COMPSCI 590N: Assignment 3

Due: September 29, 2016 at 11:55pm

Included with this assignment are two CSV files. You should write python code using NumPy to answer the following questions about these two files. **You may not use any for or while loops in your code.** Please submit to Moodle a zip file containing your code and a PDF with your answers to these questions.

1 Problem 1: Data Filtering

In computing, `NaN` is a special value that represents an undefined number such as 0/0. Operations applied to `NaN` result in `NaN`, so any `NaN` values appearing in a data matrix must be filtered before performing operations on the data. The file "boston.csv" contains a matrix which contains a number of `NaN` values. Please answer the following questions about this matrix. (Hint: see np.isnan)

1. What percentage of entries in the matrix have the value `NaN`?
2. How many rows contain at least one `NaN` value?
3. How many `NaN` values are there in each column?
4. What is the average value of each column, ignoring all rows containing at least one `NaN` value?
5. What is the average value of each column, ignoring `NaN` values, but including other features in the same row as a `NaN` value?

2 Problem 2: Data Exploration

"iris.csv" contains the Fisher Iris dataset, a canonical pattern recognition dataset. Each row in the file corresponds to a data case and contains petal and sepal measurements as well as the iris species encoded as an integer between 0 and 2. The first row in the file contains the column names. Please answer the following questions about the Fisher Iris dataset.

1. What is the average petal length for each iris species? **Copying the same line of code for each species counts as using a for loop!** (Hint: see np.bincount)

2. What is the sepal width for the five data cases with the largest sepal length? (Hint: see np.argsort)

3. We can (crudely) approximate the area of each petal by approximating the petal shape as an ellipse. Then, the petal area can be calculated as $\frac{\pi}{4} \times petal\_length \times petal\_width$. Report **sepal** measurements for the three data cases with the largest petal area according to this approximation.

4. Calculate the 4 by 4 Pearson correlation matrix for the four iris measurements. (Hint: see np.corrcoef)

5. Z-normalization is often used to ensure that the columns of a data matrix, X, have a similar scale. Let μ_f be the empirical mean of measurement f across all data cases and let σ_f be the empirical standard deviation of measurement f across all data cases. Then the Z-normalized data matrix, $\tilde{X}$, is defined as, $\tilde{X}_{if} = (X_{if} - \mu_f)/\sigma_f$ for all i and f. Calculate the Z-normalized data matrix and report the maximum value in each column of the Z-normalized matrix.