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.

Libraries

You will need scikit-learn to obtain the data and build models, pandas to analyze the data, and seaborn to visualize the data.

You may reuse code from the Jupyter notebooks accompanying the textbook 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` 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.

Note that in newer versions of scikit-learn, `fetch_california_housing()` 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 and answering each question in a Markdown cell.

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()` an `sklearn.linear_model.LinearRegression` model using AveBedrms as a predictor of MedHouseVal. How well does the model `score()`?

3. Let's take a closer look at the data. Seaborn's `pairplot()` 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 that displays the first plot incorrectly. You can work around this by passing the additional parameter `diag_kind=None`.

4. Because of the size of the dataset, graphs produced by Seaborn are rather crowded. Try the plot again using a `sample()` of 1%. How does the distribution of AveBedrms seem to affect MedHouseVal?

5. Which features seem to have a linear relationship with MedHouseVal?

6. What interesting relationship do you see between MedHouseVal and the Latitude and Longitude? Look these values up on a map of the state.

   (If you are feeling particularly ambitious, you might try plotting the values on a map.)

7. Recall that the covariance matrix shows how pairs of features in a dataset co-vary. What patterns (if any) do you observe? (Hint: use `describe()` to examine the distribution of the features before attempting to interpret the results.)

8. Covariance is difficult to interpret because the features are on very different scales. While you could standardize the features yourself, the correlation matrix is the covariance matrix of the standardized variables. Based on the correlation matrix, which features is the best predictor of MedHouseVal?

9. Repeat experiment *(2)* using the feature you found in experiment *(8)* instead of AveBedrms. How well does this model score?

10. Another way to visualize the predictive value of the two features is to compare the variance. The `seaborn.regplot()` 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 $\pm 2\sigma$.)

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.)

11. Other than the feature you found in experiment *(8)*, there appears to be only a very weak relationship between MedHouseVal and the other features. Nevertheless, fit and score a model to predict MedHouseVal using all the features at once. 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.