Aprendizagem 2024

Lab 1: Univariate Data Analysis

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 y_1 distribution using a histogram with 4 bins in [-1,1].

   Using the histogram, approximate the probability function.
2. Compute the boxplot of y_1 variable. Are there any outliers?
3. Are y_1 and y_2 variables correlated? Compare Pearson and Spearman coefficients.
4. Identify the probability mass function of y_3.
5. Assume y_3 class-conditional distributions of y_2 follow a Gaussian distribution.
   a) Identify their parameters and plot by hand the distributions.
   b) Visually annotate the discriminant rules for the classification of y_3 using y_2 values.

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

7. On supervised feature importance:
   a) According to Spearman, which numeric variable is less relevant?
   b) According to information gain, which nominal variable is less relevant?

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

9. Binarize y_1 considering
   a) equal-width/range discretization
   b) equal-depth/frequency discretization

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*).