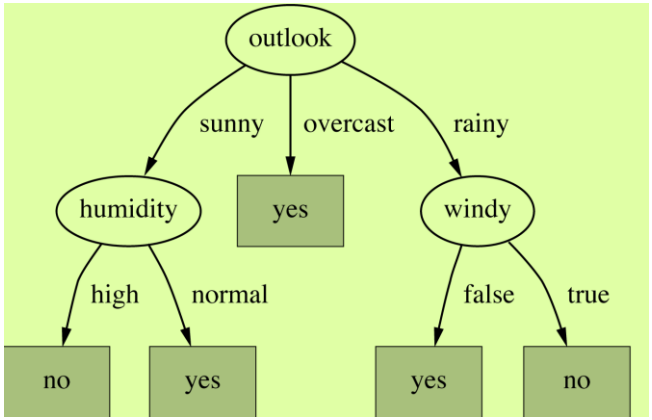
# 3. Decision tree

Decision tree is a classification model by using a "decision tree" like this..



- "Divide-and-conquer" approach produces tree
- Nodes involve testing a particular attribute
- Usually, attribute value is compared to constant
- Other possibilities:
  - Comparing values of two attributes
  - Using a function of one or more attributes
- Leaves assign classification, set of classifications, or probability distribution to instances
- Unknown instance

  is routed down the tree

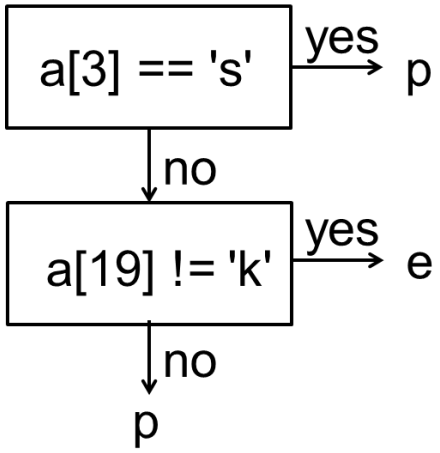


# 3. Decision tree

## 3.1. Some subclasses of decision trees

Decision stump = one level decision tree
(also called 1R in Weka)

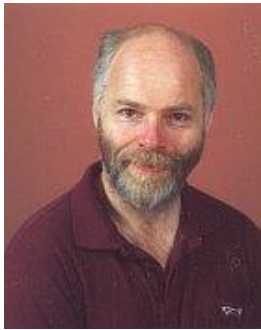
Decision list




- 1R was described in a paper by Holte (1993)
- Simplicity first pays off!

**Very Simple Classification Rules Perform Well on Most Commonly Used Datasets**

Robert C. Holte, Computer Science Department, University of Ottawa



# 3. Decision tree

## 3.2. How to construct decision trees (only intro.)

Here we discuss the very simple one for constructing a decision stump or 1R. (We will discuss an important ML algorithm (which is a basis of J4.8 in Weka) later in this course.)

```
For each attribute,
   For each value of the attribute, make a rule as follows:
      count how often each class appears
      find the most frequent class
      make the rule assign that class to this attribute-value
   Calculate the error rate of the rules
Choose the rules with the smallest error rate
```

# Example:

## Evaluating the weather attributes

| Outlook | Temp | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | High | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Sunny | Mild | High | False | No |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| Sunny | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Rainy | Mild | High | True | No |

| Attribute | Rules | Errors | Total errors |
|---|---|---|---|
| Outlook | Sunny → No | 2/5 | 4/14 |
| | Overcast → Yes | 0/4 | |
| | Rainy → Yes | 2/5 | |
| Temp | Hot → No* | 2/4 | 5/14 |
| | Mild → Yes | 2/6 | |
| | Cool → Yes | 1/4 | |
| Humidity | High → No | 3/7 | 4/14 |
| | Normal → Yes | 1/7 | |
| Windy | False → Yes | 2/8 | 5/14 |
| | True → No* | 3/6 | |

* indicates a tie

# 4. On Exercise #2

Classification rule discovery project:
- Introduction to Weka*1.
- Use 1R, Naive Bayes, and Decision Tree methods.

