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	7	0.2	0.5	A
x_2	5	0.1	-0.4	A
x_3	7	0.2	-0.1	A
x_4	10	0.9	0.8	B
x_5	2,5	-0.3	0.3	B
x_6	4	-0.1	-0.2	B
x_7	1	-0.9	-0.1	C
x_8	7	0.2	0.5	C
x_9	9	0.7	-0.7	C
x_{10}	2,5	-0.3	0.4	C

Ordered →

y_1	Rank
-0,9	1
-0,3	2,5
-0,3	2,5
-0,1	4
0,1	5
0,2	7
0,2	7
0,2	7
0,7	9
0,9	10

Q_1 ← -0,3

Q_2 = 0,15

Q_3

IQR = $Q_3 - Q_1$ = 0,5

Bounds = [−0,3 +1,5 ×0,5 , 0,2 +1,5×0,5]

= [−1.05, 0.95]

∴ No outliers

1. Approximate y1 distribution using a histogram using 4 bins in [-1,1].

   Using the histogram, approximate the probability density function.

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

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.

$$\overline{y}_1 = \frac{0.2+0.1+\dots}{10} = 0.07$$

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

→ $(y_1 - \overline{y}_1) = (0.2 - 0.07, 0.1 - 0.07, \dots)$

↓ Pearson

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

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

$$\overline{R[y_1]} = \frac{7+5+7+\dots}{10} = 5.5$$

↳ $(R[y_1] - \overline{R[y_1]}) = (7 - 5.5, 5 - 5.5, \dots)$

[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 4.5	0.5 4	A	A	A 1.5
x_2	0.1 3	-0.4 1	A	A	A 1.5
x_3	0.2 4.5	0.6 5	A	B	C 5.5
x_4	0.9 6	0.8 6	B	B	C 5.5
x_5	-0.3 1	0.3 3	B	B	B 3.5
x_6	-0.1 2	-0.2 2	B	B	B 3.5

$$\overline{y_2} = \frac{0.5-0.4+0.6+\ldots}{6} = 0.267$$

$$\overline{y_1} = \frac{0.2+0.1+\ldots}{6} = 0.167$$

$$Var(y_1) = \frac{\Sigma(y_1 - \overline{y_1})^2}{6} = \frac{\Sigma((0.2-0.167)^2, (0.1-0.167)^2, \ldots)}{6} \simeq 0.139$$

$$Var(y_2) = 0.186$$

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. $\rightarrow$ $Var(y_1) < Var(y_2)$

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

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

y_4: $\cdot\, P(A) = \frac{1}{3}$, $\cdot\, P(B) = \frac{2}{3}$

$$H(y_4) = -\sum P_i \log_2 P_i = -\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) = -\left(\frac{1}{3}(-1.585) + \frac{2}{3}(0.585)\right) = 0.918$$

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?

$$H(Y_{out} \mid Y_4 = A) = -(\log 1) = 0$$

$$H(Y_{out} \mid Y_4 = B) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = 1$$

$$H(Y_{out} \mid Y_4) = \frac{2}{6} \times 0 + \frac{4}{6} \times 1 = \frac{4}{6}$$

$$H(Y_{out} \mid Y_3 = A) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) = 0.918$$

$$H(Y_{out} \mid Y_3 = B) = 0.918$$

$$H(Y_{out} \mid Y_3) = P(Y_3 = A) \times 0.918 + P(Y_3 = B) \times 0.918 = 0.918$$

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

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

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

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, $\frac{a_{ij} - \mu}{\sigma}$: $y'_2 = (\boxed{0.494} \quad -1.413 \quad 0.706 \quad 1.130 \quad 0.071 \quad -0.989)$

$min = -0.4$

$max = 0.8$

$$0.5 \rightarrow \frac{0.5 - (-0.4)}{0.8 - (-0.4)} = 0.75$$

$$\mu = \overline{y_2} = 0.267$$

$$\sigma = \sqrt{\frac{\Sigma(y_2 - \overline{y_2})^2}{5}} = 0.472$$

$$0.5 \rightarrow \frac{0.5 - 0.267}{0.472} = 0.494$$