Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course

By Dr. Nazerfard CE5501 | Spring 2024

Teaching Assistants:

Amir Hossein Babaeayan

Assignment (3)

Outlines. In this assignment, some practical implementation skills which needed in this and other courses of this degree are noticed as well as regression topics. Remember that you may need to re-use your implementations of this assignment; so, it is suggested to code in functional.

Deadline. Please submit your answers before the end of date 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 4 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you lose 20% of the points of that assignment. After 4 days you miss all points and any submission will not be 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 for you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then reasoned about. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or researched about. These types of questions are tagged by #Theoretical. You are not allowed to use programming language or other technical tools to answer theoretical problems.

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

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

ML_03_[std-number].zip
- Report
  - ML_03_[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 strongly recommended to use python in the jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

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

Problem 1: Breast Cancer Prevention using K-Means Algorithm

Implementation

Breast cancer, a globally prevalent form of cancer, accounted for 12.5% of new cases in 2020. Early detection is possible through regular screenings and tests, despite its severity. Fine Needle Aspiration (FNA) is a reliable method for identifying breast cancer. This method involves extracting a small tissue sample with a syringe, followed by imaging. Clinicians isolate individual cells from each image to extract 30 characteristics, such as size, shape, and texture.

The objective is to use K-Means clustering to diagnose breast cancer based on features extracted from the Wisconsin Diagnostic Breast Cancer dataset.

Tasks

a. Implement a Python class, KMeansCluster, for the K-Means clustering algorithm.

The class should include:
- An initialization method (__init__) that takes the number of clusters (k), convergence tolerance (tol), and maximum iterations (max_iter).
- The following methods:
  1. fit: Fits the model to the input data matrix (X) and the initial centers (mu) iteratively.
  2. accuracy: Calculates clustering accuracy.
  3. predict: Predicts cluster labels for new data.
  4. sse: Computes the sum of squared errors (SSE).

```python
def __init__(self, k, tol, max_iter):
    # Constructor implementation
```

The clustering result should be stored in matrix C.
Use this class to cluster the data and report the accuracy of the clustering.

b. Run the algorithm 5 times using different starting points.

Record the clustering accuracy for each run.
Discuss the observations and variations in the results, if any.

c. Use the provided initial centres (init_mu) to initialize K-Means.

Each column in init_mu represents one initial center.
Report the accuracy of the clustering.

d. What happens when initialized with the true centres?

Analyze the results obtained when initializing with the true cluster centers derived after clustering.

e. Explore alternative methods for better accuracy.

1. Investigate other unsupervised learning methods.
2. Evaluate the use of a supervised learning method.

Compare the performance of these approaches to K-Means clustering.
Implement and explain any improvements.

Problem 2: Airplane crash

There has been an airplane crash recently. A lot have lost their lives, but some have survived. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

The dataset provided along with the assignment document contains passenger records. We need you to build a classifier that could helps us answer the question: “what sorts of people were more likely to survive?”

a) Load the dataset and prepare it for model training.

b) Split the data into 80% train and 20% test dataset.

c) Train an SVM model and report accuracy, precision and recall scores.

d) Use grid search with cross validation to extract the best hyper parameters.

e) Report your findings along with the reason behind hyper parameter value.

f) Train another SVM model with the best hyper parameters and report accuracy, precision and recall.

Problem 3: Credit Card

Some credit card companies have designed methods using algorithms that, based on purchase data, determine whether a purchase was made by the original cardholder or if it is fraudulent. The dataset is located at:
https://drive.google.com/file/d/1sq18MxrJKh1KbjeSo3OKzmlhD3puUD37/view?usp=sharing

This dataset contains information about purchases made by citizens over two days. If the purchase is valid, a zero is placed in the last column; otherwise, a one is placed.
In total, there were 284,807 valid purchases and 492 invalid purchases. Columns v1, v2,... represent the values of variables obtained after dimension reduction using PCA. We intend to use these columns along with SVM algorithms that you have learned to predict whether a purchase is fraudulent or legitimate.

a) Theoretically, based on the proportion of fraudulent purchases, determine the accuracy when considering all outputs as zero.
b) Considering part A and the dataset information provided to you, determine the importance of false positives and false negatives in evaluating the classifier.
c) Split the data into two sets—training and testing—using an 80:20 ratio.
d) Classify the training data using SVM (you can use other SVM types you've learned, such as SVM with kernels, in addition to linear SVM).
e) Evaluate the test data based on the classifier created and measure the accuracy of your classification. Pay particular attention to part B to ensure that the evaluation metrics used are practical and suitable.
f) What methods exist to increase accuracy in the invalid class? How does this affect the accuracy of each class? Does this method help reduce computations?

Problem 4: Diabet

Load the “pima_indians_diabets.csv” dataset that is in the folder of exercise and classify the data using below models.

a) with at least 3 values for the below parameters, train the random forest classifier on the dataset and report the accuracy on the train and test dataset and specify the best model. (n_estimators, max_features, max_depth)

b) analyze the effects of the parameters on the model performance.

c) use any of the ensemble methods and try to achieve a better accuracy on the test dataset. (3 models is enough. But it’s better to achieve a better accuracy than the previous section model.)

Problem 5: More Into Clustering

In this section, our objective is to implement the DBSCAN class for clustering tasks and apply this algorithm to the datasets "pathbased-D31-spiral-Compound." The requirement is to refrain from using any library that implements DBSCAN. After completing the implementation, we will address the following questions:

a) Calculate the purity criteria and determine the number of clusters achieved using the implemented DBSCAN class.

b) Visualize and show the data related to each cluster with a different color. Note that some data may not belong to any cluster and may be considered noise. Show noisy data with different color.

c) What effect does the types of datasets have on the performance of the DBSCAN algorithm?

Problem 6: Image Compression

The internet is filled with huge amounts of data in the form of images. People upload millions of pictures every day on social media sites such as Instagram, and Facebook and cloud storage platforms such as google drive, etc. With such large amounts of data, image compression techniques become important to compress the images and reduce storage space. In this article, we will look at image compression using the K-means clustering algorithm which is an unsupervised learning algorithm. An image is made up of several intensity values known as Pixels. In a colored image, each pixel is of 3 bytes containing RGB (Red-Blue-Green) values having red intensity value, then Blue and then green intensity value for each pixel.

Approach: K-means clustering will group similar colors together into ‘k’ clusters of different colors (RGB values). Therefore, each cluster centroid is representative of the color vector in the RGB color space of its respective cluster.

Your task is to compress the tiger image with k-means clustering and then show the output image in the doc.

In the doc, explain the method and steps took for this task.