**Midterm Study Guide**

80 minutes will be granted for the exam.

**Concept List**

1. Pandas
   1. DataFrames
      1. How to Instantiate
      2. Functions/Methods to use
         1. Drop, merge, concat, fillna, dropna
      3. How to slice and/or select data
      4. Purpose/usefulness of a Dataframe’s index
2. Matplotlib
   1. Default charting and visualization package
   2. Available with regular python install
3. Databases
   1. SQL
4. Python’s Pickle library
   1. Usefulness in data science
      1. Saving a local copy of a large remote dataset
      2. Storing a built machine learning model for later use without having to store training data
5. Python
   1. Slower than some other languages, but much easier to maintain code and takes less developer-time
6. Scikit-learn
   1. Import sklearn
   2. Classification
      1. Discrete
   3. Regression
      1. Continuous
   4. Clustering
   5. Model Selection
   6. Preprocessing
   7. KNN
   8. PCA
   9. KMeans Clustering
   10. Scaling
   11. Decision Trees
   12. Random Forests
7. Supervised vs Unsupervised Models
8. Overfitting

**Example Questions**

1. Describe an ensemble method.
2. What is a method to combat overfitting?
3. How is a PCA model helpful?
4. Give the approximate values of the correlation coefficients for each of these plots:



1. You have a DataFrame df. One of the columns is marital\_status. How would you compute *P*(marital status = 'married') from this dataset?
2. You have data in a file data.csv. Given, that this is a comma-delimited file, how would you build a Pandas Dataframe from this?
   1. Once you have built the DataFrame, how would you determine how many rows it contains?
3. SQL is one of the most-prevalent computer languages used. What is it used for?
4. What is python’s Pickle library used for?
5. What is a Jupyter Notebook and how is it useful?
6. What is the different between a DataFrame’s merge and concat functions?
7. Explain the difference between the two machine learning paradigms of supervised and unsupervised models.
8. How does a KNN Nearest Neighbor model determine an estimate Y, from an input X?
9. What is “Feature Engineering?”
10. Applying a method from scikit-learn to a dataset takes a small number of steps, amounting to only a few lines of code usually. Give these steps for a scikit classifier called "generic", starting from loading your data, and ending with estimating the accuracy of your model. Give approximate code and a short comment describing each step:
11. How can decreasing the depth of a decision tree be beneficial?