INF442 - Algorithmes pour l'analyse de données en C++ (2021-2022)

Started on	Wednesday, 20 April 2022, 7:25 AM
State	Finished
Completed on	Thursday, 21 April 2022, 4:37 AM
Time taken	21 hours 11 mins

Quiz navigation

1 2 3 4 5 6 7 8 9

Show one page at a time

Finish review

Question 1
Complete
Marked out of 1.00
Flag question

What is the setting of today's lab?

Select one:
- a. classification
- b. regression

Question 2
Complete
Marked out of 1.00
Flag question

Which algorithm should we use for email classification (see this page for the details on the dataset) with Knn?

Select one:
- a. linear scan
- b. kd-trees with backtracking search
- c. both are roughly equivalent in this setting

Question 3
Complete
Marked out of 1.00
Flag question

Use knnclassifier.py script to perform 5-neighbour classification for the audit_train.csv/audit_test.csv. What is the total number of incorrectly classified instances in the test set?

Answer: 20

Question 4
Complete
Marked out of 1.00
Flag question

Now use the same model as in the previous question (5-neighbours trained on audit_train.csv). Find the number of incorrectly classified instances if the same model is used to predict labels on the train data again.

Answer: 39

Question 5
Complete
Marked out of 1.00
Flag question

In the two previous cases (predicting on the test and train data), which error rate is significantly smaller?
(question for you: why?)

- a. When predicting on the test data
- b. When predicting on the train data

Question 6
Complete
Marked out of 1.00
Flag question

Use provided function normalize to perform also 5-neighbour classification for the normalized datasets (using method="std_mean" and method="maxmin"). Choose the setting with the highest F-score.

- a. Normalized with "mean_std"
- b. Non-normalized
- c. Normalized with "maxmin"

Question 7
Complete
Marked out of 1.00
Flag question

Use the audit_train.csv/audit_test.csv under the maxmin normalization and, for each feature, try to remove it and perform 7-neighbor (now 7, not 5 !) classification.
Removing which of the features has the highest impact (that is, the absolute value of the change is the largest) on the F-score?
Hint: you may find the method drop of the pandas DataFrame useful for this task. From a DataFrame data, you can remove a column called "A" by doing
data.drop("A", axis=1)

- a. numbers
- b. History
- c. Sector_score
- d. Loss
- e. PARA_B
- f. PARA_A
- g. Money_Value

Question 8
Complete
Marked out of 1.00
Flag question

For the maxmin-normalized audit dataset, build ROC curves for the number of neighbors k = 3, 5, 7, 9. For which of the values of k the area under the ROC curve (it is displayed on the plot) will be the largest?

- a. 3
- b. 7
- c. 9
- d. 5

Question 9
Complete
Not graded
Flag question

For the maxmin-normalized audit data trained with the number of neigbors k = 3, build the ROC curve. You can see that the plot starts with a nearly vertical line reaching almost 0.9.
What does this line mean?

- a. It means that one can reach almost 90% true positive rate if a point is considered positive as soon as it has at least one positive neighbour.
- b. It means that one can reach almost 90% true positive rate if a point is considered positive as soon as it has at least two positive neighbours.
- c. It means that one can alter the value of k to get the true positive rate being almost 90%

Finish review