CS464 Machine Learning
Fall 2018
Homework 2

Due: December 10 17:00 pm

Instructions

- Submit a soft copy of your homework of all questions to Moodle. Add your code at the end of your report while submitting it. Your report should be a .pdf file. Submitting a hard copy or scanned files is NOT allowed. You have to prepare your homework digitally (using Word, Excel, Latex etc.).
- You may code in any programming language you would prefer. In submitting the homework file, please package your report and code files as a gzipped TAR file or a ZIP file with the name `CS464_HW2_Firstname_Lastname`. Please do not use any Turkish letters for any of your files including code files and report file. The code you submit should be in a format easy to run, main script to call other functions. You must also provide us with a README file that tells us how we can execute/call your program.
- Please read the `Homework Guidelines.pdf` file that is provided inside the homework zip file which contains guidelines for homework preparation.
- Any violation of these rules may lead to significant grade deduction.

1 Linear and Polynomial Regression [30 pts]

I am Looking for an Efficient Car [15 pts]

In this question, you will perform multivariate linear regression. Multivariate linear regression is the generalized version of linear regression where your task is to predict $\hat{y}$ using multiple features. In this case, your regression line becomes a regression hyperplane in a multidimensional feature space. For this part, you are asked to predict MPG(miles per gallon) value of a car given its cylinder count, displacement, horsepower, weight, acceleration and model year. You will use `carbig.csv`[1] for this question. Columns are arranged as follows in the file: Cylinders, Displacement, Horsepower, Weight, Acceleration, Model Year and MPG.

Question 1.1 [3 pts] Derive the general closed form solution for multivariate regression model using ordinary least squares loss function given in equation 1.1. Briefly explain each matrix involved in calculation and how they are constructed.

$$J_n = \frac{1}{2}||y - X\beta||^2 = \frac{1}{2}(y - X\beta)^T(y - X\beta) \tag{1.1}$$

Question 1.2 [4 pts] Find the rank of X^TX for the given dataset in `carbig.txt` 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 1.1.

Question 1.3 [6 pts] You cannot use any machine learning libraries to train and test your model for this question. Train your model using the first 300 rows in `carbig.csv`. Reserve the remaining rows for your test set. Report your trained model's coefficients (β values). Evaluate mean squared error on training and test set separately and report them.

Question 1.4 [2 pts] After training the model for first 300 rows, comment on the coefficients you have found. What does it mean if a coefficient has negative sign? What is the relation between the magnitude of a coefficient and the predicted value?

MPG Prediction from Different Perspective [15 pts]

This time, assume that you are only provided horsepower values of cars. In this question, you will use polynomial regression to train a model. In polynomial regression, you will be using a feature x_i and its powers x_i^2, x_i^3 etc. to construct your design matrix so that the non-linear relationship between features and the ground truth values can be captured.

Question 1.5 [2 pts] Plot MPG vs. horsepower for the whole dataset (training set + test set) with horsepower on x axis and MPG on y axis. Comment on the relation between MPG and horsepower.

Question 1.6 [3 pts] What happens to the rank of X^TX as you increase p ? Find the rank of X^TX for the `carbig.csv` dataset as a function of p where $p \in \{0, 1, 2, 3, 4, 5\}$. Based on the graph you have plotted, discuss the possible problems you may encounter when you use closed form solution for β values. Now, centralize your dataset using the equation 1.2 where x_{ij} is i^{th} data instance's j^{th} feature value, μ_j is the mean value of feature j and σ_j is the standard deviation of feature j. Centralizing the dataset normalizes the data such that each feature will have a distribution with 0 mean and unit variance. After centralizing the dataset, check the rank values again for each p value. Did centralizing the data helped to increase the rank? Use the centralized dataset for the rest of this section.

$$x_{ij} = \frac{(x_i - \mu_j)}{\sigma_j} \tag{1.2}$$

Question 1.7 [5 pts] You are not allowed to use any machine learning libraries to train and test your model for this question. Construct your design matrix X such that a single row should look like the following:

$$[1\ x_1\ x_1^2\ x_1^3\ ...\ x_1^p]$$

where $p \in \{0, 1, 2, 3, 4, 5\}$ and x_1 is the centralized horsepower value of a particular car. Report mean squared error on both training and test sets for each $p \in \{0, 1, 2, 3, 4, 5\}$. On a single graph, draw all regression lines for different p values along with original training data.

