**CS3907-80/CS6444-10 Big Data and Analytics**

**Fall 2017**

**Class Project #2**

Exploring Variations in Clustering and Predictive Analysis

1. Data Set: Mushroom (on Blackboard)

These data sets are described in agaricus-lepiota.data and agaricus-lepiota.names

**Step 0:**

*Objective*: There are two classes: edibles and poisonous. But, there are some mushrooms (#TBD) which are uncertain, but are classified as poisonous. Your job is to determine of which the poisonous class are actually the uncertain category by applying clustering, classification, and prediction techniques discussed in class as well as additional functions from the packages mentioned.

You will need to divide your data set into a training set and a test set. Use samples of 50-50, 60-40, and 70-30 for the training-test ratios

**Step 1: Azeez**

Try plotting the data using several plotting functions to see what it looks like. Use pairs (e.g., 2D plots) or 3 variables (3D plots) based on the packages.

Try to filter the data by selecting samples with only certain attribute values and plotting them.

**Step 2: Pairs and check data (Reduce manually or PCA) Aashti & Joy**

You should try data reduction to eliminate some attributes through Principal Components Analysis. The idea is to try and select N attributes that will help you focus on identifying the unsure samples.

**Step 3: Clustering and Analysis. Azeez & Joy & (All)**

We discussed a number of techniques in lectures 4 and 5, but you can use other techniques from the contributed R packages.

2. This will involve some statistical analysis and some clustering. Use the R packages and functions in the notes as well as the ones below.

3. Deliverables: You will deliver your results by putting a zipfile in your group’s Blackboard file, with the following naming convention: Group-N-Project-2.zip, where N is your group number. Your deliverable should encompass the following items:

A listing of all R functions that you have written

* A document giving your results which should include:

**Step 3 cont.**

a. A description of the edible, poisonous classes and the uncertain category, respectively, based on the features using three different clustering methods such as **kmeans, k-nearest neighbor, or one other in the R contributed packages.** Clearly identify which methods you are using.

Note: The first column in the data set has e for edible and p for poisonous. All well and good but that does not tell you what features distinguish them. That is what you have to determine.

b. a clustering of the samples into N = 3, 5, 7 classes using the three different clustering methods. The idea is to see how the clustering method and its underlying assumptions changes your perspective on the data.

c. prepare a table containing the data from (a & b) with the three training-test ratios for each of the N and each clustering method

d. plots using several methods in lectures 4 and 5.

You should investigate some of the statistics of the data set

**Step 4: // Emily**

4. Try the **lm and gl**m methods to get linear fits for the data. This will not work on all attributes, so you must determine which ones it will work on. Note as discussed in class binomial (logit) expects two categories, so you might combine the two data sets into one and determine if you can distinguish between and how good the fit it.

**Step 5 : SVM & Aashti and Aakash**

5. Use SVM as well to try to separate the combined data set into two separate classes.

See the slides for book chapters to read to help you. Try different methods in glm, build a table and record the relevant data. What can you determine from the table of values?

*Approach*:

You should determine what to do and divide the work up among the team members after creating the different training and test subsets.

Remember to save your workspace! In your Group area would be a good place so all members can get to it.

Include in your Word document the results required

(use a CTRL-ALT-PrintScreen) to grab the screen

You may use Irfanview 4.40, irfanview@gmx.net. Paste in the screen image, and copy the image as JPEG to drop into your Word document.

6. Project #2 Value: 15 points

a. Document R functions that you write: 2 points

b. Table with results as specified in 3 above: 3 points

c. Discussion of results from the experiments that you run in part 3b above. Which clustering version gives the best results? What seems to be the best number of clusters for each method? --- 3 points

d. Table from 4 above – 3 points

e. Plots with discussion of results: which plot helped you understand the data best and why? – 3 points

f. Analysis of what this project helped you learn about data science, e.g., the exploration of data which is what you have been doing – 1 point. You must argue persuasively.

Review the documentation for the packages and functions that you use.

Project #2:

Some hints based on processing mushroom data which I have used in the past.

Read in the data as follows:

mr<-read.table(file.path("h:", "…/mushroom/agaricus-lepiota.data"),sep=",", header=FALSE)

> str(MRData)

'data.frame': 8124 obs. of 23 variables:

$ V1 : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...

$ V2 : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...

$ V3 : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...

$ V4 : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10 ...

$ V5 : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...

$ V6 : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...

$ V7 : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...

$ V8 : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...

$ V9 : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...

$ V10: Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3 6 8 3 ...

$ V11: Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...

$ V12: Factor w/ 5 levels "?","b","c","e",..: 4 3 3 4 4 3 3 3 4 3 ...

$ V13: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...

$ V14: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...

$ V15: Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...

$ V16: Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...

$ V17: Factor w/ 1 level "p": 1 1 1 1 1 1 1 1 1 1 ...

$ V18: Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3 3 ...

$ V19: Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 2 ...

