Aprendizagem 2023

Lab 1: Univariate Data Analysis

Prof. Rui Henriques

Practical exercises

I. Univariate statistics

Consider the following dataset:

	y_1	y_2	y_3
x_1	0.2	0.5	A
x_2	0.1	-0.4	A
x_3	0.2	-0.1	A
x_4	0.9	0.8	B
x_5	-0.3	0.3	B
x_6	-0.1	-0.2	B
x_7	-0.9	-0.1	C
x_8	0.2	0.5	C
x_9	0.7	-0.7	C
x_{10}	-0.3	0.4	C

1. Approximate y1 distribution using a histogram using 4 bins in [-1,1].
Using the histogram, approximate the probability density function.

$$\{p(-1 \leq v_1 \leq -0.5) = 0.1, p(-0.5 < v_1 \leq 0) = 0.3, p(0 < v_1 \leq 0.5) = 0.4, p(v_1 \geq 0.5) = 0.2\}$$

2. Compute the boxplot of y1 variable. Are there any outliers?

Please note that there are many variants for computing quantiles[1]. One possibility:

$$u = 0.07, median = q_n(50) = 0.15, q_n(25) = -0.3, q_n(75) = 0.2,$$
$$IQR = 0.5, bounds = [-1.05, 0.95]$$

According to the computed quartiles, there are no outliers falling outside the IQR-based bounds.

3. Are y1 and y2 variables correlated? Compare Pearson and Spearman coefficients.

$$PCC(y_1, y_2) = \frac{\sum_{i=1}^{n}(a_{i1} - \overline{y_1})(a_{i2} - \overline{y_2})}{\sqrt{\sum_{i=1}^{n}(a_{i1} - \overline{y_1})^2}\sqrt{\sum_{i=1}^{n}(a_{i2} - \overline{y_2})^2}} = 0.09$$

In the presence of ranking ties, classic Spearman is generally replaced by the PCC of the ranks. Let us compute both:

$$Spearman(y_1, y_2) = PCC([7,5,7,10,2.5,4,1,7,9,2.5],[8.5,2,4.5,10,6,3,4.5,8.5,1,7]) = 0.198$$

[1] https://en.wikipedia.org/wiki/Quantile

Variables y1 and y2 are loose-to-moderately correlated. Rank correlation (under Spearman coefficient) is higher than linear correlation (under Pearson correlation), suggesting stronger correlation in order than magnitude.

4. Identify the probability mass function of y3.

$$\{p(y_3 = A) = 0.3, p(y_3 = B) = 0.3, p(y_3 = C) = 0.4\}$$

II. Data preprocessing

Consider the following dataset:

	y_1	y_2	y_3	y_4	y_{out}
x_1	0.2	0.5	A	A	A
x_2	0.1	-0.4	A	A	A
x_3	0.2	0.6	A	B	C
x_4	0.9	0.8	B	B	C
x_5	-0.3	0.3	B	B	B
x_6	-0.1	-0.2	B	B	B

where y1 and y2 are numeric variables in [-1,1], y_3 and y_4 are nominal, and y_{out} is ordinal

5. On unsupervised feature importance:

a) Considering standard deviation, which numeric variable is less relevant?

Variable y_1 has lower variability than y_2, therefore should be removed.

b) Considering entropy, which nominal variable is less relevant?

$$E(y_3) = 1, \qquad E(y_4) = 0.918$$

Variable y_4 has lower entropy than y_3, therefore should be removed.

6. On supervised feature importance:

a) According to Spearman, which numeric variable is less relevant?

$$\text{Spearman}(y_1, y_{out}) < \text{Spearman}(y_2, y_{out})$$

Variable y_1 is less correlated with the output variable, therefore is less relevant (candidate to be removed)

b) According to information gain, which nominal variable is less relevant?

$$IG(y_{out}|y_j) = E(y_{out}) - E(y_{out}|y_j)$$

$$E(y_{out}) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 1.585$$

$$IG(y_{out}|y_3) = 1.585 - 0.918 = 0.667, \qquad IG(y_{out}|y_4) = 1.585 - \frac{4}{6} = 0.918$$

Variable y_3 has lower information gain, therefore should be removed.

7. Normalize y_2 using min-max scaling and standardization. Compare the results

Considering min-max scaling, $\frac{a_{ij}-min}{max-min}$: $y'_2 = (0.75 \quad 0 \quad 0.833 \quad 1 \quad 0.583 \quad 0.167)$

Adjusting y_2 to a standard Gaussian, $\frac{a_{ij}-\mu}{\sigma}$: $y'_2 = (0.494 \quad -1.413 \quad 0.706 \quad 1.130 \quad 0.071 \quad -0.989)$