CST 383: Intro to Data Science
Dr. Glenn Bruns

Correlation and conditional probability

Instructions. This assignment is based on material in class lectures, and also on the section 'Looking for Correlations' in Chapter 2 of "Hands-On Machine Learning".

Answer the following questions by downloading and editing probs.txt. These questions are based on the famous 'iris' data set. Your can read the data using Panda's read_csv() function directly using this URL:

https://raw.githubusercontent.com/grbruns/cst383/master/iris.csv

You will want to write Python code to answer these questions, but I only need to see your answers. Assume the iris data was collected by randomly selecting irises. If a number is required, give 3 significant digits in your answers. Answers within 2% of the correct value are accepted.

1. Use the Pandas dataframe corr() function on the iris data (but omit the 'species' column). Name the two columns that are most strongly correlated, either positively or negatively. You can provide the column names in either order.
2. What is the marginal probability that an iris is of species “setosa”?
3. What is the conditional probability that an iris is of species “setosa” given its sepal length is less than 5?
4. What is the conditional probability that an iris has sepal length less than 5 given its species is “setosa”?
5. What is the conditional probability that an iris has a sepal length greater than 6 and sepal width less than 2.6 given its species is “versicolor”?

Submission. Submit probs.txt on iLearn.