Homework 6: Multivariate Data Mining

Updated March 2016

Table of Contents

Multivariate Data Mining

Data mining is the computational process of discovering patterns in large data sets involving machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Mining includes data pre-processing, model and inference considerations, statistical characterization, post-processing of discovered structures, and visualization.

One key method of data mining is “data reduction” such as principal components analysis or PCA. In this homework, we will explore a large dataset and reduce the data via record subsetting, and variable selection, and reduction via PCA. Further, we will use clustering to aggregate observations into like groups, and then test whether those types conform with other parameter designations. Both data manipulation and data visualization are key components of data mining, so make sure that you include diagnostic plots of your work.

Homework Exercise (3 Points total)

These exercises rely on the lab data found in the files section of CATCOURSES, `meadows.dbf`. This file has over 17,000 records. We are going to create a subset of these data for this exercise.

Your document should have the following sections, and provide written explanations formatted in RMarkdown that explains your code, output and graphics in the following format:

NAME
CLASS
DATE

Homework Assignment 6

Objective Statement: [What are you trying to accomplish?]

Methods: [In general terms, what analyses are you doing?]
Data: [What are the data and where did they come from?]
Code: [In specific terms, what is the code that was used to conduct the analysis?]

Results: [What do the results show? Numerical evidence and graphic evidence are required.]

Discussion: [What do the results mean?]

Limitations: [What are the limitations, caveats, and assumptions of the analysis?]

OBJECTIVE

You have been asked to validate the independent "hydrogeomorphic type" assigned by the US Forest Service to a select set of meadows in the Sierra Nevada. Read through the General Technical Report which was provided to you to determine which variables will be important to statistically designate "HGM" types.

Your deliverable will be: an HGM validation using cluster analysis

Step 1 - EDA and scatter-plot matrices

Overview: Read in data using `foreign` package and `read.dbf()` function. With the imported data set, subset to records having "HGM" values; explore relevant parameters using scatterplots.

Detailed Methods:
Install and load the `foreign` package and use the `read.dbf()` function to load the data "`Sierra_Nevada_MultiSource_Meadow_Polygons_Compilation_v1.dbf`". Look over the metadata (`SNMMPCv1_Metadata.txt`) and view the structure of the dataframe to become familiar with the dataset. Notice that R has limitations on the column name length, and some attribute names have been truncated. Reduce the data records to only those that have values for the Hydrogeomorphic type ("HGM_TYPE") using the following:

```
mdwhgm<- mdws[!is.na(mdws[,"HGM_TYPE"]),]
```

[Quiz Question: What does `!is.na` mean in words?]

Add new columns with names that are more appropriate (and more intuitive) using the suggestions below (where `mdwhgm` is the dataframe having only those with HGM designations):

```
mdwhgm$area.sqkm = mdwhgm[,"Shape_Area"]/1000000 # m^2 to km^2
mdwhgm$catch.sqkm = mdwhgm[,"CATCHMENT_"]/1000000# m^2 to km^2
mdwhgm$elev_m = mdwhgm[,"ELEV_MEAN"]
mdwhgm$elev_r = mdwhgm[,"ELEV_RANGE"]
mdwhgm$lat_dd = mdwhgm[,"LAT_DD"]
mdwhgm$lon_dd = mdwhgm[,"LONG_DD"]
mdwhgm$slope.pct = mdwhgm[,"FLOW_SLOPE"]
mdwhgm$edge.comp = mdwhgm[,"EDGE_COMPL"]
mdwhgm$clay = mdwhgm[,"ClayTot_r"]
mdwhgm$soil.kf = mdwhgm[,"Kf"]
```

After reading the metadata, are there other columns that should be included for further analysis?

Perform exploratory data analysis to determine which variables are relevant to this project. This is actually not so trivial as it usually requires you to be very familiar with the data, so after EDA write up, you may just accept the new columns suggested above as the most relevant - You can keep track of them using the column indices or loading their column names in a list. For example:

```
#Optional method for keeping track of the relevant variables
rel_cols = c("area.sqkm", "catch.sqkm", "elev_m", "elev_r", "lat_dd",
"lon_dd", "slope.pct", "edge.comp", "clay", "soil.kf")
```

Create a scatterplot matrix of the relevant variables (Hint: Due to the number of variables in the scatter plot (~10) it may help to manually set the size of the figures using the codeblock options in r markdown. On the top of the codeblock, try something like `{r, fig.width=9, fig.height=9}`). Based on a visual inspection, which variables seem to have the greatest variability? Which variables appear to be correlated with each other? Hint, use `pairs()` to see these scatterplots.

Step 2 - Clustering and Clustering Output

Overview: Cluster using the relevant variables using hierarchical (`hclust()`) and k-means (`kmeans()`) clustering.

Hierarchical Clustering:

For hierarchical clustering, first find the distances of each variable using `dist()` then run `hclust`:

```
#dist using euclidean
mdwhgm.dist<- dist(x = mdwhgm[,rel_cols],method = "euclidean")
#hclust using ward.D
mdwhgm.hc<- hclust(mdwhgm.dist,method="ward.D")
```

Plot the results from hclust and draw 6 rectangles around the hierarchical clusters using the `rect.hclust()` (where `tree` is your cluster and `k` is 6. See `?rect.hclust`). Notice the different cluster groups.

