CS464 Introduction to Machine Learning
Spring 2019
Homework 2

Due: December 11, 2019 11:59 PM

Instructions

- For this homework, you may code in 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 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_sheldon_cooper`). 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.

1 PCA & Cats [25 pts]

In this question, you are expected to analyze cat images using PCA. You will use the dataset provided within the homework zip file as `cats`. The dataset is composed of 5000 images of cats. For this question, use of any library for PCA calculations is not allowed.

Flatten all images of size $64 \times 64 \times 3$ to get 4096×3 matrix for each image. Note that all images are 3-channel RGB. Create a 3-D array, X, of size $5000 \times 4096 \times 3$ by stacking flattened matrices of the images provided in the dataset. Slice X as $X_i = X[:, :, i]$ where i corresponds to the first three index. thus obtaining each color channel matrix independently for all images. Reshape all X_i to obtain matrices, instead of 3D arrays.

Question 1.1 [10 pts] Apply PCA on X_i's to obtain first 10 principal components for each X_i. Report proportion of variance explained (PVE) for each of the principal components. Discuss your results.

Question 1.2 [5 pts] Using the first 10 principal components found for each color channel, reshape each principal component to a 64×64 matrix. Stack corresponding principal components of each color channel to obtain 10 RGB images of size $64 \times 64 \times 3$ and display all. Discuss your results.

Question 1.3 [10 pts] Describe how you can reconstruct an original cat 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 $k \in \{1, 50, 250, 500\}$. Discuss your results.

2 Linear Regression [25 pts]

Dataset

The dataset required for this question is in `q2-train-features.csv` file. The dataset for this question consists of monthly average of USD/TRY exchange rates from November 2014 to August 2019. In addition, you can also find CPI (Consumer Price Index, or TÜFE in Turkish) and unemployment rates of each month from November 2014 to August 2019. Please read question instructions carefully. Provide proper title, axis labels and legend for each plot requested. You will lose points for unformatted plots. You are not allowed to use any machine learning libraries for any question in this section. The dataset is constructed using the information provided by Central Bank of the Republic of Turkey and Turkish Statistical Institute.

Question 2.1 [3 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 [5 pts] Using the formula you have derived for Question 2.1, train a linear regression model by using only "Months Past Since November 2014" feature which is given to you in the dataset. Consider all of the dataset as training set. Report your coefficients and interpret these coefficients. In addition, plot the graph of USD/TRY exchange rate vs. "Months Past Since November 2014" along with your model's predictions on the same plot. Finally, also provide the training MSE for your model. This model will be referred as model A in the following questions.

Question 2.3 [5 pts] This time, you will be adding CPI value as a new feature to model A. Modify your feature matrix and train your model again using "Months Past Since November 2014" and "CPI" features. Consider all of the dataset as training set. Report your model coefficients and interpret them. Plot the graph of USD/TRY exchange rate vs. "Months Past Since November 2014" along with your model's predictions on the same plot. Please also provide your new model's training MSE. This model will be referred as model B in the following questions.

Question 2.4 [5 pts] Similar to the previous two questions, you will be constructing another model by adding "Unemployment Rate" in addition to "CPI" and "Months Past Since November 2014" features. Modify your feature matrix accordingly and train your model with this new feature matrix. Consider all of the dataset as your training set. Report your coefficients and comment on them. Plot the graph of USD/TRY exchange rate vs. "Months Past Since November 2014" along with your model's predictions on the same plot. Please also provide your new model's training MSE. This model will be referred as model C in the following question.

Question 2.5 [5 pts] For this question, split your dataset into training and test sets as follows: consider the last 3 rows as the test set (August 2019, July 2019 and June 2019 data) and consider the remaining rows as the training set. Train model A, model B and model C on this new training set and then test all these models on the test set. Report test set MSE values for each model and explain which model is better at predicting USD/TRY exchange rate using MSE values you have found. Would you agree that these models are promising models for predicting exchange rates? Explain your stance clearly.

