**Module 2.3 – COMP 642-901 – Christian Ruiz, cr72**

4. Download the data for the California Census and answer the following

a. What are the attributes for each district?

It seems that most of the districts contain float64 data on longitude, latitude, median age, total rooms, total bedrooms, population, households, median income, and median house value while ocean proximity is the only categorical feature.

b. What attributes are confusing to you?

It seems that the names of some of the attributes may be confusing such as the median age. Also, ocean proximity may be hard to interoperate looking at <1H OCEAN and NEAR OCEAN. The two definitions seem difficult to differentiate.

c. Without graphing tools, what observations can you make about the data?

- There are some missing values for total bedrooms

- There are only 5 ISLAND homes and the largest count attribute is <1H OCEAN at 9,136 observations

- ISLAND homes have the greatest median age with INLAND homes having the lowest median age

- ISLAND homes have the lowest number of total rooms and total bedrooms compared to the other attributes

- ISLAND has the lowest number of households (unsure what this feature means)

- ISLAND has the lowest median income but also has the lowest std of median income compared to the other homes

- INLAND has the lowest median house values

5. Suppose you have two datasets A and B, they have exactly the same structure that each row contains a single feature x and its corresponding label y and you want to train a linear regression model to fit two datasets.

a. Draw the possible predicted regression line for dataset A. (You can draw by hand)

A graph with a red line

Description automatically generated

b. Draw the possible predicted regression line for dataset B. (You can draw by hand)

A graph with a red line

Description automatically generated

c. Which data points in dataset B are possible outliers, can you draw a circle around them?

A diagram of a line with dots and a red line

Description automatically generated

d. Draw the possible predicted regression line for dataset B after removing the outliers.

(You can draw by hand)

A graph with blue dots and red line

Description automatically generated

e. Suppose we use the mean square error to evaluate the model on dataset B, how does

removing possible outlier affect the mean square error?

The mean square error, compared to the root mean squared error, considers outliers more in it's evaluation by squaring simply the mean error calculation. Removing the outliers would essential lower the error value.

A black background with white text

Description automatically generated

6. Run the notebook of housing price prediction and answer the following questions.

1. Question 1: Frame the Problem

Given a dataset like this, how can it be framed as a machine learning problem (try to frame it in different ways other than predicting housing price)? Is the problem you want to solve supervised learning or unsupervised learning? Classification problem or regression problem?

ANSWER:

There are a few predictions and/or approaches that we could make on this dataset.

1. We can try to use a supervised learning regression approach to predict the median income based on the other features of the data

2. We can also create unsupervised classification mini models of each ocean proximity to determine if there are underlying segments of data present within this categorical data.