# CPSC 483 - Introduction to Machine Learning

Project 4, Fall 2020

due November 9 (Section 02) / November 12 (Section 01)

*Last updated Friday November 6, 4:10 pm PST*

In this project we will see some of the challenges of working with a “real-world” dataset, and see the importance of exploratory data analysis to understand the features.

The project may be completed individually, or in a group of no more than three (3) people. All students on the team must be enrolled in the same section of the course.

## Platforms

The platform requirements for this project are the same as for [previous projects](https://docs.google.com/document/d/1gAvnkp62x00YTYp9Vwo0qbuQQqpxY9pGykYewmgAMAg/edit?usp=sharing).

## Libraries

You will need [scikit-learn](https://scikit-learn.org/) to obtain the data and build models, [pandas](https://pandas.pydata.org/) to analyze the data, and [seaborn](https://seaborn.pydata.org/) to visualize the data.

You may reuse code from the [Jupyter notebooks accompanying the textbook](https://github.com/sdrogers/fcmlcode/tree/master/notebooks) and from the documentation for the libraries. All other code and the results of experiments should be your own.

## Dataset

While they are not included directly with scikit-learn, the [sklearn.datasets](http://scikit-learn.org/stable/datasets/index.html) module includes the ability to fetch some larger “real-world” datasets for experimentation. In this project we will continue our earlier task of trying to predict median values of homes, but this time from the [California Housing dataset](https://scikit-learn.org/stable/datasets/index.html#california-housing-dataset).

Note that in newer versions of scikit-learn, [fetch\_california\_housing()](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html) includes an as\_frame parameter that will add a .frame attribute containing a pandas DataFrame.

## Experiments

Run the following experiments in a Jupyter notebook, performing each action in a [code cell](https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Running%20Code.html) and answering each question in a [Markdown cell](https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Working%20With%20Markdown%20Cells.html).

1. Load and examine the California dataset’s features, target values, and description.
2. Recall that when we originally discussed housing prices, we suggested that the price of a house might depend on how many bedrooms it has. Create and [fit()](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.fit) an [sklearn.linear\_model.LinearRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) model using AveBedrms as a predictor of MedHouseVal. How well does the model [score()](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.score)?
3. Let’s take a closer look at the data. Seaborn’s [pairplot()](https://seaborn.pydata.org/generated/seaborn.pairplot.html) function can be used to plot pairs of features against each other. Plot MedHouseVal as a function of each of the features.

Note that older versions of Seaborn (including Google Colab) may have a [bug](https://github.com/mwaskom/seaborn/issues/2260) that displays the first plot incorrectly. You can work around this by passing the additional parameter diag\_kind=None.

1. Because of the size of the dataset, graphs produced by Seaborn are rather crowded. Try the plot again using a [sample()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html) of 1%. How does the distribution of AveBedrms seem to affect MedHouseVal?
2. Which features seem to have a linear relationship with MedHouseVal?
3. What interesting relationship do you see between MedHouseVal and the Latitude and Longitude? Look these values up on a [map of the state](https://www.mapsofworld.com/usa/states/california/lat-long.html).

(If you are feeling particularly ambitious, you might try [plotting the values on a map](https://jakevdp.github.io/PythonDataScienceHandbook/04.13-geographic-data-with-basemap.html#Example:-California-Cities).)

1. Recall that the [covariance matrix](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.cov.html) shows how pairs of features in a dataset co-vary. What patterns (if any) do you observe? (Hint: use [describe()](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html) to examine distribution of the features before attempting to interpret the results.)
2. Covariance is difficult to interpret because the features are on very different scales. While you could [standardize](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html) the features yourself, the [correlation matrix](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html) is the [covariance matrix of the standardized variables](https://en.wikipedia.org/wiki/Covariance_matrix#Relation_to_the_correlation_matrix). Based on the correlation matrix, which features is the best predictor of MedHouseVal?
3. Repeat experiment *(2)* using the feature you found in experiment *(8)* instead of AveBedrms. How well does this model score?
4. Another way to visualize the predictive value of the two features is to compare the variance. The [seaborn.regplot()](https://seaborn.pydata.org/generated/seaborn.regplot.html) function can be used to create a scatter plot, add a regression line, and plot a 95% confidence interval in a single step. (Recall that 95% corresponds to *±2*𝜎.)

Plot AveBedrms as a predictor of MedHouseVal, then use the feature you found in experiment *(8)*. What difference do you see? (Don’t forget to use the sample you created in experiment *(4)*, or your graph will be difficult to interpret.)

1. Other than the feature you found in experiment *(8)*, there is only a [very weak relationship](https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/) between MedHouseVal and the other features. Fit and score a model to predict MedHouseVal using all the features. Are you surprised by the result? What accounts for the difference from experiment *(9)*?

## Submission

Submit your Jupyter .ipynb notebook file through Canvas before class on the due date. Your notebook should include the usual identifying information found in a README.TXT file.

If the assignment is completed by a team, only one submission is required. Be certain to identify the names of all students on your team at the top of the notebook.