CS464 Machine Learning
Spring 2021
Homework 2

Due: April 12, 17:00

Instructions

- For this homework, you may use any programming language of your choice.
- You are NOT allowed to use any machine learning packages, libraries or toolboxes for this assignment (such as scikit-learn, tensorflow, keras, theano, MATLAB Statistics and Machine Learning Toolbox functions, e1071, nnet, kernlab etc.) unless otherwise stated.
- Submit a soft copy of your homework to Moodle.
- Upload your code and written answers to the related assignment section on Moodle (.TAR, .TAR.GZ or .ZIP). Submitting hard copy, handwritten or scanned files is NOT allowed.
- The name of your compressed folder must be "`CS464_HW2_Section#_Firstname_Lastname`" (i.e., `CS464_HW2_1_john_doe`). Please do not use any Turkish characters in your compressed folder name.
- Your code should be in a format that is easy to run and must include a driver script serving as an entry point. You must also provide a README file with clear instructions on how to execute your program.
- This is an individual assignment for each student. That is, you are NOT allowed to share your work with your classmates.
- If you do not follow the submission routes, deadlines and specifications (codes, report, etc), it will lead to significant grade deduction.
- Plots generated for this homework should be properly formatted. That is include title, axis labels and legend (if needed) in your plots.
- **You may ask your questions regarding this homework to your TA, Salman Mohammad (salman.mohammad@bilkent.edu.tr)**

1 PCA & Digits [35 pts]

In this question, you are expected to compress digit [1] images using PCA and decompress it back to restore original images. For this question, you will use digits dataset which is provided to you within homework zip file as `digits.csv` . The first column represents the digit while the other columns represent the features of the digits. Note that you are not allowed to use PCA from libraries. [Hint: Please make your data mean centered before applying PCA]

Question 1.1 [15 pts] Apply PCA to obtain the first 10 principal components. Report the proportion of variance explained (PVE) for each of the principal components. Reshape each of the principal component to a 28x28 matrix and show them. Discuss your results.

Question 1.2 **[5 pts]** Obtain first k principal components and report PVE for $k \in \{8, 16, 32, 64, 128, 256\}$. Plot k vs. PVE and comment on it.

Question 1.3 **[15 pts]** Describe how you can reconstruct an image using the principal components you obtained in question 1.1. Use first k principal components to analyze and reconstruct the first image in the dataset where $k \in 1, 3, 5, 10, 50, 100, 200, 300$. Discuss your results.

2 Linear & Polynomial Regression [30 pts]

Dataset

Your dataset contains information of house prices in King County from [2]. Our aim is to understand which features are responsible for higher property value. There are 4 features provided. You will use the following files:

- `question-2-features.csv which contains 4 features.`
- `question-2-labels.csv which contains the ground truth i.e the total price of the house`

Question 2.1 **[4 pts]** Derive the general closed form solution for multivariate regression model using ordinary least squares loss function given in Eqn. 2.1. Briefly explain each matrix involved in calculation and how they are constructed.

$$J_n = ||y - X\beta||^2 = (y - X\beta)^T(y - X\beta) \tag{2.1}$$

Question 2.2 **[4 pts]** Find the rank of X^TX for the given dataset using built-in library functions of your language (rank() for MATLAB,numpy.linalg.matrix rank() for numpy etc.). What does the rank tell you about the solution you have found for Question 2.1.

Question 2.3 **[10 pts]** Using the formula you have derived for Question 2.1, train a linear regression model by using only "sqftliving" feature which is given to you in the dataset. Consider all of the dataset as training set. Report and interpret the coefficients of the trained model. In addition, plot price vs. "sqftliving" along with your model's predictions on the same plot. Finally, calculate MSE using your model's predictions and ground truth labels. Comment on your results.

Question 2.4 **[12 pts]** In this part, again assume that you are only provided with "sqftliving" as a feature. You will use polynomial regression to train your model. In this part, you will be using the feature x_i and its powers x_i^2. In addition, plot the graph of price vs. "sqftliving" along with your model's predictions on the same plot. Finally, provide the coefficients and the training MSE for your model. Assume the whole dataset is your training set. Comment on your results.

3 Logistic Regression [35 pts]

For this part of the question, your dataset is a subset of the credit card fraud detection dataset [3]. The dataset contains only numerical input variables which are the result of the PCA transformation, i.e. features V1, V2, ... V28 are the principal components obtained from PCA. The only feature which has not been transformed with PCA is Amount. The Class column is the response variable and it takes value 1 in case of fraud and 0 otherwise.

You will use the following files:

- `question-3-features-train.csv`
- `question-3-labels-train.csv`
- `question-3-features-test.csv`
- `question-3-labels-test.csv`

Question 3.1 **[15 pts]** You will implement full batch gradient ascent algorithm to train your logistic regression model. Initialize all weights to 0. Try different learning rates from the given logarithmic scale [10^-5, 10^-4, 10^-3, 10^-2, 10^-1] and choose the one which works best for you. Use 1000 iterations to train your model. Report the accuracy and the confusion matrix using your model on the test set given. Calculate and report averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores. (Hint: Check the Amount feature. It needs to be normalized before moving to logistic regression)

Question 3.2 **[15 pts]** You will implement mini-batch gradient ascent algorithm with batch size = 100 and stochastic gradient ascent algorithm to train your logistic regression model. Initialize all weights to random numbers drawn from a Gaussian distribution N(0, 0.01). Use the learning rate you have chosen in Question 3.1 and perform 1000 iterations to train your model. Report the accuracies and the confusion matrices using your models on the given test set. Calculate and report averages of precision, recall, negative predictive value (NPV), false positive rate(FPR), false discovery rate (FDR), F1 and F2 scores.

Question 3.3 **[5 pts]** In what cases, NPV, FPR, FDR, F1 and F2 would be more informative metrics compared to accuracy, precision and recall alone? Explain.

References

[1] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[2] "House prices - advanced regression techniques."

[3] M. L. G. ULB, "Credit card fraud detection," Mar 2018.