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 (4)

Outlines. In this assignment, SVM, Ensemble Models and Clustering are noticed (6 Questions).

Deadline. Please submit your answers before the end of July 1 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_04_[std-number].zip
        Report
                AML_04_[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 (20 + 4 pts)

#Theorical

a) How would you find the optimal number of random features to consider at each split?

b) What are the trade-offs between the different types of Classification Algorithms? How would do you choose the best one?

c) How does Ensemble Learning tackle the No-Free Lunch Dilemma?

d) Can you use the LASSO method for Base Learner Selection?

e) What's the similarities and differences between Bagging, Boosting, Stacking?

f) What is a dual and primal problem and how is it relevant to SVMs?

g) Can an SVM classifier outputs a confidence score when it classifies an instance? What about a probability?

h) If you train an SVM classifier with an RBF kernel. It seems to underfit the training dataset: should you increase or decrease the hyper-parameter γ (gamma)? What about the C hyper-parameter?

i) What technique can be used to solve the optimization problem cast by Support Vector Machines?

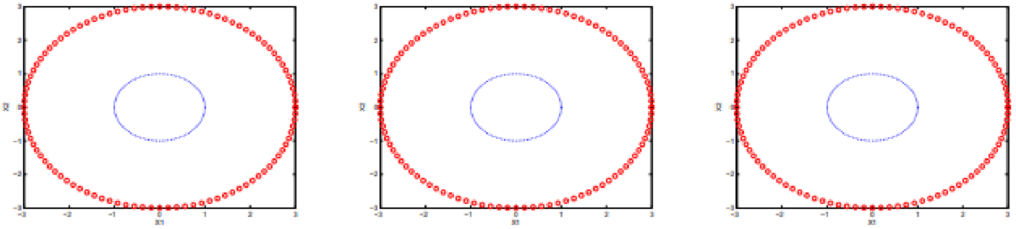
j) How probabilities are calculated for SVM model?

k) In class we learnt that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $K(x, z) = \varphi(x)\ T\ \varphi(z)$, where $\varphi(x)$ is a feature mapping. Let K1 and K2 be Rn × Rn kernels, K3 be a Rd × Rd kernel and $c \in R+$ be a positive constant. $\varphi1 : Rn \rightarrow Rd$, $\varphi2 : Rn \rightarrow Rd$, and $\varphi3 : Rd \rightarrow Rd$ are feature mappings of K1, K2 and K3 respectively. Explain how to use $\varphi1$ and $\varphi2$ to obtain the following kernels.

   a. $K(x, z) = cK1(x, z)$

   b. $K(x, z) = K1(x, z)K2(x, z)$

l) You are given the following 3 plots, which illustrates a dataset with two classes. Draw the decision boundary when you train an SVM classifier with linear, polynomial (order 2) and RBF kernels respectively. Classes have equal number of instances.



Problem 2: Custom Dataset | SVM (10 pts)

# Implementation

a) Use the make_classification function to generate a custom dataset. split the dataset into 80% training data and 20% validation data.

b) Get the best parameters for the model using svm and grid search.

c) What is mkl and explain its parameters and types and how does it differ from using a single kernel in SVM?

d) Model your dataset with mkl and get the desired outputs such as precision, recall, etc.

Problem 3: Credit Card Fraud | SVM (15 pts)

# Implementation

"Suppose you have a dataset called 'Credit Card Fraud' that consists of financial transactions. This dataset contains both normal transactions and fraudulent transactions. It contains a total of 10,000 transactions, where 99% of the transactions are normal and 1% are fraudulent. Your task is to train an SVM algorithm on this dataset to detect fraudulent transactions.
Dataset: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

a) Load data set and split the dataset into 80% training data and 20% validation data then Implement the SVM algorithm on the training data and tune the parameters to achieve the best performance
b) What is one-class SVM and how can it be used in outlier detection? run the one-class SVM on training data. (Set the nu, kernel and gamma).
c) Make outlier predictions on the validation dataset. Which of the metrics such as precision, recall, etc. can show how accurately our model was able to detect outlier data? Explain your reasons.

