Adam Thelen and An Nguyen

EE425X

3/11/2020

Project Proposal

Updated Project Plan as of 4/14/2020

The previous project proposal strayed too far from the guidelines. Trying to implement a neural network along with increasing the size of a classification problem would have been too much to study bias-variance tradeoff. The complex challenges of simply implementing the neural net would have taken away from the time needed to fine tune a model and deep dive into the bias-variance relationship of the data. For this reason, we will now be using a dataset found online that contains data of house sales in King County, United States. The data contains 21 features and upwards of 20,000 data points about the sale of the house, including the sale price. The data is for the sale of houses between May 2014 and May 2015. The close range of dates allows us to ignore inflation. Because the feature for the sale date will prove hard to work with, we plan to create a feature to replace it. We propose creating a feature that is the sale date less the year the house was built. This new "age" feature should prove easier to work with in python.

We propose using a multivariate linear regression model. With this model, we will be able to study the bias-variance tradeoff. A multivariate linear regression model will allow us to look into different combinations of features and discern which have the greatest impact on model performance. It is believed that many of the features will lack strong correlation with the house sale price. For this reason, we aim to reduce the total number of features to roughly half the original, (10/21). Reducing the features before beginning more advanced techniques will save us a lot of time in creating and testing many different combinations.

A more detailed test plan is below:

Project Plan:

1. Data simulation:

   We use a similar data simulation process as in homework 4:

   - Dependent variables are generated from a multivariate normal distribution to account for the fact that some of our features are correlated. (n = 100, m = 1000)
   - The theta vector contains a few 0 values to account for the case that some of the features do not contribute to house Price.
   - The noise terms & dependent variable is generated in the same way as in homework 4.

2. Exploratory Analysis

   Plot features and determine which features are most correlated with the price of houses. (can also be done mathematically)

   a. Remove features that are not highly correlated with the price of the house

b. Aiming for roughly 10 remaining features from the initial 21
c. Begin taking different combinations of the 10 remaining features to determine which combination yields the lowest test-MSE (bias and variance)
d. This will be one method of analyzing bias-variance and a step in determining the optimal model with least test-MSE.

3. Model selection:

   We will experiment with both PCA and L1 regularization as methods for model selection.

   PCA:

   - Since one of the key assumptions in linear regression is that regressors are independent, we think that using PCA to project the data onto orthogonal principal components will improve the performance of our model.
   - We will experiment with the number of components to be included (r) to find the model with the lowest test-MSE.
   - Consider combining the PCA preprocessing step with the best model from the previous step and seeing if there is any improvement in Test-MSE.

   L1 Regularization:

   - We decided to go with L1 regularization rather than L2 because this technique will shrink the less important features' coefficients all the way to 0, which has a very clear interpretation. (e.g. on average, having a beach view does not affect the house price.)
   - We will experiment with *lambda* to find the model with the lowest test-MSE.

4. Explain in detail the final model and test-MSE results

Explanation for choosing PCA as a method to improve performance:

PCA processing was chosen because it is an effective way of quickly detecting linearly dependent features and rectifying them. For example, the number of bathrooms in a house is closely related to the number of bedrooms a house has. PCA processing detects the nearly linear relationship between the bathrooms and bedrooms and alters the subspace of the data to reflect this. In this way, PCA processing acts to reduce the dimensionality of the underlying data. The newly crafted data is reduced in dimensionality, which is generally considered an excellent way to improve the performance of linear classifies and avoid overfitting [1].

Current Status of Completed Tasks:

1. Done
2. Done
3. Working on L1 regularization. Codes for the experiment with PCA can be reused from the homework.
4. Pending