Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course By

Dr. Nazerfard

CE5501 | Spring 2023

Teaching Assistants

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

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Ali Amirian (ali.amiryan@aut.ac.ir)

Ghazaleh Gholinejad (Ghazaleh.gholinejad@aut.ac.ir)

Assignment (2)

Outlines. In this assignment, Preprocessing Dataset and Decision Tree are noticed (5 Questions).

Deadline. Please submit your answers before the end of April 14nd 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 10 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:

```
AML_02_[std-number].zip
        Report
                AML_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 (29 pts)

#Theorical

a) What are the challenges associated with **imbalanced** datasets, and how can they impact the performance of machine learning models? Additionally, what preprocessing techniques can be employed to address these challenges and improve the model's performance?

b) How does **data normalization** impact the performance of machine learning models, and what are some common techniques for scaling data to a similar range? how can the choice of normalization technique affect the results of a machine learning model, and what factors should be considered when selecting a normalization technique for a particular dataset?

c) How can preprocessing techniques be used to transform **categorical variables** into numerical features, and what are the implications of different encoding methods on the accuracy of machine learning models? Additionally, how can the choice of encoding technique affect the computational complexity and interpretability of a machine learning model, and what factors should be considered when selecting an appropriate encoding technique for a particular dataset?

d) Is the decision tree a parametric model or non-parametric? Why?

e) How do you choose the appropriate splitting criteria for a decision tree?

f) Consider the following table, apply decision tree classification with ID3 algorithm and plot it. Show all the steps as well as the formulas used. Which of the following features is the most important one?

Id	F1	F2	F3	F4	F5	Target
1	-	-	-	+	+	1
2	-	+	-	+	+	1
3	+	-	-	+	-	0
4	+	+	+	+	-	1

g) What is pruning in the decision tree? Does applying pruning improve the results or not? Try for the above part.

h) Can decision trees be used for multi-class classification problems?

i) What is the difference between information gain and gain ratio in decision tree algorithms?

j) How can decision trees be used for multi-output regression problems?

k) What is the difference between CART and C4.5 decision tree algorithms?

l) What is the difference between decision trees and rule-based systems?

m) What is the effect of tree depth on the bias-variance tradeoff in decision trees?

Problem 2: Cancer | Preprocessing Dataset (20 pts)

# Implementation

Cancer is a major health concern worldwide, and it is one of the leading causes of death globally. According to the World Health Organization (WHO), cancer is responsible for an estimated 9.6 million deaths worldwide in 2018, making it the second leading cause of death after cardiovascular diseases. The cancer dataset is a well-known dataset that is often used in machine learning research. It contains measurements of various features of breast mass samples that can be used to predict whether a mass is benign (non-cancerous) or malignant (cancerous). The dataset was originally compiled by Dr. William H. Wolberg. The dataset contains a total of 569 instances, with 212 instances corresponding to malignant masses and 357 instances corresponding to benign masses. cancer.csv has been provided in the exercise file, which is a collection of personal information and the likelihood of each person's cancer between 0 and 1. Some columns have been removed from this dataset for various reasons, and of course, many data are incomplete. In this exercise, we ask you to perform the following tasks on the dataset in the best possible way:

a) Load the data and convert the cancer probability from a number to an optimal 5-category. Explain how you do this.

b) Describe the problems with the height column data and try to implement the best solution for it and apply it to the data.

c) Remove the rows that have empty or invalid data in the current_smoking column.

d) Fill the remaining empty or invalid data in other columns with the mean of all data.

e) Remove additional columns.

f) Divide the data into two parts of training and testing (70/30) and standardize the data.

g) Train the data using the knn algorithm with k=5 and classify it.

h) Obtain the Confusion matrix, R2_Score, and Accuracy.

i) Try another Preprocessing methods to improve results.

Problem 3: Use ML in your Field| Preprocessing Dataset (11 + 10 pts)

# Implementation

One of the goals of the machine learning course is for you to be able to apply the methods you have learned in your field. One of the most important tasks of a machine learning engineer is to be able to pre-process the dataset well and prepare it for applying machine learning and neural network methods. In this question, you need to find a data set related to your field and apply the pre-processes reviewed in the course on this data set. This data set should contain at least 70 data.

a) Describe the data set and explain its characteristics

b) Load the selected data set and display the first seven data.

c) Apply the necessary pre-processing on the dataset.

d) Teach a model using one of the methods learned in the lesson.

e) Evaluate the model.

Problem 4: How much will my score be? |Preprocessing & Decision Tree (20 pts)

# Implementation

In this question, we want to predict the passing or failing status of the students in the exam by using the six features that we have from the students of a class and check whether these features can be used to predict the passing or failing status or not. You can see the information of the students in the Tatt file.

a) Load the dataset.

a) Preprocess the data as necessary (e.g. handle missing values, encoding categorical variables,..).

b) Split the data into training and test sets. Which of the methods presented in the class is better for division?

c) Initialize a decision tree classifier object with default parameters.

d) Fit the classifier to the training data.

e) Evaluate the performance of the model on the test data, using metrics such as accuracy, precision, recall, and F1 score.

f) Visualize the decision tree using a library such as Graphviz or scikit-learn's plot_tree function.

g) Experiment with different hyperparameters of the decision tree (e.g. max_depth, min_samples_split) to see if you can improve the performance of the model.

h) Use grid search to find the best hyperparameters for the decision tree.

i) Evaluate the performance of the best decision tree on the test data. Can the decision tree determine the pass or fail status?

Problem 5: Passengers of the Titanic | Decision Tree (20 pts)

# Implementation

Titanic dataset is another well-known dataset that contains information about passengers on the Titanic and whether or not they survived.

Suppose you are given the Titanic dataset and are tasked with building a decision tree model to predict the survival of passengers on the Titanic. Using scikit-learn library, answer the following questions.

a) Load the Titanic dataset into a Pandas dataframe. What preprocessing steps do you think is needed for this data? Apply them to the data.

b) Split the dataset into training and testing sets using a 80/20 split ratio.

c) Perform hyperparameter tuning to optimize the performance of the decision tree model. what hyperparameters would you tune?

d) With parameters you got in previous part train a decision tree and report the accuracy, precision, recall and F1score on test dataset.

e) Visualize the decision tree using Graphviz or Matplotlib. what insights can you gain from the visualization?

f) Perform pruning on the decision tree model to prevent overfitting.

g) compare the performance of the pruned decision tree model to the unpruned model, and what insights can you gain from the comparison?