Problem 4: Data Science Salaries | Ensemble & Decision Tree (20 pts)

# Implementation

Consider the dataset Data Science Salaries 2023[1]. In this question, you have to train the data using a decision tree with depth 2, then train the data that is in the leaf using any other arbitrary algorithm. Follow the steps below.

a) Data preprocessing.
b) Extract useful information using data visualization and considering the “job_title” column.
c) split the dataset into train and test.
d) Train the model.
e) Report model accuracy.
f) Explain how you implemented this model.

Problem 5: Price Pattern Extraction | K-means (20 pts)

# Implementation

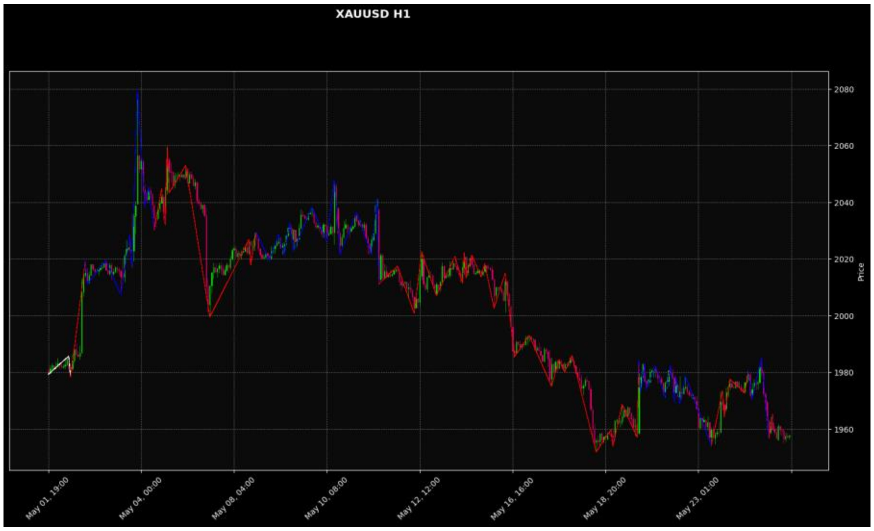
The Zigzag indicator is a technical analysis tool used to identify significant changes or reversals in price trends. It helps traders and analysts visualize the most prominent price swings in a financial time series, filtering out smaller fluctuations and noise.

The Zigzag indicator connects significant highs and lows in a chart by drawing lines or segments, highlighting the major price movements. It aims to simplify price analysis by focusing on the turning points or pivot levels that indicate potential trend changes.

In this problem, our objective is to utilize the K-means algorithm for extracting price patterns from the Zigzag indicator. We have already been provided with the implementation of the Zigzag indicator, which we will use exclusively for pattern extraction. The output of this indicator is visualized in the form of shape (a). In this implementation we also specified the trend for each leg.

[1] **https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023**

a) The Zigzag lines included start and end point of each leg and each point contains its date and price. What do you think is a better way to represent a leg that is fine to be used in k-means algorithm? Implement it and use it for the next parts of problem.
b) Using Zigzag lines create three datasets that contain the previous 2, 3 and 4 legs. (The trend of current leg should be included)
c) Run K-means algorithm for each dataset. You should split data for each trend and find 5 clusters for each one. Plot the patterns you found (There should be 6 different sets of patterns).
d) Show the path of centers movement only for dataset that contains 2 legs. (You can use a 2D chart because there's only two features in this dataset)
e) Find the best hyperparameter for K-means using the elbow method. Extract and plot the patterns you found with the new number of clusters.
f) Is there any difference between bullish and bearish patterns that you found?



(a)

Problem 6: Synthetic Clusters | DBSCAN (15 pts)

# Implementation

a) Write a command using sklearn.datasets.make_blobs to generate synthetic clusters and standardize the data.
b) Provide the command to create a scatter plot to visualize the generated data.
c) Implement the DBSCAN algorithm using sklearn.cluster.DBSCAN on the standardized data.
d) Write the command to access the resulting cluster labels and calculate the estimated number of clusters and noise points.
e) Write a command to evaluate the clustering performance using various metrics such as homogeneity, completeness, V-measure, adjusted Rand Index, adjusted mutual information, and silhouette coefficient using sklearn.metrics on the true labels and the DBSCAN cluster labels.