Codage attributs-valeurs

Fabien Torre

Université de Lille

Mercredi 30 septembre 2009

Codage attributs-valeurs

Couple attribut-valeur

- Attribut : un nom et un type :
  - des attributs continus ;
  - des attributs discrets ;
- valeur : ... une valeur !

Exemples attributs-valeurs

- Un exemple est un vecteur de taille fixée pour le problème ;
- les exemples d'un problème ont le même nombre d'attributs ;
- il peut y avoir des valeurs manquantes ;
- une valeur manquante est souvent symbolisée par un point d'interrogation.

Format pratique

- Le format C4.5 : trois fichiers (`.names`, `.data`, `.test`) ;
- le repository UCI [Blake and Merz, 1998] propose des problèmes variés en attributs-valeurs ;
- du code est disponible pour charger ces fichiers (Perl ou Java).

Fichier agaricus-lepiota.txt I

```
1. Title: Mushroom Database
2. Sources:
    Mushroom records drawn from The Audubon Society Field Guide t
    American Mushrooms (1981). G. H. Lincoff (Pres.), New York: A
3. Past Usage:
    Schlimmer,J.S. (1987). Concept Acquisition Through Representa
    Adjustment (Technical Report 87-19).  Doctoral disseration, D
    of Information and Computer Science, University of California
    Disjunctive rules for poisonous mushrooms:
      P_1) odor=NOT(almond.OR.anise.OR.none)
           120 poisonous cases missed, 98.52% accuracy
      P_2) spore-print-color=green
           48 cases missed, 99.41% accuracy
```

Fichier agaricus-lepiota.txt II

```
      P_3) odor=none.AND.stalk-surface-below-ring=scaly.AND.(stalk-col
           8 cases missed, 99.90% accuracy
      P_4) habitat=leaves.AND.cap-color=white, 100% accuracy
4. Relevant Information:
   This data set includes descriptions of hypothetical samples
   corresponding to 23 species of gilled mushrooms in the Agaricus an
   Lepiota Family (pp. 500-525).  Each species is identified as
   edible or poisonous.
5. Number of Instances: 8124
6. Number of Attributes: 22 (all nominally valued)
7. Attribute Information: (classes: edible=e, poisonous=p)
   1.cap-shape:                bell=b,conical=c,convex=x,flat=f,
                               knobbed=k,sunken=s
   2.cap-surface:              fibrous=f,grooves=g,scaly=y,smooth=s
```


Fichier agaricus-lepiota.txt III

```
   3.cap-color:                brown=n,buff=b,cinnamon=c,gray=g,gr
                               pink=p,purple=u,red=e,white=w,yello
   4.bruises?:                 bruises=t,no=f
   5.odor:                     almond=a,anise=l,creosote=c,fishy=y
                               musty=m,none=n,pungent=p,spicy=s
   6.gill-attachment:          attached=a,descending=d,free=f,notc
   7.gill-spacing:             close=c,crowded=w,distant=d
   8.gill-size:                broad=b,narrow=n
   9.gill-color:               black=k,brown=n,buff=b,chocolate=h,
                               green=r,orange=o,pink=p,purple=u,re
                               white=w,yellow=y
  10.stalk-shape:              enlarging=e,tapering=t
  11.stalk-root:               bulbous=b,club=c,cup=u,equal=e,
                               rhizomorphs=z,rooted=r,missing=?
```


Fichier agaricus-lepiota.txt IV

```
  12.stalk-surface-above-ring:ibrous=f,scaly=y,silky=k,smooth=s
  13.stalk-surface-below-ring:ibrous=f,scaly=y,silky=k,smooth=s
  14.stalk-color-above-ring:  brown=n,buff=b,cinnamon=c,gray=g,orange=
                              pink=p,red=e,white=w,yellow=y
  15.stalk-color-below-ring:  brown=n,buff=b,cinnamon=c,gray=g,orange=
                              pink=p,red=e,white=w,yellow=y
  16.veil-type:               partial=p,universal=u
  17.veil-color:              brown=n,orange=o,white=w,yellow=y
  18.ring-number:             none=n,one=o,two=t
  19.ring-type:               cobwebby=c,evanescent=e,flaring=f,large=
                              none=n,pendant=p,sheathing=s,zone=z
  20.spore-print-color:       black=k,brown=n,buff=b,chocolate=h,green=
                              orange=o,purple=u,white=w,yellow=y
  21.population:              abundant=a,clustered=c,numerous=n,
```


