AMATH 482/582: HOMEWORK 2

SATHVIK CHINTA

ABSTRACT. Using the Fast Fourier Transform algorithm alongside a Guassian Noise filter, we were able to succesfully de-noise acoustic data of a submarine located within Puget Sound and track its position over a 24 hour period of time.

1. Introduction and Overview

This time, we have a very interesting (and fun) problem to solve. We want to create a model that can identify hand-written digits! The data itself is already split into training and test data, so we won't have to worry about that. Taking a look at the data itself, we can see that there are 2000 images in the train data and 500 in the test, each with 256 modes. Knowing that we most likely won't need to use all the modes to represent our data, we can think about using PCA analysis in order to figure out how many modes we truly need.

2. Theoretical Background

Imagine we are given some dataset in a very high dimensional space. We want to apply a model to identify classes on said dataset, but the complexity of multiple higher dimensions makes this task very difficult. This is where PCA analysis comes in. We check the covariance of each dimension to the other dimensions and then find the eigenvalues and eigenvectors of the covariance matrix. We then sort the eigenvalues in descending order and use the corresponding eigenvectors to project our dataset into a lower dimensional space. This is called PCA analysis. It allows us to also know the variance of each dimension and how much of the variance is explained by each dimension. So, if we know that the variance of some dimension is 20 percent of the total variance, we can say that 20 percent of the data is explained by that dimension. Knowing this, we can remove the dimensions that are not important to our model. This is called dimensionality reduction. Our current dataset has 256 modes of data, so we should perform PCA analysis and see how many of these dimensions we can remove and still have a good model.

Once we have done so, we can use Ridge regression to find the best weights for our model. We can then use the weights to predict the class of a new data point within our dataset. In Ridge regression, we estimate the weights and assume that the dimensions are highly correlated with each other. This is why it's so important to perform PCA analysis before hand. In ridge regression, we attempt to minimize the following equation:

$$minimize_{\beta}||A\beta - Y||^2 + \lambda||\beta||^2$$

Where A is our prediction, y is the true value and λ is our regularization parameter.

3. Algorithm Implementation and Development

I used Python for this project. The major package that I used was sklearn, but I also used numPy and matplotlib.

Within sklearn I used the following functions:

Date: February 10, 2022.

`sklearn.decomposition.PCA()` to perform PCA analysis.
`sklearn.singular values` to find the singular values of the data.
`sklearn.decomposition.PCA().fit()` to fit the PCA model to the training data.
`sklearn.linearModel.Ridge()` to initialize the ridge regression model.
`sklearn.metrics.meanSquaredError()` to find the mean squared error of the model.
`sklearn.linearModel.Ridge().fit()` to fit the ridge regression model to the training data.
`sklearn.linearModel.Ridge().predict()` to predict the class of a new data point.

Within numPy I used the following functions:
`np.load()` to load in both the train and test data.
`np.cumsum()` to get the cumulative sum while looking at the total variance explained by each dimension.

Within matplotlib I used the following functions:
`plt.plot()` to plot all graphs/data

4. Computational Results

5. Summary and Conclusions

Acknowledgements

I am thankful to Professor Hosseini for introducing us to the concept of Fourier Transforms and explaining their significance. I am also grateful to the YouTube Channel "3Blue1Brown" by Grant Sanderson for providing a visual explanation for the intuition behind Fourier Transforms/series, and how they can be used to find signals within data.

I am very thankful to my peers taking the class alongside me, they have helped me understand the material as well as provide a reference to compare my results against. I interacted with them both through Canvas discussion boards as well as Discord chat.