Question 1.8 [5 pts] You are not allowed to use any machine learning libraries to train and test your model for this question. Assume now that besides horsepower values, you are also provided model year of each car. Using a similar approach as in Question 1.7, construct your design matrix such that a single row looks like the following:

$$[1\ x_1\ x_1^2\ x_1^3\ ...\ x_1^p\ x_2\ x_2^2\ ...\ x_2^p]$$

where x_1 is the centralized horsepower value of a car and x_2 is the centralized model year of a car. For each $p \in \{1, 2, 3\}$, calculate and report mean squared error for training and test sets. Compare your results against the results you have obtained for Question 1.7.

2 Logistic Regression and Model Evaluation [40 pts]

Question 2.1 [**15 pts**] You are not allowed to use any machine learning libraries to train and test your model for this question. For this part of the question you will use the file `ovariancancer.csv`[2] and `ovariancancer_labels.csv`. In `ovariancancer.csv`, there are 4000 predictors (features) to determine if a patient has ovarian cancer and each row contains information about a single patient. In `ovariancancer_labels.csv`, 1 denotes a person with ovarian cancer and 0 denotes a healthy person in terms of ovarian cancer.

You will be implementing gradient ascent algorithm to train your logistic regression model. Your hyper-parameters are iteration count and learning rate. Divide your dataset such that the test set contains the first 20 people that have ovarian cancer and the first 20 people that do not have ovarian cancer in `ovariancancer.csv`. Then, your training set consists of the remaining people. You will perform 5-fold cross validation on training set to select optimum values for your hyper-parameters. Use iteration count and learning rate values as follows: Iteration Count $\in \{500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$ and Learning Rate $\in$ { 0.001, 0.002, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03 }

Report the hyper-parameters you found and the confusion matrix using your model on the 40 people test set you have separated.

Question 2.2 [**15 pts**] You are not allowed to use any machine learning libraries for this question. Implement forward selection and backward elimination algorithms separately to find optimal subset of features by maximizing accuracy on the training dataset. To find the accuracy of a particular feature set, use 5-fold cross validation mean accuracy. You now have two models based on the subset of features you have obtained using forward selection and backward elimination. Evaluate both of your models on the test set (with 40 people) and report corresponding confusion matrices. Are the two subsets you have obtained using forward selection and backward elimination the same? If not, explain the reason.

Question 2.3 [**10 pts**] You will be using the two feature sets you have obtained for Question 2.2 as two separate models. In this question, you will be building ROC and Precision-Recall curves to compare the performances of these two models. Use the procedure described below to obtain ROC and Precision-Recall curves on 40 people test set. You are not allowed to use any library or function that automatically draws ROC and Precision-Recall curves.

Create two tables (could be matrices, you do not have to report it) for each classifier such that for each test instance you are going to provide $P(Y = 1|X = x)$ and that instance's ground truth label. Sort values in descending order by $P(Y = 1|X = x)$ values. Now you have a knob that determines $k \in [0, 1]$ where an instance with $P(Y = 1|X = x) \geq k$ will be classified as $Y = 1$. Start with $k = 0$ so that every test instance will be classified as $Y = 1$. At each step, you tune k such that next instance above and all instances below it will be classified as $Y = 0$. You tune k until every instance is classified as $Y = 0$, i.e. $k = 1$ (Remember that you started where every instance is classified as $Y = 1$). Now for each value of k (there should be n such k values where n is number of test instances), calculate true positives, false positives, false negatives and true negatives. Using these values, plot ROC and Precision-Recall curves for each classifier and compare their performance using the curves you have plotted. You may use libraries or functions to find area under a curve.