## Task #1: Try to use Weka

(a) Prepare m8126.arff.
(b) Compute classification models on m8126.arff.

## Task #2: Understand basics on classification tasks and decision trees

(c) Use breast-cancer.arff (given as a sample data in Weka) to learn how to create an appropriate decision tree.

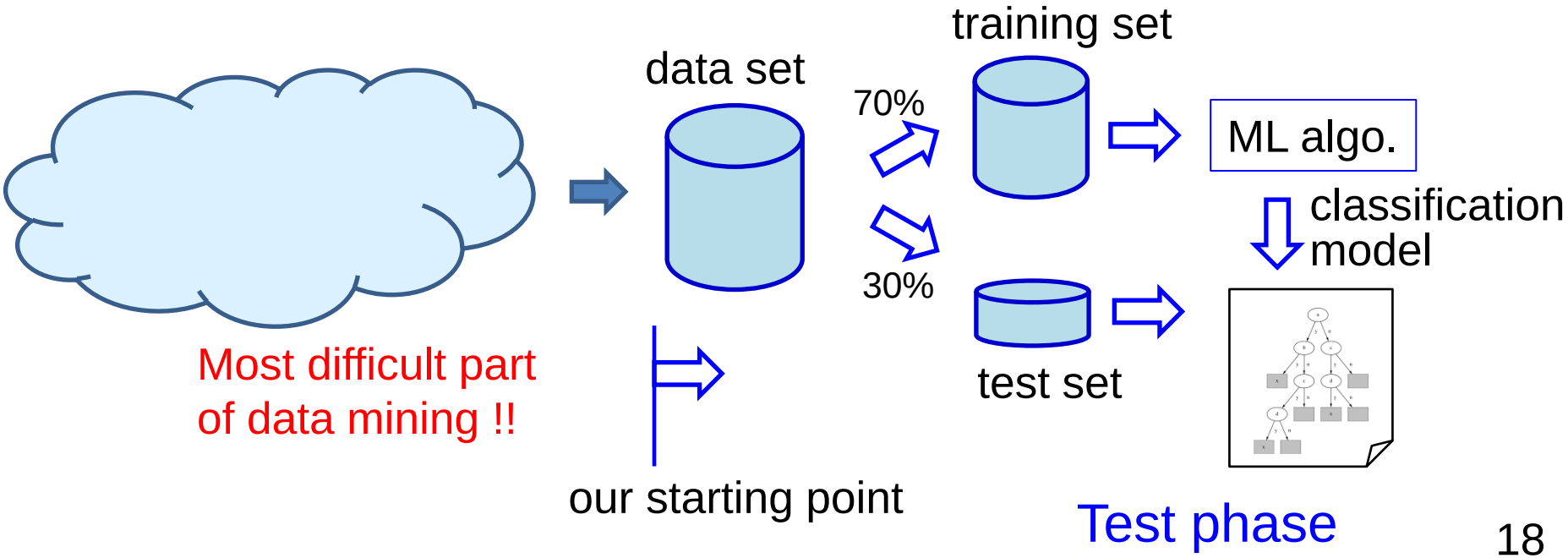
# 4. On Exercise #2

## Intro. to classification task (cont.)

### Task #2: Understand basics on classification tasks and decision trees

(c) Use breast-cancer.arff (given as a sample data in Weka) to learn how to create an appropriate decision tree.

A standard flow of classification tasks



# 4. On Exercise #2

## Intro. to classification task (cont.)

Important Points:

1. The performance of the obtained model should be tested on a separate data, i.e., test set.

e.g., a decision tree with 300 leaves for 300 instances

2. Avoid overfitting. It is rather easy to obtain a model that is good (or almost perfect) on a training set, but it is not useful in general.

The simpler is the better.

Occam's Razor

Plurality must never be posited without necessity