Use `cutree()` to store cluster group identity to your dataframe:

```
mdwhgm$hc6 <- cutree(mdwhgm.hc, k=6) #store group # in hc6
```

k-means Clustering

For k-means clustering, simply run k-means using 6 centers creating the following object:

```
mdwhgm.km6 <- kmeans(mdwhgm[,rel_cols],centers = 6)
```

Now add the k-mean cluster group identity to your dataframe:

```
mdwhgm$km6 <- mdwhgm.km6$cluster #store group # in km6
```

Compare the results from each cluster method for consistency:

```
table(mdwhgm$hc6, mdwhgm$km6)
```

Discuss the output of this table. Make 2 plots using the `DEM.tif` as a background (Remember the `raster` and `gdal` packages? Also, here it may help to display the raster in grayscale by adding the following color parameter: `col=gray.colors(10, start=0.9, end=0.3)`). Add the points from the dataframe and color based on cluster class (`hc6` for the first plot and `km6` for the second plot).

Step 3 - Principal Components Analysis PCA

Overview:

Run a principal components analysis (PCA) on the meadows data using the `prcomp()` function. View the results and determine which have the greatest variability.

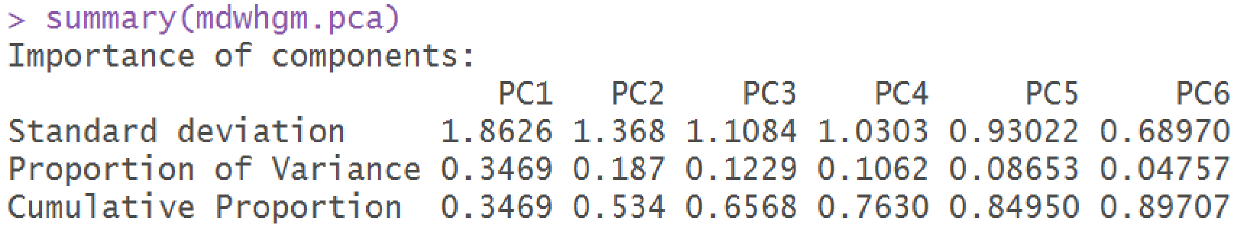
Detailed Methods:

Run the principal component analysis on the dataframe:

```
mdwhgm.pca <- prcomp(x = mdwhgm[,rel_cols], scale=TRUE, retx = TRUE,
center = TRUE, scores=TRUE)
```

View and discuss the `summary()` of the PCA, plus the visual diagnostic plots (e.g., `screeplot()`). Which are the 2 most important components of the PCA? An alternative to the `screeplot()` is `plot(pca, type="lines")`.

Note: typically users are looking to explain 75-90% of the variance in the first few components. In the case below, the first 5-6 components appear helpful.



```
> summary(mdwhgm.pca)
Importance of components:
                          PC1   PC2    PC3    PC4     PC5     PC6
Standard deviation     1.8626 1.368 1.1084 1.0303 0.93022 0.68970
Proportion of Variance 0.3469 0.187 0.1229 0.1062 0.08653 0.04757
Cumulative Proportion  0.3469 0.534 0.6568 0.7630 0.84950 0.89707
```

Next, evaluate the loadings to determine which of the parameters have the greatest influence on the PCA axes. For example, `print(mdwhgm.pca$rotation)`. These are typically called loadings. Which parameters are driving the variability in the meadow dataset (i.e., highest value)? Are these positive or negative loadings?

Use the `biplot()` function to examine how the loadings relate to the axes.
An example is below where `choices=i:j` reflect different axes of the PCA. (Use the `xlim` and `ylim` delimiters to rescale your axes as needed). Cycle through the most important axes.

```
biplot(mdwhgm.pca, choices=1:2, cex=0.5, xlim=c(-0.1,0.2), ylim=c(-
0.1,0.2))
```

How is this different from using `pairs()` on the PCA axes?

```
pairs(mdwhgm.pca$x[,1:3],col=mdwhgm$km6)#colored by Kmeans group
pairs(mdwhgm.pca$x[,1:3],col=mdwhgm$hc6)#colored by HClust group
```

Step 4 -- Contingency Analysis of Hydrogeomorphic Type

Perform a contingency analysis to examine how the cluster groupings compare to the designated USFS HGM Type from the field observations. Does there appear to be relationship based on counts? Is there a statistical relationship? [Hint, use the `chisq.test()` with pairings enumerated from `table()`.] Could you refactor the HGM_TYPE values to six overall HGM groups and get a better result? What do the other nominal variables, such as dominant rock type and dominant vegetation type, look like in comparison to your clusters?

Step 5 -- Summarize the Data by National Forest

Summarize a reported value of total acres by your meadow cluster type by National Forest (see [OWNERSHIP] column). Could managing for climate change vulnerability by meadow type be simplified for each National Forest focusing on one meadow type? In other words, is there any consistency in those cross-tabulation data? Rely on the assigned reading to help form your answer.

Resources

Data

- `Sierra_Nevada_MultiSource_Meadow_Polygons_Compilation_v1.dbf` on UCMCROPS (FILES→HOMEWORK→HOMEWORK 6)
- `SNMMPCv1_Metadata.txt` on CATCOURSES (Resources→HOMEWORK→HOMEWORK 7)
- `DEM.tif` on CATCOURSES (FILES→HOMEWORK→HOMEWORK 6)

Reading

- Principal components and factor analysis

Keywords

`read.dbf, hclust(), kmeans(), rect.hclust(), cutree(), prcomp(), biplot()`