Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfadr

CE5501 | Fall 2022

Teaching Assistants

Mohsen Ebadpour(head) (M.Ebadpour@aut.ac.ir)

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Mohamadreza Jafaei (Mr.Jafaei@aut.ac.ir)

Assignment (2)

Outlines. In this assignment, K-NN and Decision Tree are noticed as well as evaluation metrics.

Deadline. Please submit your answers before the end of November 13th in courses.aut.ac.ir. Other methods like sending via email or in social networks are not accepted and will not be considered.

Assignment Manual

Delay policy. During the semester, you have extra 7 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you loss 20% points of that assignment. After 4 days you miss all points and any submit doesn`t acceptable. Remember that saving this time doesn`t have any extra point.

Sharing is not caring. Students are free to discuss and share their ideas about problems with others. But sharing source codes, solutions, answers and other results is not allowed and based on university`s rule, both sides will be graded zero.

Problems are waiting you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then discuss about it. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or research about them. These types of questions are tagged by #Theorical. You are not allowed to use programming language or other technical tools to answer theorical problems.

Report is the key. All students' explanations, solutions, results, discuss and answers must be compacted into a single pdf report. A clean and explicit report is expected and may followed by extra pts; so, you may need to write any related detail or experience during the solving problems. Report file should started within a cover page that it includes course and assignment information as well as identical details like name, student number and email address. Second page should be table of contents that indicates student`s answer to each question. Please repeat your name and student number in left side of footer in other pages. Also, you are free to write in Persian or English. If typing is bothering you, so write in a paper and put its picture with acceptable readability quality in report file.

Organize the upload items. Students should upload their implementation source codes as well as results and report. You should upload a single .zip file with the following structure:

ML_02_[std-number].zip
- Report
  - ML_02_[std-number].pdf
  - [other material and results]
- Source codes
  - P[problem-number]_[a-z].py
  - P[problem-number]_[a-z].ipynb
  - …

Python is the power. Students are free to use any programming language like python, matlab, C++ , etc. However it is recommended strongly to use python in jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

Feel free to contact. If you have any question or suggestion, need guide or any comment be comfortable to ask via email as well as Telegram group.

Problem 1: why and how (30 pts)

#Theorical

a) K-NN classification is based on distance and neighbor calculation. Three important distance metrics are Euclidean, CityBlock, and Cosine. How they work? Give explanation. In K-NN classification with R^2 domain, is it possible that selected neighbor according different distance metrics will be different? Why? Give examples.

b) Assume that you are given face recognition task with face's vector as dataset; which one of above distance metrics should be chosen? Why?

c) K-NN and Decision Tree algorithms can be used for regression task. How they work?

d) In Decision Tree classification, describe "pruning" term and explain why it be used?

e) Consider to below table, apply Decision Tree classification with ID3 algorithm and plot its graph. Indicate all calculation steps as well as used formulas. Which feature has most separability?

#ID	F1	F2	F3	F4	F5	F6	Target
1	+	-	+	+	+	-	1
2	+	-	-	+	+	-	1
3	-	+	-	+	+	+	0
4	+	-	-	+	-	+	1

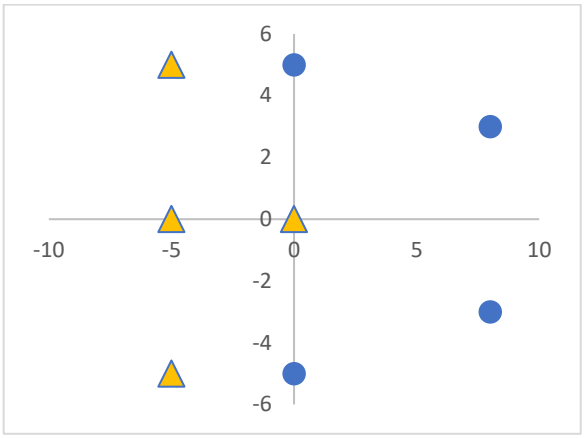
f) Add below sample to above dataset; then, repeat previous part.

#ID	F1	F2	F3	F4	F5	F6	Target
5	+	-	+	-	+	-	0

g) Suppose it takes 1 hour to train a Decision Tree for a dataset with 1 million samples, roughly how much time will it take to train 10 million instances? Justify your answer.

h) We have a dataset of size N, where each instance consists of M binary features. A Decision Tree trained using this dataset, how many leaves has at most?

i) In classification problem of the figure below by 1-NN method, with distance metric $d_1(x, y) = \max_i |x_i - y_i|$ and $d_2(x, y) = \sum_{i=1}^{D} |x_i - y_i|$, what is the result of classification of point $[5, 0]^T$? repeat 4-NN for $[1, 1]^T$ and $[-2, 1]^T$?



j) Evaluation metrics like accuracy or F1-Score and be reported in Macro and Micro stages. Research about their impacts and explain what is different.

