Homework 2

PStat 131/231 - Spring 2016

Due April 29th, 2016

Instructions

Use **R Markdown** to produce your solution to the hw and provide a pdf version of it.

Download the **HW template.rmd** from GauchoSpace and open it in R-studio. Insert your text and R code (the R code in an R chunk). Answer all questions in the order given. Clearly show the structure of your answers by recalling the specific question you're answering .

Use the packages `ggplot2` and `data.table` when appropriate.

Question 1

Perform a PCA of the Iris data set and answer the following questions:

a. Give a summary of the PCA results in R
b. Looking at the loadings of the second PC could we interpret it as "Dimension of the Sepal"? Explain why or why not.
c. Use the biplot to visualize the first two PC loading vectors and scores. Use colors to separate the three species on the graph.

   Try to describe the main features of the graph, for example, based on the graph how would you say the Setosa is different from the other two species? How could the virginica and the Versicolor be possibly distinguished?

   (Hint: to reproduce the colors it would be better to define a data set with the factor scores and the Species and use ggplot, to add the loadings vector is necessary to create a different data set and add a layer to the scatterplot of the scores)
d. What is the PVE of the first two components? Produce the scree plots for your analysis

Question 2

With this exercise we try to compare two classification techniques, namely `knn` and `LDA` using the Iris data.

a. Perform a `knn` classification. Choose the optimal `K` using cross validation and provide eplicitly the value of `k` minimizing the cross validation test error rate (Hint: you could use the function `knn.cv` from the package `class`, which provides a LOOCV; see the R help window to get more information on that function)
b. Perform an LDA analysis and evaluate the test error rate by LOOCV.
c. based on the LOOCV estimate of the test error rate in `knn` and `LDA` which classification rule would you choose? If this criterion is not useful, but a choice has to be done what would you choose? Explain your reasoning.
d. If I have two flowers with the following measurements:

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1:            4         2.5          3.0         0.5
## 2:            6         4.0          1.8         1.5
```

```
what is their predicted Species with `knn`? And with `LDA`?
```

Additional exercises for PStat 231

Consider a field of application of your interest. Starting your search in google scholar (https://scholar.google.com/) find one paper which applies a classification technique we have seen in class to a problem connected to the field you are interested in.

Read carefully the paper and summarize (at most one sided page):

1. the problem under consideration and the objectives of the analysis
2. The variables used in the analysis
3. The main findings
4. Briefly discuss of any issue which may concern you (for example, do you agree with the analysis done by the authors of the paper? Do you think the variables used are appropriate for the problem at hand? Should, or could, the authors have considered other variables?)

As an example, if you're interested in the use of LDA in analysis of problems related to credi card use, try typing in Google scholar "credit card discriminant analysis". A list of links to papers (mostly directly downloadable from a UCSB computer) appears. Briefly scan them to find a paper suitable for you and the homework.