COMPSCI 762 2022 S1 Week 5 Questions – Regression & Preprocessing

Luke Chang

March 17, 2022

Question 1 – Regression

Consider a time series data as the following:

X	8	3	2	10	11	3	6	5
Y	4	12	1	12	9	4	9	6

1. Use the least square method to determine the equation of line of best fit for the data. Then plot the line.
2. Given a new data point $X = 9$, what is the predicted value of Y?

Question 2 – Data Cleaning

You sent a survey to collect data about customers who buy lunch in the university cafe. Given the following data for the attribute – *Age*:

$$13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70$$

1. Use *smoothing by bin means* to smooth these data with 3 bins. Illustrate your steps. Comment on the effect of this technique for the given data.
2. How do you determine *outliers* in the data?
3. What other methods are there for data smoothing?

Question 3 – Preprocessing on Real-World Data

Your team decides to participate a Kaggle competition on predicting car price. Here is the URL for the Kaggle page: `https://www.kaggle.com/ersany/car-price-prediction?resource=download&select=car_price.csv`.

1. What preprocessing techniques would you apply in order to train a regression model?
2. There is no missing value in the dataset. What if 20% of `body_type` are missing? What imputation technique should you use?

3. What if 20% of `body_color` are missing? What imputation technique should you use? Can you use the same imputation technique for both attributes?

Note: The documentation of `scikit-learn` preprocessing is a good starting point. The link can be found here: `https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing`. Writing code for Question 3 can certainly help you to determine the optimal solution. However, it is not required. You must explain why a particular technique has been selected.