Boğaziçi University
Department of Computer Engineering
Instructor: İnci Baytaş
Teaching Assistants: Berrenur Saylam, Tolga Kerimoğlu

Spring 2022
Assignment 2

CMPE 462 Project - Phase 2

This assignment consists mainly of programming questions. Familiarity with Python, Jupyter Notebooks, and relevant libraries are assumed, namely **numpy** and **matplotlib**.

For theoretical questions, we expect the answers in LaTeX. One option is to merge your theoretical answers and code into a single Jupyter Notebook where you can use markdown cells to insert LaTeX into the notebook. This is our preferred method of delivery, and we believe it would be easier for you as well. If you want to submit your code and solutions separately, you need to provide .py files with instructions on reproducing the results for each question.

In this phase, you will continue working with the **3D Shapes** dataset provided to you in the first phase.

Question 1. *Logistic Regression and Interpreting Feature Importance*

In this question, you will implement a logistic regression classifier for the **3D Shapes** dataset. Please implement the logistic regression from scratch. You may use the features you extracted in Phase 1.

(a) Before you start training logistic regression, visualize your features using t-SNE (This method is developed by Geoffrey Hinton and Laurens van der Maaten, their research paper is here.) If the dimensionality of your features is higher than three, you may reduce its size to three or two using t-SNE. Visualize your features with a 2-D or 3-D scatter plot where you indicate different categories with different colors. Comment on the discriminability of your features. Do they look linearly separable?

(b) Derive the gradient of the loss function with respect to model vector, $\nabla_{\mathbf{w}} E_{in}$.

(c) Determine the step size of the full batch gradient descent using 5 fold cross-validation.

(d) Report the training and test classification accuracy, which can be calculated as given below:

$$ACC = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}$$

(e) If you observe overfitting, try using $L2$-norm regularization. Determine the value for the regularization parameter by 5 fold cross-validation.

(f) If you observe underfitting, try to extract more discriminative features and repeat the t-SNE plot with the new features.

(g) After obtaining classifier, $\mathbf{w}$, with the best accuracy you could get, investigate the weight values and their corresponding features. Can you interpret the contributions of each input feature?

(h) Implement logistic regression with $L1$-norm regularization from scratch. Logistic regression with $L1$-norm regularization should give you a sparse solution, $\mathbf{w}$. Since $L1$-norm is not smooth, you need to use **soft thresholding**. Learn how to implement soft thresholding and incorporate it into your implementation. Determine the regularizer λ using 5 fold cross-validation. Which attributes do have non-zero weights? Compare your results with the result in (f) regarding feature significance.

Question 2. *Naive Bayes*

Train a Naive Bayes classifier with the features you used in Question 1. In this question, you may use a built-in function from a library.

(a) Report your training and test classification accuracy. Compare the performance with Question 1.

(b) Does conditional independence assumption hold for the features you used in this phase? Elaborate on your answer.

(c) What are the numbers of parameters you need to estimate with and without the conditional independence assumption for this dataset and your features?