Homework-3

March 27, 2017

1 Predicting Individual's Salary

To be completed INDIVIDUALLY and due on March 31 at 3 pm.

In this assignment, we will work on US census data. The goal is to determine whether a person makes over $50K. The dataset was extracted from the 1994 US census data. You can find the details of the dataset here. The data is stored in a tab separated value file, and each line represents a person.

This dataset contains categorical values. Most of the algorithms we learn in class can only handle numeric values and so we need to create a new variable (dummy variable) for every unique value of the categorical variables. To convert the categorical variable into dummy/indicator variables, we will use pandas.get_dummies. Here is an example on how to convert categorical variables to dummy variables:

```
In [1]: import pandas as pd
        raw_data = {'age': [23, 62, 31, 48, 59],
                'salary': [60000, 100000, 120000, 150000, 95000],
                'education': ['Bachelor', 'Masters', 'PhD', 'Jd', 'Masters']}
        df = pd.DataFrame(raw_data, columns = ['age', 'salary', 'education'])
        df_edu = pd.get_dummies(df['education'], prefix = 'edu')
        df_new = pd.concat([df, df_edu], axis=1)
        df_new = df_new.drop( ['education'], axis = 1 )
        df_new
```

```
Out[1]:    age  salary  edu_Bachelor  edu_Jd  edu_Masters  edu_PhD
        0   23   60000             1       0            0        0
        1   62  100000             0       0            1        0
        2   31  120000             0       0            0        1
        3   48  150000             0       1            0        0
        4   59   95000             0       0            1        0
```

Now your job is to convert all categorical features in the dataset to indicator variables.

(10 pts)

```
In [ ]:
```

Now that we know how to convert categorical variables to numerical dummy variables, we can use the algorithms we learned in the class. For the first part, we will use KNeighborsClassifier.

KNeighborsClassifier is an instant-based classification which simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point.

Report the accuracy of your prediction.

(20 pts)

```
In [ ]: from sklearn.neighbors import KNeighborsClassifier
```

For this part of the homework, we will use Decision Trees to determine a person makes over $50K. Report accuracy score using decision trees.

Compare your results with the results you obtained in Part 1.

(20 pts)

```
In [ ]: from sklearn import tree
```