Fall 2021: CSEE5590/490 – Special Topics

Python and Deep Learning - ICP-6

Lesson Overview:
In this lesson we will review regression techniques
Regression techniques

a. Linear Regression
b. Multiple Regression
c. Clustering
d. PCA

Source Code:
Provided in the assignment repo & Canvas use-case file.

Regression Assignment:
For question 1 use the same dataset used in the source code (House Prices).

Delete all the outlier data for the GarageArea field **(for the same data set in the use case: House Prices)**.

*for this task you need to plot GaurageArea field and SalePrice in scatter plot, then check which numbers are anomalies.

2. Evaluate the model using MAE, MSE, RMSE and R2 score.

3. Using simple regression select one feature that is positively correlated with 'SalePrice' create a regression model and Plot the regression line between the two features.

For questions 2 and 3 use the Restaurant Revenue Prediction dataset "rest_data.csv" described here:
https://www.kaggle.com/c/restaurant-revenue-prediction/data

- Id : Restaurant id.
- City Group: Type of the city. Big cities, or Other.
- Type: Type of the restaurant. FC: Food Court, IL: Inline, DT: Drive Thru, MB: Mobile
- P1, P2 **-** P37: There are three categories of these obfuscated data. Demographic data are gathered from third party providers with GIS systems. These include population in any given area, age and gender distribution, development scales. Real estate data mainly relate to the m2 of the location, front facade of the location, car park availability. Commercial data mainly include the existence of points of interest including schools, banks, other QSR operators.
- **Revenue**: The revenue column indicates a (transformed) revenue of the restaurant and **is the target of predictive** analysis. Please note that the values are transformed so they don't mean real dollar values.

2. Create Multiple Regression for the "**Restaurant Revenue Prediction**" dataset.

(bonus) 3. Find top 5 most correlated features to the target label(revenue) and then build a model on top of those 5 features. Evaluate the model using MAE, MSE, RMSE and R2 score and then compare the result with the RMSE and R2 you achieved in question 2.

K-means & PCA Assignment:

1. Apply K means clustering to credit card dataset: CC.csv

   - Remove any null values by the mean.
   - Use the elbow method to find a good number of clusters with the K-Means algorithm
   - Calculate the silhouette score for the above clustering.

2. Try feature scaling and then apply K-Means on the scaled features. Did that improve the Silhouette score?

*** **Bonus points**

2. Visualize the clustering of first question.
3. Apply PCA on the same dataset. Apply K-Means algorithm on the PCA result and report your observation if the silhouette score improved or not?

Maximum bonus points = 10 points.

** Follow the IPC rubric guidelines.

Submission Guidelines:

1. Once finished present your work to TA during class time.
2. Once evaluated submit your source code and documentation to GitHub and represent the work in a ReadMe file properly (short summary for the ICP).

After class submission:

1. Complete your work and submit to your repo before the deadline.
2. Record a short video (1~3) minute, explaining the technical part and method used.
3. Add video link to ReadMe file.

Note: *Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL: https://catalog.umkc.edu/special-notices/academic-honesty/*