HW5 SPR Fall 2022

Deadline 14 Bahman

To do this homework, you must implement K-means, GMM, and SVM. You should go through these concepts.

K-means

Dataset: "tsnv.csv, blobs.csv, elliptical.csv, moon.csv, circle.csv"

- K-means is a popular machine learning and data mining algorithm that discovers potential clusters within a dataset. Finding these clusters in a dataset can often reveal interesting and meaningful structures underlying the distribution of data. K-means clustering has been applied to many problems in science and still remains popular today for its simplicity and effectiveness.

- You are given 5 datasets to see whether k-means can cluster them or not.
  - You are asked to perform a k-means clustering on all given datasets using the Euclidean distance as the distance function. Which of the datasets is clustered better with k-means?
  - How do we choose K? In this problem, we will investigate various ways of evaluating the quality of a clustering assignment.
    - **1.** Use the elbow method to evaluate the best choice of the number of clusters, plotting the total within-cluster variation against the number of clusters for k-means clustering with $k \in (1, 2, \ldots, 15)$.
    - **2.** Use the average silhouette to evaluate the choice of the number of clusters for k-means clustering with $k \in (1, 2, \ldots, 15)$. Plot the results.(You can use library for this part)
  - Plot the clusters for each dataset after you performed the k-means algorithm on them with different k's.
  - After analyzing the plots produced by elbow method and silhouette method, discuss the number of clusters that you feel is the best fit for each of the given datasets. Defend your answer with evidence from these two parts and their produced plots, and what you surmise about these datasets.

GMM

Dataset: "tsnv.csv, blobs.csv, elliptical.csv, moon.csv, circle.csv"

- The Gaussian mixture model (GMM) is well-known as an unsupervised learning algorithm for clustering. Here, “Gaussian” means the Gaussian distribution, described by mean and variance; mixture means the mixture of more than one Gaussian distribution. GMM uses Expectation Maximization (EM) to train a GMM model. A GMM model can be employed to estimate the PDF of some samples (like a parametric density estimator).

  - You are asked to perform a GMM method on these datasets for clustering.
  - Plot the training and testing data (Different colors for each class).
  - Construct a GMM for clustering, with K = 1,5,10 Gaussian components, and train on Train Data.
  - For each k, plot the test and train data classified (clustered) by the GMM algorithm.
  - How do we choose the number of gaussian components? Search various ways of evaluating the quality of a clustering assignment for the GMM algorithm. Explain at least two metrics.
  - Report the best k based on the two metrics you found.
  - Compare the results you've gotten from k-means with GMM. For each dataset determine which method works better? Why?

SVM

Dataset: “Dataset1.mat, Dataset2.mat”

- In this exercise, you will be using support vector machines (SVMs) with various example 2D datasets. Experimenting with these datasets will help you gain an intuition of how SVMs work and how to use a Gaussian kernel with SVMs. (You can use **optimizer** libaraies for this part)

- You will try using different values of the C parameter with SVMs. Informally, the C parameter is a positive value that controls the penalty for misclassified training examples. A large C parameter tells the SVM to try to classify all the examples correctly.
  - Train the SVM using three different values of the penalty parameter (C=1 and C=100 and C=1000) on **“Dataset1.mat”.**
  - Plot the train and test data with the decision boundary and marginal boundary on **“Dataset1.mat”**.
  - Report the train and test accuracies for C=1, 100 and 1000 on **“Dataset1.mat”.**

- In the next part by using the Gaussian kernel with the SVM, you will be able to learn a non-linear decision boundary that can perform reasonably well for the dataset. In general, SVM is a linear classifier. When data are not linearly separable, Kernel SVM can be used. Here, you will utilize SVM with RBF kernel for non-linear classification. Perform the following step for **“Dataset2.mat”** and **“Dataset1.mat”** datasets.
  - Train SVM with the penalty parameter C and the standard deviation σ for RBF kernel. Your task is to use the ten-time-ten-fold cross-validation set to determine the best C parameter to use.

  - In the next part you should find the best value for both C and σ simountinously, we suggest trying values in multiplicative steps (e.g., 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30). Note that you should try all possible pairs of values for C and σ (e.g., C = 0.3 , and σ = 0.1). For example, if you try each of the 8 values listed above for C and for σ, you would end up training and evaluating (on the cross-validation set) a total of 8^2 = 64 different models to select the best model.
  - Plot train and test accuracies and their corresponding variances of ten-time-ten-fold cross-validation for different values of C and σ.
  - Plot the train and test data and the non-linear decision boundary for both datasets (for the best model)
  - Report the train and test accuracies using the selected model for both datasets (best C and σ)boundary

✓ Pay extra attention to the due date. It will not extend.

✓ Be advised that submissions after the deadline would not grade.

✓ Prepare your full report in PDF format and include the figures and results.

✓ Do not use sklearn or any similar library and write your own code.

✓ Use only the python programming languages.

✓ Submit your assignment using a zipped file with the name of:

"FirstName_LastName_ TeammateFirstName _Teammate LastName.zip"

✓ Email your zip file to 'arezoo7697@gmail.com'.

✓ Using other students' codes or the codes available on the internet will lead to zero grades.