Introduction:

In this project, by following the steps that we have prepared for you, we are going to strengthen our understanding in support vector machines and apply the learned knowledge in practice. For this purpose, the MNIST handwritten number dataset is considered. This collection contains images consisting of 10 figures. We are going to create a machine that can finally classify each digit in the corresponding class. Download the MNIST data set in two sets of training and testing (validation) along with the label of each, from this link(MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges). These images are binary coded. By reading the guide on the site, you can download the images in two sets of training and testing and decode.

Project Description

1. Randomly display 10 images of the tutorial along with their labels.

   To reduce the training time of the model, randomly create subsets of 5000 members from the training set and subsets of 1000 members from the test set. You should choose the same number of each class so that the problem of class imbalance does not arise. Consider the new sets the training and test sets, respectively.

2. By searching the internet, study the PCA method and explain how it works. You are supposed to perform feature extraction with this method.

   Consider the following features and prepare for model input.

   a. Data pixel values in full
   b. The features obtained by reducing the dimensions of pixels with the PCA method to 100 dimensions
   c. The features obtained by reducing the dimensions of pixels with the PCA method to 40 dimensions

3. Implement the support vector machine for classification of MNIST images using the Singer Crammer [1] method from scratch. To speed up your work and mental order, you can complete the notebook attached in the project folder. In all cases, set the value of epsilon equal to 0.01.
   a. Considering the given following states and hyperparameters, train the support vector machine and calculate and report the following metrics on the test set and for each state. Enter the completed tables in your report.
      i. Test Accuracy
      ii. Precision
      iii. Recall
      iv. F1_score
      v. AUC
      vi. Confusion matrix

vii. Number of support patterns

4. See the Persian Project Description for this task
5. See the Persian Project Description for this task
6. See the Persian Project Description for this task
7. In each of sections 4 to 6, analyze the effect of each of the hyperparameter values and the results obtained. Finally, compare the results of the three sections.

Consider the best model obtained and answer the following questions according to it.

8. Suppose we want to find 2 numbers, which criterion(s) should we consider? Why?
9. Analyze and compare the number of support patterns in the specified colored houses (houses of the same color).
10. Did you notice the same sensitivity to parameter changes in original dimension data and reduced dimension data? Why? Name the other ways you can present the model with the above generalization.
11. Explain how your Singer Crammer support vector machine is able to recognize and classify digits.
12. Another multiclass classification method is the rest versus one method, which uses the aggregation of binary classifications. Read about this method. State with reasons what interfaces exist between hyperparameter C in this machine and B in the machine using the singer crammer method.
13. Put the hyperparameters related to the best result obtained in the previous sections as the input of the SVM model using the versus one rest method. Report the mentioned criteria for it. Compare the results of the two models. (Use the libraries available in Python language.)
14. Suppose you are asked to perform data cleaning using support vector machine. Explain how you do this. (point section)

[1] K. Crammer and Y. Singer, “On the Algorithmic Implementation of Multiclass Kernel□based Vector Machines,” 2001.

Additional notes:

- The time needed to run each machine will be between 5 and 20 minutes. For better planning in the project, include a day for implementation and recording the results.
- The project must be done individually and group solutions are not allowed
- Provide a complete report of your project in each section. In addition to the implementation of the project, the full score depends on the description of its implementation and the submission of a comprehensive report on the results.
- Submit your answers separately, including the project report in PDF format and the code implemented in Python along with the result cells, in the Python notebook format (ipynb.) in a zipped folder with the PR_Project1_YourStudentID.zip format in the university's LMS system Upload by the specified deadline.
- You can raise your doubts and questions in the following two ways:

- Telegram.me/TBehjat
- Mahdiyeh.bhjt@gmail.com