Inclass 8 Part 1

October 20, 2014

1. Create a new ipython notebook Inclass8_1.ipynb.
2. Install the pyper library.
3. Import the necessary libraries: pandas, pyper, matplotlib.
4. Read in the Iris data (csv file on Scholar/Resources/Data) as a pandas DataFrame.

The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa = 0*, *Iris virginica=1* and *Iris versicolor = 2*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, we would like to be able to distinguish the species from each other.

5. Examine the data and get numerical summaries using pandas capabilities: .head(), .shape, .describe, value_counts.
6. Create an R instance with pyper. Be sure to include the full path to your R installation.
7. Pass data from Python to R.
8. Now use the "rdata" and R capabilities to: examine the data (head() function, names() function), get numerical summaries (correctly, depending on if a variable is numeric or factor).
9. Access the help file for the R's *princomp* function.
10. Calculate the principal components of the iris data and assign the result to an R object that you can name "p".
11. Display the "names" in the object "p".
12. The actual principal components are saved in the "scores" name in the object "p". All the principal components are calculated by default (as many as the number of columns in the data set used to calculate the principal components). Display the first 6 rows of the principal components.
13. Now pass the principal components you have just visualized into a pandas data frame. Be sure to give proper names to columns using the pd.DataFrame's argument "columns".
14. Examine the newly created pandas dataframe with .head().
15. Now create a basic scatterplot of the first two principal components (with matplotlib.pyplot capabilities). Is there a pattern in the data?
16. Next use the example code from last lecture and build the scatterplot for the first two principal components with the three species differentiated by color. You will have to figure to modify the code slightly to assign different colors to the different species and to display the correct legend. What patterns do you observe now? Are the first two principal components good at summarizing the info in the dataset and discriminating between the three species?
17. Submit your notebook with code, results, plots and comments to Drop Box and Github.