Fichier agaricus-lepiota.txt V

```
                              scattered=s,several=v,solitary=y
  22.habitat:                 grasses=g,leaves=l,meadows=m,paths=
                              urban=u,waste=w,woods=d
8. Missing Attribute Values: 2480 of them (denoted by "?"), all f
   attribute #11.
9. Class Distribution:
    --    edible: 4208 (51.8%)
    -- poisonous: 3916 (48.2%)
    --     total: 8124 instances
```

Fichier agaricus-lepiota.names

```
e,p.

cap-shape: b,c,x,f,k,s.
cap-surface: f,g,y,s.
cap-color: n,b,c,g,r,p,u,e,w,y.
bruises?: t,f.
odor: a,l,c,y,f,m,n,p,s.
gill-attachment: a,d,f,n.
gill-spacing: c,w,d.
gill-size: b,n.
gill-color: k,n,b,h,g,r,o,p,u,e,w,y.
stalk-shape: e,t.
stalk-root: b,c,u,e,z,r.
stalk-surface-above-ring: f,y,k,s.
stalk-surface-below-ring: f,y,k,s.
stalk-color-above-ring: n,b,c,g,o,p,e,w,y.
stalk-color-below-ring: n,b,c,g,o,p,e,w,y.
veil-type: p,u.
```


Fichier agaricus-lepiota.data I

```
x,s,n,t,p,f,c,n,k,e,e,s,s,w,w,p,w,o,p,k,s,u,p
x,s,y,t,a,f,c,b,k,e,c,s,s,w,w,p,w,o,p,n,n,g,e
b,s,w,t,l,f,c,b,n,e,c,s,s,w,w,p,w,o,p,n,n,m,e
x,y,w,t,p,f,c,n,n,e,e,s,s,w,w,p,w,o,p,k,s,u,p
x,s,g,f,n,f,w,b,k,t,e,s,s,w,w,p,w,o,e,n,a,g,e
x,y,y,t,a,f,c,b,n,e,c,s,s,w,w,p,w,o,p,k,n,g,e
b,s,w,t,a,f,c,b,g,e,c,s,s,w,w,p,w,o,p,k,n,m,e
b,y,w,t,l,f,c,b,n,e,c,s,s,w,w,p,w,o,p,n,s,m,e
x,y,w,t,p,f,c,n,p,e,e,s,s,w,w,p,w,o,p,k,v,g,p
b,s,y,t,a,f,c,b,g,e,c,s,s,w,w,p,w,o,p,k,s,m,e
x,y,y,t,l,f,c,b,g,e,c,s,s,w,w,p,w,o,p,n,n,g,e
...
```


Fichier iris.txt I

```
1. Title: Iris Plants Database
2. Sources:
     (a) Creator: R.A. Fisher
     (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
     (c) Date: July, 1988
3. Past Usage:
   1. Fisher,R.A. "The use of multiple measurements in taxonomic proble
      Annual Eugenics, 7, Part II, 179-188 (1936).
   2. Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene A
      (Q327.D83) John Wiley & Sons.  ISBN 0-471-22361-1.  See page 218
4. Relevant Information:
   --- The data set contains 3 classes of 50 instances each,
       where each class refers to a type of iris plant.
```


Fichier iris.txt II

```
   --- Predicted attribute: class of iris plant.
5. Number of Instances: 150 (50 in each of three classes)
6. Number of Attributes: 4 numeric, predictive attributes and the
7. Attribute Information:
   1. sepal length in cm
   2. sepal width in cm
   3. petal length in cm
   4. petal width in cm
   5. class:
      -- Iris Setosa
      -- Iris Versicolour
      -- Iris Virginica
8. Missing Attribute Values: None
9. Class Distribution: 33.3% for each of 3 classes.
```

Fichier iris.names I

```
Iris-setosa, Iris-versicolor, Iris-virginica.

SepalLength: continuous.
SepalWidth: continuous.
PetalLength: continuous.
PetalWidth: continuous.
```


Fichier iris.data I

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
```


Fichier iris.data II

```
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
```


Méthodes

- Bayes naïf ;
- arbres de décisions (séance 2) ;
- moindres généralisés (séance 3) ;
- réseaux de neurones ;
- *Support Vector Machines* (SVM) ;
- etc.

Bibliographie I


Blake, C. and Merz, C. (1998).
UCI repository of machine learning databases
[http ://archive.ics.uci.edu/ml/].