Aprendizagem 2024

Lab 1: Univariate Data Analysis

Practical exercises

I. Univariate statistics

Consider the following dataset:

	y_1	y_2	y_3
$\mathbf{x}_1$	0.2	0.5	A
$\mathbf{x}_2$	0.1	-0.4	A
$\mathbf{x}_3$	0.2	-0.1	A
$\mathbf{x}_4$	0.9	0.8	B
$\mathbf{x}_5$	-0.3	0.3	B
$\mathbf{x}_6$	-0.1	-0.2	B
$\mathbf{x}_7$	-0.9	-0.1	C
$\mathbf{x}_8$	0.2	0.5	C
$\mathbf{x}_9$	0.7	-0.7	C
$\mathbf{x}_{10}$	-0.3	0.4	C

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

   Using the histogram, approximate the probability function.

   $\{p(-1 \leq y_1 \leq -0.5) = 0.1,\ p(-0.5 < y_1 \leq 0) = 0.3, p(0 < y_1 \leq 0.5) = 0.4, p(y_1 \geq 0.5) = 0.2\}$

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.

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

$$PCC(y_1, y_2) = \frac{\sum_{i=1}^{n} (a_{i1} - \bar{y}_1)(a_{i2} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{n} (a_{i1} - \bar{y}_1)^2} \sqrt{\sum_{i=1}^{n} (a_{i2} - \bar{y}_2)^2}} = 0.09$$

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

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

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.

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

5. Assume y2 distribution is conditional to y3 classes and follows a Gaussian assumption.

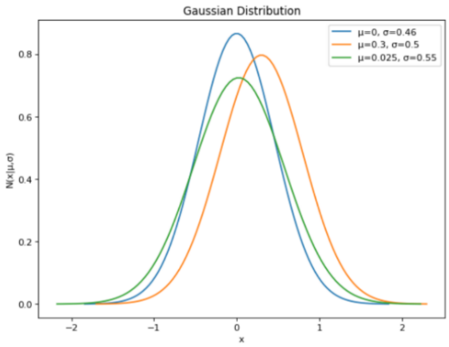
   a) Identify their parameters and plot by hand the distributions.

   Considering that the provided data is a sample of a larger population, corrected standard deviation is necessary.

$$N_{y_2|c=A}(u_{y_2|y_3=A} = 0, \sigma_{y_2|y_3=A} = 0.46)$$

$$N_{y_2|c=B}(u_{y_2|y_3=B} = 0.3, \sigma_{y_2|y_3=B} = 0.5)$$

$$N_{y_2|c=C}(u_{y_2|y_3=C} = 0.025, \sigma_{y_2|y_3=C} = 0.55)$$

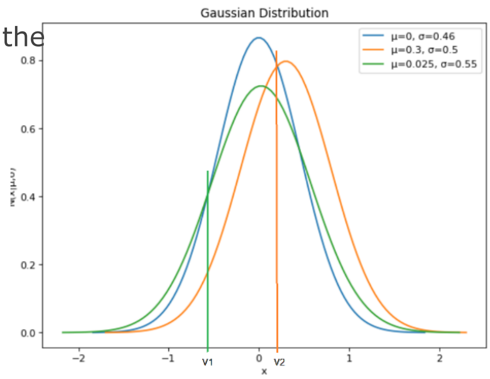


   b) Visually annotate the discriminant rules for the classification of y3 using y2 values.

$$y_2 < v_1 \Rightarrow C$$

$$v_1 < y_2 < v_2 \Rightarrow A$$

$$y_2 > v_2 \Rightarrow B$$



II. Data preprocessing

Consider the following dataset:

	y_1	y_2	y_3	y_4	y_{out}
$\mathbf{x}_1$	0.2	0.5	A	A	A
$\mathbf{x}_2$	0.1	-0.4	A	A	A
$\mathbf{x}_3$	0.2	0.6	A	B	C
$\mathbf{x}_4$	0.9	0.8	B	B	C
$\mathbf{x}_5$	-0.3	0.3	B	B	B
$\mathbf{x}_6$	-0.1	-0.2	B	B	B

where y_1 and y_2 are numeric variables in [-1,1], y_3 and y_4 are nominal, and $yout$ is ordinal

6. 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?

$$H(y_3) = 1,\ \ H(y_4) = 0.918$$

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

7. 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?

$$IG\left(y_{out}\middle|y_j\right) = H(y_{out}) - H\left(y_{out}\middle|y_j\right)$$

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

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

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

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

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

Adjusting y_2 to a standard Gaussian, $\dfrac{x_{ij} - \mu_j}{\sigma_j}$:

$$y'_2 = (0.494 \quad -1.413 \quad 0.706 \quad 1.130 \quad 0.071 \quad -0.989)$$

9. Binarize y_1 considering

a) equal-width/range discretization

Assuming $y_1 \in [-1,1]$, then $\mathbf{y'}_1 = (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0)$

b) equal-depth/frequency discretization

$\mathbf{y'}_1 = (1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0)$

Programming quest

10. Given the *breast.w.arff* dataset and the provided Jupyter notebook on *Data Exploration*, explore the dataset and rank input variables according to their information gain (*mutual_info_classif*).