DS403 - Introduction to Statistical Learning

Assigned Date: 02/10/2023

11:59pm, Due Date:10/10/2023

---

1 Real time problems - Linear/Ridge/Logistic Regression - Data Standardization and Normalization [30 Marks]

- Download the datasets 'heights-weights.csv', 'advertising.csv', and 'banknote-authentication.csv' https://drive.google.com/drive/folders/1eHRq7x7PPPyrkf-CAnvU1wNJ2a61kToy?usp=sharing from google drive.
- You are **NOT** allowed to use machine learning libraries such as 'scikit-learn' to build the models for this assignment.

Multiple linear/ridge regression for sales prediction

1.1 Without Data Normalization

1. Use linear/ridge regression model to learn the relationship between sales and advertising budget for a product. The 'advertising.csv' dataset contains statistics about the sales of a product in 200 different markets, together with advertising budgets in each of these markets for different media channels: TV, radio, and newspaper. The sales are in thousands of units and the budget is in thousands of dollars.
2. Plot the 3-d input vectors (Use tsne tool and visualize the data)
3. Split the data into train and test set. Use the training data to learn the function. Train linear and ridge regression models to predict the sales of the product given the TV, radio, and newspaper ad budgets. compute training and testing error for both the models. (fix the $\lambda = 0.01$ for the ridge regression)
4. Experiment with λ value in ridge regression case: plot the test/training error Vs $ln(\lambda)$.

1.2 With Data Normalization & Standardization

1. Models can be sensitive to different scales of the features/independent variables. Hence, it is important to normalize them. Apply data normalization and data standardization and fit the models. Refer this link: https://courses.cs.washington.edu/courses/cse446/21wi/sections/04/section04.pdf
2. Plot the 3-d input vectors after data normalization and standardization (Use tsne tool and visualize the data). Compare the plots with 2nd question in 1.1.
3. Compare the training and test error with 3rd question in 1.1.

2 Logistic Regression [15 Marks]

Multiple logistic regression to identify forged banknotes.

1. Train logistic regression model to identify forged banknotes. The 'banknote-authentication.csv' dataset has been created from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400×400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images. Train a logistic regression model to distinguish between genuine and forged banknotes given features extracted from their images." Compute the classification accuracy of the model trained.

2. If you have implemented 'Logistic-Regression' correctly, the **test accuracy** should be > 0.98. Please verify.

3 Understanding Bias Variance Trade off [30 Marks]

Toy Problem - Apple Stock Prediction - Polynomial Fitting
Find the apple stock data :
https://drive.google.com/drive/folders/1eHRq7x7PPPyrkf-CAnvU1wNJ2a61kToy?usp=sharing

1. Load the data and plot the stock value Vs time.

2. Split the data into train and test. (data upto 2022, consider as training data, remaining data can be used as testing data)

3. Apply polynomial fit to predict the current stock value from the past values.

$$Y_t = \beta_{1,t-1}Y_{t-1} + \beta_{2,t-1}Y_{t-1}^2 + - - - - - + \beta_{p-1,t-k}Y_{t-k}^{p-1} + \beta_{p,t-k}Y_{t-k}^p$$

4. Add more features using the higher powers of the past stock values in the dataset. We have two parameters in hand i) k- past window length ii)p-order of the polynomial. Play with $k = 1, 2, 3$ and $p = 1, 2, 3$. k and p should be atleast 3.

5. Computer the prediction accuracy for the test data and plot the predicted stock value and original stock value Vs time on a single plot. (For all values of k and p). i) Observe the change in parameter values β w.r.t p and k ii) comment on bias and variance.

6. Plotting: i) Plot the prediction error Vs p (order of the polynomial fit - model complexity) ii) Plot the prediction error Vs k iii) 3-D plot Test error (prediction accuracy) Vs k and p. k should be y- axis and p should be x-axis.