Question 2.6 [2 pts] Discuss how would you perform exchange rate prediction task in a more successful way. You may suggest a modification to the models you have constructed before or you may propose a completely new model.

3 Support Vector Machines (SVMs) [25 pts]

Dataset

You are a teacher in a secondary school and you are tasked with choosing eligible students for a special scholarship. Your school has hundreds of students and you notice that you do not have enough time to go over every student's grades and documents to see whether they are eligible for the scholarship. Hence, you and one of your colleagues from another secondary school use your love of computer science to find a solution and decide to train a machine learning model to decide whether you should give a student the scholarship or not.

Your dataset contains answers of the survey given to students in the math course in two secondary schools from previous years [1, 2]. The dataset has the following attributes. These attributes are given in the same order in the feature files as follows.

- school - student's school (binary: '1' - Gabriel Pereira or '0' - Mousinho da Silveira)
- gender - student's gender (binary: '1' - female or '0' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: '1' - urban or '0' - rural)
- famsize - family size (binary: '0' - less or equal to 3 or '1' - greater than 3)
- Pstatus - parent's cohabitation status (binary: '1' - living together or '0' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (one-hot encoded: '4' - teacher, '3' - health care related, '2' - civil services (e.g. administrative or police), '1' - at home or '0' - other)
- Fjob - father's job (one-hot encoded: '4' - teacher, '3' - health care related, '2' - civil services (e.g. administrative or police), '1' - at home' or '0' - other)
- reason - reason to choose this school (one-hot encoded: '1' - close to home, '2' - school reputation, '3' - course preference or '0' - other)
- guardian - student's guardian (nominal: '1' - mother, '2' - father or '0' - other)
- travel time - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- study time - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - ¿10 hours)
- failures - number of past class failures (numeric: n if $1 <= n < 3$, else 4)
- school support - extra educational support (binary: '1' - yes or '0' - no)
- family support - family educational support (binary: '1' - yes or '0' - no)
- paid - extra paid classes within the course subject (binary: '1' - yes or '0' - no)
- activities - extra-curricular activities (binary: '1' - yes or '0' - no)
- nursery - attended nursery school (binary: '1' - yes or '0' - no)
- higher - wants to take higher education (binary: '1' - yes or '0' - no)
- internet - Internet access at home (binary: '1' - yes or '0' - no)
- romantic - with a romantic relationship (binary: '1' - yes or '0' - no)

- family relations - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- free time - free time after school (numeric: from 1 - very low to 5 - very high)
- go out - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)

The data have been already split into two subsets: a 280-sample subset for training and a 115-sample subset for testing (consider this as your validation set and imagine there is another test set which is not given to you). You will use the following files:

- `q3-train-features.csv`
- `q3-train-binary-labels.csv`
- `q3-train-multiclass-labels.csv`
- `q3-test-features.csv`
- `q3-test-binary-labels.csv`
- `q3-test-multiclass-labels.csv`

The files that end with features.csv contain the features and the files ending with labels.csv contain the ground truth labels. In the feature files, each row contains the feature vector for a student. The j-th term in the row i is the information related to j-th attribute of the i-th student. The binary-label files include the ground truth label, i.e. whether the corresponding student is eligible for the scholarship. The order of the student rows is the same in the features and the labels files. That is, the i-th row in both files corresponds to the same student's information. You will use these files, train-binary-labels.csv and test-binary-labels.csv, for Question 3.1 and Question 3.2. The multiclass-label files include the ground truth label, i.e. whether the corresponding student is accepted, on-hold or not-eligible for the scholarship. Again, the order of the student rows is the same in the features and the labels files. You will use these files, train-multi-labels.csv and test-multiclass-labels.csv, for Question 3.3.

In this question, you will train one soft margin and one hard margin SVM classifier on the dataset explained above. You must perform 5-fold cross validation WITHOUT using any libraries but you CAN use libraries or software packages to train your SVM.

