Machine Learning - Assignment 1

You must submit to GitHub and give the repository link. Make sure to include clear explanations for each step of your code. Copied code simply rejected, and CIE, assignment 2 will not be allowed.

1. Regression

CO1,2 B4

For this exercise, you will experiment with regression, regularization, and cross-validation. Choose appropriate Dataset.

(a) Load the data into memory. Make an appropriate X matrix and y vector.

(b) Split the data at random into one set (X_{train}, y_{train}) containing 80% of the instances, which will be used for training + validation, and a testing set X_{test}, y_{test}) (containing remaining instances).

(c) Give the objective of logistic regression with L_2 regularization.

(d) Run logistic regression on the data using L_2 regularization, varying the regularization parameter $\lambda \in \{0, 0.1, 1, 10, 100, 1000\}$. Plot on one graph the average cross-entropy for the training data and the testing data (averaged over all instances), as a function of λ (you should use a log scale for λ). Plot on another graph the L_2 norm of the weight vector you obtain. Plot on the third graph the actual values of the weights obtained (one curve per weight). Finally, plot on a graph the accuracy on the training and test set. Explain briefly what you see.

(e) Re-format the data in the following way: take each of the input variables, and feed it through a set of Gaussian basis functions, defined as follows. For each variable (except the bias term), use 5 univariate basis functions with means evenly spaced between -10 and 10 and variance σ. You will experiment with σ values of 0.1, 0.5, 1, 5 and 10.

(f) Using no regularization and doing regression with this new set of basis functions, plot the training and testing error as a function of σ (when using only basis functions of a given σ). Add constant lines showing the training and testing error you had obtained in part c. Explain how σ influences overfitting and the bias-variance trade-off.

(g) Add in *all* the basis function and perform regularized regression with the regularization parameter $\lambda \in \{0, 0.1, 1, 10, 100, 1000, 10000\}$. Plot on one graph the average cross- entropy error for the training data and the testing data, as a function of λ (you should use a log scale for λ). Plot on another graph the L_2 norm of the weight vector you obtain. Plot on a different graph the L_2 norm of the weights for the set of basis functions corresponding to each value of σ, as a function of λ (this will be a graph with 5 lines on it). Explain briefly the results.

(h) Explain what you would need to do if you wanted to design a set of Gaussian basis functions that capture relationships between the inputs. Explain the impact of this choice on the bias-variance trade-off. No experiments are needed (although you are welcome to explore this on your own).

(i) Suppose that instead of wanting to use a fixed set of evenly-spaced basis functions, you would like to adapt the placement of these functions. Derive a learning algorithm that computes both the placement of the basis function, μ_i and the weight vector w from data (assuming that the width σ *isfixed*. You should still allow for L_2 regularization of the weight vector. Note that your algorithm will need to be iterative.

(j) Does your algorithm converge? If so, does it obtain a locally or globally optimal solution? Explain your answer.

2. Experiment on any complex datasets to demonstrate the Linear REGRESSION and its versions, and logistic regression (CLASSIFICATION) along with complete data preprocessing steps.

   Note: You should not show your lab practices for this question. CO1,2 B5

---PRACTICE and SUBMITE--