COMPSCI 762 2022 S1 Week 5 Solution

Luke Chang

April 7, 2022

Question 1

- Using partial derivative to solve least square problem.
- Fit the data into a straight line $y' = ax + b$, by minimizing:

$$f(a,b) = \sum (y_i - (ax_i + b))^2$$

- Computing partial derivative:

$$\frac{\partial f}{\partial a} = \sum 2(y_i - (ax_i + b))(-x_i)$$

$$\frac{\partial f}{\partial b} = \sum 2(y_i - (ax_i + b))(-1)$$

- To find the extrema, we assign the two equations above to 0, then we have:

$$\sum (x_i^2 a + x_i b - x_i y_i) = 0$$

$$\sum (x_i a + b - y_i) = 0$$

- After simplification, we have:

$$(\sum x_i^2)a + (\sum x_i)b = \sum x_i y_i$$

$$(\sum x_i)a + nb = \sum y_i$$

- $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $(\sum x_i^2)$ are constants, so we have a 2×2 linear system.
- Rewrite the equations above by moving unknowns to the left:

$$a = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum(x_i^2) - (\sum x_i)^2}$$

$$b = \frac{\sum y_i - a\sum x_i}{n}$$

- In this case, $\sum x_i$, $\sum y_i$, $\sum x_i y_i$ and $(\sum x_i^2)$ equal to 48, 57, 385 and 368 respectively. Thus,

$$y' = 0.54x + 3.90$$

- When $x = 9$, $y' = 8.7$.
- We can also solve the same problem using pseudoinverse. In this case we treat X as a $2 \times n$ matrix, where the 1st column filled with 1. The intuition is we remove the bias term by integrating it into X.

Question 2

- $27/3 = 9$, sort the list, each bin will have 9 data points.
- After smoothing by bin means, the value of bins are: 18, 28, 44.
- To compute the standard deviations (σ)

$$\sigma = \sqrt{\frac{1}{N}\sum_{N}^{i=1}(x_i - \mu)^2}$$

- Mean ans SD for each bin:

	Mean	σ	2σ
Bin 1	18	2.9	6
Bin 2	28	4.4	9
Bin 3	44	10.9	22

- Compute the absolute deviation, $|x - \mu|$:
  - **Bin 1:** $5,3,2,2,1,2,2,3,4$
  - **Bin 2:** $6,3,3,3,3,1,5,5,7$
  - **Bin 3:** $9,9,9,8,4,1,2,8,26$
- Let outlier be the data points outside 95% confidence interval (2σ), then we identify the data point "70" is an outlier.
- **Note:** If the question does not provide the rule for outliers, we must define it by ourself. Thus, this question may have different solutions depend on the criteria of outliers.
- Alternative methods:
  - Binning: By mean, median and mode;
  - Regression: Not the best choice for 1D data;
  - Clustering: Unsupervised learning, e.g., K-means algorithm;
  - Moving average with fixed window size;
- Moving average with window size = 3:

$$x'_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}$$

- For the 1st data point, x_{i-1} is not available. In this case, we use x_i to fill the empty space. We also use the same technique on the last data point.
- **Note:** When applying a machine learning algorithm for smoothing (e.g., K-means, regression), we cannot use the target value (y). Because y will be unavailable at inference time.

Question 3

- 'Audi A2' only appeared once. If we remove it, then "make_model" becomes binary.
- The attribute "hp" has one unique value.
- The attribute "Extras" contains comma separated strings; Split the string and apply encoding on 'Extras';
- 6% "Gearing Type" are neither "Manual" nor "Automatic". Based on the fact that there are only 2 models, and "Tiptronic" and "Semi-automatic" are just alternative name for "Automatic", we can replace others with "Automatic".
- There are multiple ways to encode "body_type", the imputation method will limit by the encoding strategy we selected.

- If we convert "body_type" into ordinal data, then we can sort it based on the size of the vehicle. Thus, impute with median value is valid. Otherwise, only mode is available for a simple imputer.
- We apply more advanced imputation methods, only if the simple imputer performs poorly. In this case, we don't know. A machine learning pipeline contains many moving parts (parameters), we do not want to over engineering a step unless there is a good reason to do so.
- "body_color" is nominal data (categorical but unordered). We can impute the missing value with mode. Mean and median are not visible, since we cannot add or sort the data.
- We can order color based on hue or color spectrum, but it is difficult for white, black and silver, which are the majority of "body_color". Once again, do not over complicate the task, unless you have an unconfirmed hypothesis. E.g., we want to predict car accidents, you hypothesize color spectrum relates to the visibility of the vehicle at night. Therefore, using ordinal encoding may provide additional information to the model.
- When analyzing features, we must consider the data type. The default `corr()` method in `Pandas` is for continuous data. When mean and SD are absent, we need to consider nonparametric tests which often based on ranking.