Ex1: Mini. Data Mining

Tokyo Tech.
Intro. to Comp. & Data
Exercise&hw week1

Try some mini. data mining project:
Classification rule discovery of poisonous mushrooms

1. Data description and our goal.

2. Homework assignment #1.

   * submit through ~~OCW~~ *before* week2lect *of each campus*

   please send one pdf file via email to
   Suzukakedai: watanabe.o.aa-cd18s@ml.m.titech.ac.jp
   Ookayama: watanabe.o.aa-cd18o@ml.m.titech.ac.jp

3. Our tools (i.e., python programs) and dataset.

   for ex1 materials, see
   http://tcs.c.titech.ac.jp/DataMining/index.html

1. Data description and our goal

Discover the rule for classifying poisonous mushrooms

From the mushroom characteristics (attributes), discover a rule (binary decision rule) for determining whether the mushroom is poisonous or not.

Cap surface

Size

Cap shape

About 20 attributes

Smell

Spots

General shape

Cap color

General color

Stem color

Gills

data or dataset

||

sample = collection of examples (or instances)

||

a tuple of attribute values + class value

class value

p: poisonous
e: edible

22 attribute values

```
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
                    :
```

cap color

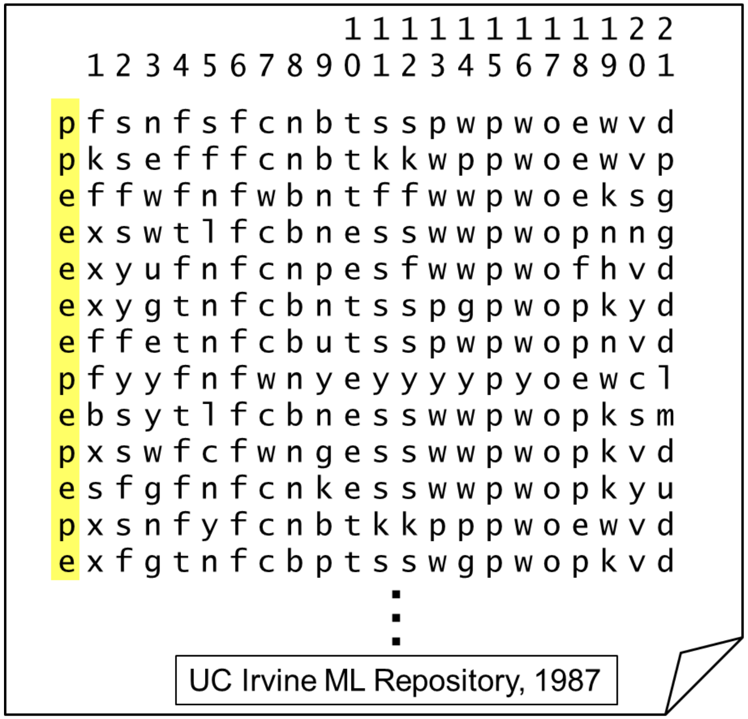
y: yellow
p: pink
b: brown
⋯

8000+ mushroom instances

Use only these data!

UC Irvine ML Repository, 1987

goal = to find a *good* rule for detecting the class value of a given mushroom instance



```
                 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
                   ⋮
```

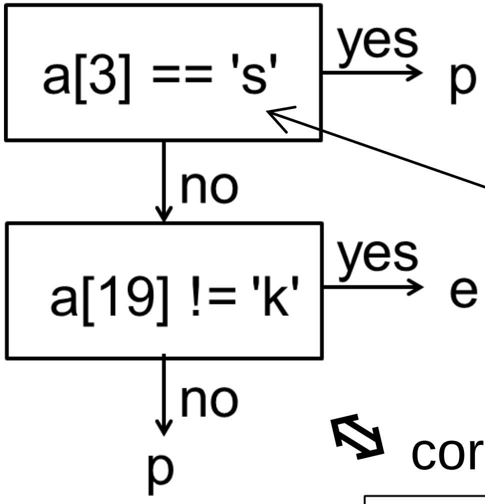
UC Irvine ML Repository, 1987

rule = a Boolean expression that determines a given mushroom's toxicity

more specifically

rule = a decision list

e.g.,



each base predicate is to ask whether a[*k*] == *val* or not

use some appropriate # of base predicates

⇔ corresponding Boolean expression

a[3] == 's' ∨ (a[3] != 's' ∧ a[19] == 'k')

Good rule = a decision list with low error rate

Accuracy =
(or, success rate)

$$\frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

Error rate =

$$\frac{\text{Number of } \textit{in}\text{correctly classified instances}}{\text{Total number of instances}}$$

False positive = Incorrectly classified as positive

False negative = Incorrectly classified as negative

in our mushroom data, let us call

p: poisonous = +1, positive
e: edible = -1, negative

Better to avoid false negative!

for our mushroom classification task

2. Homework assignment #1: Task

training set

Your task:

(a) Obtain a decision list using 2000 instances of the mushroom data (m8124org.txt) with accuracy > 90% on the whole dataset.

page 7

demo. in the ex. session

* Use only the provided python programs.
* You may modify these programs *as you like!!*

(b) Understand the mechanism of the provided python programs.

2. Homework assignment #1: Report

submit through OCW *before* week2lect

Required items that you need to explain:

Japanese is OK!!
About 1 page for each item, please!

(1) + a decision list that you obtained,

+ its corresponding Boolean expression (used in test.py), and
+ its statistical data, that is,
  + accuracy, true positive rate, true negative rate
    both on the training set and on the whole data set.

$$\text{Ture positive rate} = \frac{\text{Number of correctly classified positive instances}}{\text{Total number of positive instances}}$$

$$\text{Ture negative rate} = \frac{\text{Number of correctly detected negative instances}}{\text{Total number of negative instances}}$$

2. Homework assignment #1: Report (Cont.)

Required items that you need to explain: (Cont.)

(2) Explain a way to obtain your decision list at.

- * The outline of what you did (or what your program did) for obtaining your decision list.

(3) Explain a key program (e.g., count.py) that you used.

- + explanation of the program outline, and
- + the source code of the program with explanation on what is computed at each key statement.

↑ hand written comments are enough!!

3. Our tools and dataset

http://tcs.c.titech.ac.jp/DataMining/index.html

Dataset: from UCI repository

https://archive.ics.uci.edu/ml/datasets/mushroom

- m8124org.txt: 8124 mushroom instances
- mushroom-spec.txt: explanation on this dataset

Tools: simple python programs

- shuffle.py: permute lines (i.e., instances) randomly
- test.py: test accuracy of a Boolean expression
- count.py: count # of instances for a basic predicate
- select.py: select instances not satisfying a basic predicate

Usage of these tools

demo. in the ex. session

```
C:> python xxx.py number of instances < input file > output file
```

Example: selection of 2000 instances for a training set

```
C:> python shuffle.py  8124 < m8124org.txt > m8124rnd.txt
C:> head -2000 < m8124rnd.txt > m2Krnd.txt
```

References

Some Terminal commands

Command	Example	Meaning
mkdir	mkdir ex1	Create a folder ex1
cd	cd ex1	Move to the ex1 folder
	cd **..**	Move to the parent folder
	cd **../..**	Move to the parent of the parent folder
dir	dir	Display files of the current folder
rm	rm foo.py	Delete foo.py (It is impossible to undo this command!)
python	Python foo.py	Run a program in the machine code