Question 3.1 [8 pts] Use files that end with binary-label.csv for this question. In this part, you will train a linear SVM model with soft margin (no kernel). Your model's hyper-parameter is C. Using 5-fold cross validation on your *training set*, find the optimum C value of your model. Look for the best C value within the interval from $[10^{-3}, 10^{-2}, 10^{-1}, 10, 10^{1}, 10^{2}]$ and calculate accuracy on the left-out fold. For each value of C, calculate mean cross validation accuracy by changing the left-out fold each time and plot it in a nice form. Report your optimum C value. Then, run your model on the *test set* with this C value and report test set accuracy along with the confusion matrix. Calculate and report micro and macro averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores.

Question 3.2 [8 pts] Use files that end with multiclass-label.csv for this question. This time, use radial basis function (RBF) kernel to train your hard margin SVM model on the processed data set. RBF kernel is defined as in Eqn. 3.1

$$K(x,x') = exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right) \tag{3.1}$$

In RBF kernel formula, $\gamma = -\frac{1}{2\sigma^2}$ is a free parameter that can be fine-tuned. This parameter is the inverse of the radius of the influence of samples selected by the model as support vectors. Similar to linear SVM part, train a SVM classifier with RBF kernel using same training and test sets you have used in linear SVM model above. γ is your new hyper-parameter that needs to be optimized. Using 5-fold cross validation and calculating mean cross validation accuracy as described in Question 3.1, find and report the best γ within the interval from the logarithmic scale $[2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}]$. After tuning γ on your *training set*, run your model on the *test set* and report your accuracy along with the confusion matrix. Calculate and report micro and macro averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores.

Question 3.3 [9 pts] As an additional task, you are later asked to separate students into three classes: accepted, on-hold or not-eligible. Use files that end with multiclass-label.csv for this question. Train a hard margin SVM with RBF kernel for this three-class classification problem using *one vs. all* approach. Set your hyper-parameter γ to its optimal value and report class based accuracy, micro and macro averages of precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1 and F2 scores for all classes. Discuss your results. What would happen if you used *all vs. all* approach?

4 Logistic Regression [25 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:

- `q4-train-dataset.csv`
- `q4-test-dataset.csv`

Question 4.1 [10 points] Implement full batch gradient ascent algorithm to train your logistic regression model. Initialize all weights to random numbers drawn from a Gaussian distribution $N(0, 0.01)$. Try different learning rates and choose the one which works best for you. Use 1000 iterations to train your model. Print model weights in each iteration $i \in \{100, 200, ..., 1000\}$. Report the 10 most important features and discuss the relation between weights and individual importance of features. Report the accuracy, precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1, F2 scores and the confusion matrix using your model on the given test set. Discuss your results. (Hint: Analyze the dataset and determine if normalization is required.)

Question 4.2 [12 points] Implement mini-batch gradient ascent algorithm with *batch size = 32* and stochastic gradient ascent algorithm to train two logistic regression models. Initialize all weights to random numbers drawn from a Gaussian distribution $N(0, 0.01)$. Use the learning rate you have chosen in Question 4.1 and perform 1000 iterations to train your models. Report the class based accuracies and the confusion matrices using your models on the given test set. Calculate and report precision, recall, negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), F1, F2 scores for each model. Discuss your results. (Hint: Analyze the dataset and determine if normalization is required.)

Question 4.3 [3 points] In what cases, NPV, FPR, FDR, F1 and F2 would be more informative metrics compared to accuracy, precision and recall alone? Explain.

References

[1] P. Cortez and A. M. G. Silva, “Using data mining to predict secondary school student performance,” 2008.

[2] D. Dua and C. Graff, “UCI machine learning repository,” 2017.

[3] J. B. Pozzolo, Caelen, “Credit card fraud detection: a realistic modeling and a novel learning strategy,” 2015.