Problem 2: Moons (30 pts)

Implementation

Train a Decision Tree for the moons dataset by below steps: make_moons function from sklearn library generates 2D binary classification datasets that are challenging to certain algorithms e.g. centroid-based clustering or linear classification, including optional Gaussian noise. Also, they are useful for visualization.

a) Use make_moons(n_samples=10000, noise=0.4) to generate a moons dataset. Plot the obtained dataset.

b) Use train_test_split() to split the dataset into a train and test sets. (test size = 0.2)

c) Explain how "grid search" works then use it with cross-validation to find good hyperparameter values for a Decision Tree classifier; Report best estimator. Also, you are allowed to use built-in functions in sklearn (**Hint**: try various values for max_leaf_nodes and max_depth)

d) Train it on the full training set using these hyperparameters, and measure your model's performance on the test set. Report accuracy and confusion matrix.

e) Generate 1,000 subsets of the training set with make_moon which, each one is containing 100 instances selected randomly.

f) Train one Decision Tree on each subsets, using the best hyperparameter. Evaluate these 1,000 Decision Trees on the test set. (**Hint**: you can use sklearn.base.clone for this part.)

Since they were trained on smaller sets, these Decision Trees will likely perform worse than the first Decision Tree, achieving only about 80% accuracy.

g) For each test set instance, generate the predictions of the 1,000 Decision Trees, and keep only the most frequent prediction (you can use SciPy's mode() function for this). This approach gives you majority-vote predictions over the test set. (These methods are same to ensemble approaches that will be considered in next assignment)

h) Evaluate these predictions on the test set and compare the accuracy of this model with the model in part d.

Problem 3: Lung Concern (20 pts)

Implementation

Lung concern detection based on clinical reports is one of the primary tasks considered last decade. You are given a dataset that includes 15 clinical features for 309 people who visited Shahid Modares hospital in Shahrivar 1401. In this problem, you are asked to implement the ID3 Decision Tree to classify whether people have lung concerns.

a) Implement a Decision Tree and train a model with 25% of the dataset selected randomly; then, report the obtained tree (graph information like depth, leaves, etc.), accuracy, and confusion matric for the test set. (Note: You are only allowed to use "Numpy," "Pandas," and "math" libraries)

b) Repeat part (a) seven times and report mean accuracy; why are the results different? (In each model, select the training set randomly and independent of others)

c) Repeat parts (a-b) for a situation the training set is 45%, 65%, and 85% of the dataset. In your report, discuss the training set`s size effects on obtained tree`s graph and model performances or not.

d) Now, select 75% of the dataset for the training set and the remaining for testing. Implement pruning operation on the test set. Plot a line chart. The Y-axis is the accuracy measure, and X-axis is the max depth of the tree. Discuss obtained chart. (Hint: you can assume a threshold for deciding whether it is valuable to keep a node. The threshold can be applied on one or many metrics like the model`s accuracy in un/pruned states)

e) [5 Extra Point] Repeat parts (a-d) using sklearn built-in functions and compare them with your results; Discuss differences.

Problem 4: Optical Recognition of Handwritten Digits (20 pts)

Implementation

In this problem, K-NN classification will be used to recognize the handwritten digits. Load dataset by sklearn "load_digits" function. (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits)

a) Load the dataset; then, shuffle it. Select 16 random sample from dataset and show them in one 4*4 plot with related label. (Note: each sample is a vector from R^{64}; only for plotting part you should convert it to a 8*8 matric that indicating related gray-scale image)

b) Implement K-NN classification algorithms with two distance metrices (Euclidean and Cosine).

c) Split dataset into train and test sets with ratio of 8:2; using Elbow technique, find optimal K for each distance metric. Report Elbow charts and optimal Ks. Discuss about obtained results.

d) Which distance metric has better performance? Why? Justify your answers.

e) Select a random sample from test set and plot its neighbor digits based on obtained optimal K for both metrics. Discuss about selected neighbors based on different metrics.

f) Report confusion matrices for optimal K and both metrics.

g) Based on obtained confusion matrices, for optimal K and Euclidean distance calculate TP, TN, FP, FN, F1-score. (First, assume that number 8 is true; second, assume that number 3 is true)

h) Repeat previous part for Cosine. (First, assume that number 6 is true; second, assume that number 4 is true)