Aprendizagem 2021

Lab 1: Univariate Statistics and Evaluation

Practical exercises

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

I. Univariate statistics

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

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

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

4. Identify the probability mass function of y3.

5. Assume y3 class-conditional distributions of y2 follow a Gaussian distribution.

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

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

II. Lazy learning

6. Assuming a *k*-nearest neighbor with *k*=3 applied within a leave-one-out schema:

a) Let y_3 be the output variable (categoric). Considering an Euclidean ($l2$) distance, provide the classification estimates for x_1 and x_7.

b) Let y_2 be the output variable. Considering equally weighted numeric-categoric variables with Manhattan ($l1$) and Hamming distances, provide the mean estimates for x_4 and x_9.

7. Consider a weighted-distance *k*-nearest neighbor with *k*=5 and the input data as training, compare the classification estimates for < –0.2, 0.5, **?** > assuming:

a) Chebyshev ($l\infty$) distance

b) cosine dissimilarity

III. Evaluation

8. Consider the following y_3 paired estimates, $\hat{y}_3 = [B\ B\ A\ C\ B\ A\ C\ A\ B\ C]$.

a) Draw the confusion matrix

b) Compute the *k*NN accuracy and sensitivity/recall per class

c) Considering class *C*, identify its precision and F-measure

d) Identify the accuracy, sensitivity, and precision of the random classifier

9. Consider the following y_2 paired estimates, $\hat{y}_2 = [0.3\ \text{-}0.5\ 0.5\ 0.5\ 0.4\ \text{-}0.2\ 0.1\ 0.5\ \text{-}0.9\ 0.4]$.

a) Compute the mean absolute error and root mean squared error

b) Perform a residue analysis to assess the presence of systemic biases against y1

Programming quests

10. Using **sklearn**, replicate the following code for learning and assessing the kNN classifier. Replace the iris dataset by the provided dataset for this lab.
Resource: https://medium.com/@jebaseelanravi96/machine-learning-iris-classification-33aa18a4a983

11. Using **plotly**, plot the class-conditional distributions of each input variable of the iris dataset. By visual analysis, identify which variable appears to be a better predictor of the class.
Resource: https://plotly.com/python/histograms/ (overlaid histograms)