Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course By

Dr. Nazerfard

CE5501 | Spring 2023

Teaching Assistants

Mohamadreza Jafaei (HEAD) (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 (1)

Outlines. In this assignment, Regression and KNN are noticed (5 Questions).

Deadline. Please submit your answers before the end of March 17rd 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_01_[std-number].zip
- Report
  - AML_01_[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 (26 pts)

#Theorical

a) Can KNN be used for multi-class classification problems? If yes, how?

b) How does the curse of dimensionality affect the performance of KNN?

c) How can KNN be used in unsupervised learning tasks?

d) How does KNN differ from other classification algorithms, such as Naive Bayes and Decision Trees?

e) What is the effect of the choice of distance metric on the performance of KNN? Describe the characteristics of the Euclidean, Manhattan, and cosine distance metrics, and explain when each metric might be appropriate to use in KNN.

f) Describe the process of selecting the optimal value of K in KNN. What are some methods for selecting K, and what are the advantages and disadvantages of each method? How does the choice of K affect the bias-variance tradeoff in KNN?

g) Explain the concept of locally weighted regression in KNN. How does locally weighted regression differ from standard KNN regression, and what are the advantages and disadvantages of using locally weighted regression in KNN? What are some applications of locally weighted regression in machine learning?

h) Suppose you are given the following training set:

Instance	Feature1	Feature2	Class
1	2	4	A
2	3	5	A
3	4	6	A
4	5	7	B
5	6	8	B
6	7	9	B

And the following test instance:

Instance	Feature1	Feature2
7	3	4
8	6	5

Using the KNN algorithm with K=3 and the City Block distance metric, predict the class of the test instance. Show your work.

i) What is the difference between a model parameter and a learning algorithm's hyperparameter?

j) Can you name four of the main challenges in Machine Learning?

k) If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

l) What can go wrong if you tune hyperparameters using the test set?

m) Suppose you are using Polynomial Regression. You plot the learning curves and you notice that there is a large gap between the training error and the validation error. What is happening? What are three ways to solve this?

n) Why would you want to use:
I. Ridge Regression instead of plain Linear Regression (i.e., without any regularization)?
II. Lasso instead of Ridge Regression?
III. Elastic Net instead of Lasso?

Problem 2: How much does my laptop cost | Regression **(19 pts)**

Implementation

In this question, train a **polynomial** regression to predict the price of laptops using their features. For this purpose, do the following steps (**All the steps below must have a report**):

a) Load CSV file of Laptop Price and show 10 random item

b) Data visualization (at least 5 plots. Interpret the plots)

c) Dataset cleaning and Feature Engineering

d) Data preparation for ML (e.g., encoding, scaling, train/test/validation split)

e) Train the model with K-fold cross-validation for choose the best degree. Draw a plot to show the changes in MSE with increasing regression degree on the validation data.

f) Train the final model with best degree and report MSE and R2 on test data.

g) In your opinion what is the most important feature? Why?

Problem 3: We Need Fresh Air | Regression Help Us (18 pts)

Implementation

Download Air Quality Dataset[1]. Describe this dataset. Here we want to test different types of regression models to predict "PT08.S1(CO)". For this purpose, do the following steps (**All the steps below must have a report**):

a) Data visualization (at least 5 plots. Interpret the plots)

b) Dataset cleaning and Feature Engineering.

c) Data preparation for ML (e.g., scaling, train/test/validation split)

d) Train Ridge, Lasso and Elastic Regression. (Find the best hyperparameters for each model by trial and error. (Alpha for Ridge and Lasso. Alpha and l1_ratio for Elastic).

e) Train the three final model with best hyperparameters and report MSE and R2 on test data.

f) Compare all three models. Which model has worked better? Why?

[1] https://archive.ics.uci.edu/ml/datasets/air+quality

Problem 4: Handwritten Digits Detection | KNN (19 pts)

Implementation

Handwritten digits dataset is used for image classification tasks and consists of 8x8 pixel images of handwritten digits from 0 to 9. The dataset consists of 1797 instances, where each instance is an image of a digit and the goal is to predict the digit in the image. This dataset are readily available in scikit-learn, and you can load it using the 'load_digits' function (**All the steps below must have a report**).

a) Load the Digit dataset from scikit-learn and split it into training and test sets using a 70/30 split. Use a random seed of 42 for reproducibility.

b) Preprocess the data by scaling the pixel values to the range [0, 1] using the MinMaxScaler from scikit-learn.

c) Train a KNN classifier on the training set using K=5 and the default distance metric. Evaluate the performance of the classifier on the test set by computing the accuracy, precision, recall, and F1-score. You can do it with 'classification_report' function.

d) Use 10-fold cross-validation to evaluate the performance of the KNN classifier with K=5 and the default distance metric. Compute the average accuracy, precision, recall, and F1-score, along with their standard deviations.

e) Plot a graph that shows how the average accuracy of the KNN classifier varies with different values of K. Use 10-fold cross-validation and the default distance metric. Choose K to range from 1 to 30. Determine the optimal value of K based on the elbow method.

Problem 5: Iris flowers | KNN (18 pts)

Implementation

Iris flowers is a classic dataset for classification tasks and is often used to demonstrate machine learning algorithms, including KNN. The dataset consists of 150 instances with 4 features, where each instance is a measurement of an iris flower, and the goal is to predict the species of the flower. You can load this dataset using 'load_iris' function.

You want to use KNN to classify new instances of iris flowers, and you want to determine the optimal value of K for this task. Answer the following questions (**All the steps below must have a report**):

a) Load the Iris dataset from scikit-learn and split it into training and test sets using a 80/20 split. Use a random seed of 42 for reproducibility.

b) Preprocess the data by scaling the features to have zero mean and unit variance using the StandardScaler from scikit-learn.

c) Train a KNN classifier on the training set using different values of K ranging from 1 to 20, and evaluate the performance of the classifier on the test set by computing the classification accuracy for each value of K. Choose the optimal value of K based on the accuracy results.