TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework I
Deadline 29/9/2024 20:00
Submit on Fenix as pdf

I) Correlation (2 pts)

Compute the correlation (Pearson correlation) and **Spearman's rank** for two variables **x**1 and **x**2. (Indicate all computational steps!)

(a) (1 pts)
x1 = (1, 3, 4, 6)
x2 = (-30, -10, 0, 20)
Why are the same?

(b) (1pts)
x1 = (1, 3, 4, 6)
x2 = (-3, -0.5, 29, 30)
Why are they different?

II) Decision Trees (5 pts)

F1	*F2*	*F3*	*F4*	*Output*
c	*a*	*b*	*x*	*n*
a	*a*	*c*	*a*	*t*
a	*b*	*b*	*a*	*t*
c	*b*	*c*	*x*	*m*
a	*b*	*b*	*c*	*f*

(a) (2 pts)
Determine the root of decision tree using the ID3 algorithm with the target "Output". Indicate the calculation. (Indicate all computational steps!)

(b) (2 pts)
Determine the decision tree using the ID3 algorithm with the target "Output". Indicate the calculation and draw your decision tree.

(c) (1 pts)
Draw the training confusion matrix for the learnt decision tree.

TÉCNICO LISBOA

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework I
Deadline 29/9/2024 20:00
Submit on Fenix as pdf

III Software Experiments (3pts)

Download the jupyter notbook HM1_DT.ipynb.

We will use the build in wine data set:

Using chemical analysis to determine the origin of wines.
These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Scikit-learn only supports **binary splits** and **numerical variables** for now

`(a) (1pts)`
Split the data using the command (in the notebook)

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=**value**,stratify=y,random_state=**your_group number**)

Partition data with train_test_split, with values 0.2, 0.5, 0.7 and indicate the depth and the accuracy of each the decision tree (if you have no group yet, put the last three digits of student nr).

`(b) (1pts)`
Draw the decision tree for the value 0.7 (copy the drawing into your document)

(c) (1pts)
Now perform the same experiment without the command stratify=y,with the value 0.7

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=**0.7**, ,random_state=**your_group number**)

Draw the decision tree (for the value 0.7, copy the drawing into your document)

Are the result different, what is the meaning of the command stratify=y and stratified fashion? (See scikit-learn manual) and write in **one** sentence pls, do not copy the whole section (only the important part, if you copy…)