3 Support Vector Machines [30 pts]

Question 3.1 [**8 pts**] Consider the simple dataset given at Fig. 1. For this dataset, consider a soft margin SVM model. Find and report an interval for C such that the hard margin on this dataset is enforced. Additionally find and report α_i values of each data point for all possible margins. Show all your work on your report to receive full credit.

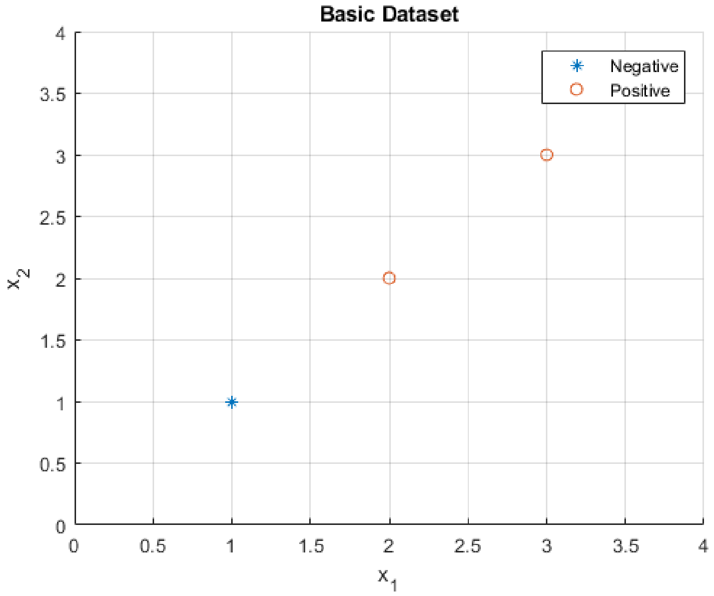

Figure 1: Basic Dataset with 3 Data Points

Question 3.2 [**2 pts**]
Considering a soft margin SVM model, what happens when C value (margin violation penalty) is set to ∞ ?

Training SVM on Real Data [20 pts]

In this question, you will train one soft margin and one hard margin SVM classifiers on the `UCI_Breast_Cancer.csv` dataset[3]. This dataset has 9 predictors for breast cancer. The first column is the ID number of a patient which can be ignored during training and test. The next 9 columns are the predictors and the last column is the class label. In terms of class labels, 2 denotes benign tumor and 4 denotes malignant tumor.

You can use libraries to train SVM. However, you must perform k-fold cross validation without using any libraries. Take first 500 rows as your training set and the rest as your test set.

Question 3.3 [**10 pts**] Without using any kernels, you will train a linear SVM model with soft margin. Your model's hyper-parameter is C. Using k-fold cross validation, find the optimum C value of your model (You can choose C values from a logarithmic scale, e.g. 10^{-3}, 10^{-2}, ... ,10^{1},10^{2} etc.). Plot your train set accuracies while tuning for the optimum C value (i.e. for each value of C, calculate mean cross validation accuracy among all folds and plot it in a nice format). Report your optimum C value. Then, run your model on the test set and report test set accuracy with the confusion matrix.

You are free to choose k as you like. In your report, please state your k value and explain why you chose that value.

Question 3.4 [**10 pts**] This time, use RBF kernel to train your hard margin SVM model. RBF kernel is defined as

$$K(x,x') = exp\Bigg(- \frac{||x - x'||^2}{2\sigma^2}\Bigg) \qquad (3.1)$$

In RBF kernel formula, $\gamma = -\frac{1}{2\sigma}$ is a free parameter that can be fine-tuned. 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 you new hyper-parameter that needs be optimized. Using k-fold cross validation, find the best γ within a specified interval. (again, γ values can be chosen from logarithmic scale such as 2^{-4}, 2^{-3}, ... , 2^{0},

2^1 etc.). After tuning γ, run your model on the test set and report your accuracy along with the confusion matrix. Additionally, report your optimum γ.

References

1. Auto MPG Dataset https://archive.ics.uci.edu/ml/datasets/auto+mpg
2. MATLAB Ovarian Cancer Dataset
3. UCI Breast Cancer Dataset
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29