$ V20: Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5 5 5 5 ...

$ V21: Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3 4 3 3 ...

$ V22: Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...

$ V23: Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...

Convert to a data frame

mr.df<- as.data.frame(MRData)

> mr.df

class cshape csurface ccolor bruises odor gattach gspace gsize gcolor sshape sroot

1 p x s n t p f c n k e e

2 e x s y t a f c b k e c

3 e b s w t l f c b n e c

4 p x y w t p f c n n e e

5 e x s g f n f w b k t e

6 e x y y t a f c b n e c

7 e b s w t a f c b g e c

8 e b y w t l f c b n e c

9 p x y w t p f c n p e e

10 e b s y t a f c b g e c

11 e x y y t l f c b g e c

Rest of listing has been omitted

**Attribute Information:**

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

3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y

4. bruises?: bruises=t,no=f

5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s

6. gill-attachment: attached=a,descending=d,free=f,notched=n

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,gray=g, green=r,orange=o,pink=p,purple=u,red=e, 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=?

12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s

13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s

14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y

15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, 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=l, none=n,pendant=p,sheathing=s,zone=z

20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y

21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y

22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Attribute Names have been somewhat abbreviated

>names(mr)<-c(“class”, "cshape", "csurface", "ccolor", "bruises", "odor", "gattach", "gspace", "gsize", "gcolor", "sshape", "sroot", "ssabove", "ssbelow", "scabove", "scbelow", "vtype", "vcolor", "rnumber", "rtype", "spcolor", "popnum", "habitat")

> names(mr)

[1] "class" "cshape" "csurface" "ccolor" "bruises" "odor" "gattach" "gspace"

[9] "gsize" "gcolor" "sshape" "sroot" "ssabove" "ssbelow" "scabove" "scbelow"

[17] "vtype" "vcolor" "rnumber" "rtype" "spcolor" "popnum" "habitat"

Plotting the data, you get a very hard to read set of plots.

The file MRDataPlot.jpg shows an expanded view.

So, what we can do is plot pairs of attributes to see what we can glean, then construct a formula from the interesting ones and plot that as well.

>pairs(class ~ cshape + csurface + ccolor, data=mr)

What do we learn from this?

1. We already knew there were just two classes and this reinforces that notion with the plots of class against the other factors in the top row.

2. For cshape, we see linearity - either vertically or horizontally, which tells us the values of shape are closely correlated with the surface. We didn’t know this before, but might have guessed it.

3. Further, it appears that surface and color are also closely correlated for the discrete values.

Now:

plot(class ~ sshape + sroot, data=mr)

Here is the plot for class vs sroot:

So, what we see here is the following:

stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?

So, sroot having the attribute club, equal, or rhizomorphs characterize both edible and poisonous whereas bulbous only characterizes edible.

So, this could lead to a rule that says: if the sroot is bulbous, it is edible. BUT, that is not the only rule!

THUS: You should consider doing pairwise comparisons and seeing what conclusion you can draw.

Now, you can setup two data sets: mrtrain.df and mrtest.df using the methods we discussed in lecture 4.

In fact, as required by project #2, you have to have three pairs of these sets with the ratios 50-50, 60-40, and 70-30.

Then, you need to convert letter values to numbers to support the clustering analysis.

So, the 70-30 training set is:

>train.df <- sample(nrow(mr), 0.7\*nrow(mr))

Yields (for me) 5686 records.

The training set is

> mrtrain.df <- mr[train.df,]

The test set is:

> mrtest.df <- mr[-train.df,]

Which yields (for me) 2438 records

> table(mrtrain.df$class)

e p

2925 2761

> table(mrtest.df$class)

e p

1283 1155

An almost equitable distribution of edible versus poisonous.

Converting letters to values within the data set:

The first thing to note is that the read functions seem to return factors. You need to convert the values of the attributes to characters.

So, do the following:

mrf<-as.data.frame(mr, stringAsFactors=FALSE)

> mrf[,c(1,23)]<-sapply(mrf[,c(1,23)],as.character)

> mrf$class[mrf$class=='e'] <- 0

> mrf

class cshape csurface ccolor bruises odor gattach gspace gsize gcolor sshape sroot

1 p x s n t p f c n k e e

2 0 x s y t a f c b k e c

3 0 b s w t l f c b n e c

4 p x y w t p f c n n e e

5 0 x s g f n f w b k t e

6 0 x y y t a f c b n e c

7 0 b s w t a f c b g e c

8 0 b y w t l f c b n e c

9 p x y w t p f c n p e e

10 0 b s y t a f c b g e c

Notice that all the ‘e’s in the class column were made 0.

Similarly, for the ‘p’s in the class column.

Consider the first attribute:

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

You might consider substituting b =1, c =3, x = 5, f = 7, k=9, s =11

Note that in the clustering algorithms, the comparison is between records, not between attributes. Thus, the distance measures take into account all of the factors in each record, assuming you have omitted the class.