CS5785 Applied Machine Learning

Homework 0

Disheng Zheng, Dan Terry

08/28/2017

Summary

This lab analyzed the Iris Flower database from UC Irvine, the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a species of Iris. The attributes are sepal length, sepal width, petal length, and petal width.

In this lab, we used Anaconda Jupyter Notebook, a web-based interface to write and run Python code. We then parsed the dataset with the pandas library, separating the columns based on ',' signs. We assigned each column with respect to their appropriate attributes: four numerical attributes representing measurements of the flowers and a fifth categorial variable to represent the species of flower.

We created a matrix of bi-dimensional scatter plots (Figure 1.) covering each pair of the four numerical values. The color of the dots on the graph correspond to the three different species in the fifth column. The diagonal entries in the graph show the relative distribution of that one particular variable. The graph is attached below, and the source code is provided in a separate .py file. The ipynb file is included, although it does not contain any additional information. The diagonal graphs are the distributions of each corresponding attributes.

Citations

Libraries and Documentation

http://pandas.pydata.org/

http://www.numpy.org/

http://matplotlib.org/

General Reference

https://stackoverflow.com/questions/31328526/scatter-plot-by-category-in-pandas

Figure 1. Iris Flower Septal and Petal width and length for 3 species.