Homework 1

MTH 994 – Machine Learning
Due date: Friday, Oct 12, 2018

(6 problems/2 pages)

1 Handwritten Homework

Note All problems in this section requires the handwritten answers.

Problem 1 (10pts). Assume the training data is given as follows: (x_1, y_1), (x_2, y_2), $\ldots$, (x_M, y_M). The predictor of the linear regression is defined as

$$p_{\mathbf{c}}(x) = c_0 + c_1 x$$

a) Find the loss function associated with the predictor $p_{\mathbf{c}}(x)$.

b) Find the optimal values c_0 and c_1. **(Note: show all of your steps to receive a full credit)**

Problem 2 (15pts). Assume the training data for the classification task is given as follows: (x_1, y_1), (x_2, y_2), $\ldots$, (x_M, y_M), with $y_i \in \{0, 1\}$, $i = 1, 2, \ldots, M$. The logistic regression is employed to learn this dataset.

a) What is the predictor for a given input x?

b) Show all of steps for constructing the loss function of the logistic regression method?

c) What is the parameter vector $\mathbf{c}$ in the predictor after the first iteration in the gradient descent. (choose your own initial values)

Problem 3 (10pts). a) What is the purpose of the regularization?

b) State the loss functions of linear regression and logistic regression under the regularization.

Problem 4 (10pts). Assume the features of our training data is given as

$$(1, 20), (-3, 40), (-2, 10), (0, 30)$$

a) Use two different ways of normalizing features to scale all the feature values in the training data.

b) If the test data is given by $(4, 25), (2, 15)$. Find the normalized features of the test set corresponding to each normalization approach.

2 Programming Homework

Note Write your codes in Azure notebook or similar kinds. Each question is in a separate notebook and submit all of them via a dropbox in D2L. (**You are allowed to use the available machine learning libraries in Python**)

Problem 2.1 (30pts). Given training data: `X_house_train.csv` (feature values), `y_house_train.csv` (labels) and test data: `X_house_test.csv` (feature values), `y_house_test.csv` (labels) . File `House_feature_description.csv` describes the meaning of each column in the data set.

a) Program a linear regression (LR) model to predict the labels in the test data. And explicitly write down the representation of model's predictor (**note**: type down your formulation in the notebook)

b) Program a Tikhonov-regularized linear regression model to predict the labels in the test data. Compare this model to the previous one, and comment on the choice of regularized parameters.

Problem 2.2 (25pts). Given the Iris dataset. It has been split into training data: `X_iris_train.csv` (features), `y_iris_train.csv` and test data `X_iris_test.csv` (features), `y_iris_test.csv`. File `Iris_description.csv` describes the meaning of each column in the data set.

a) Program a LASSO-regularized logistic regression model to predict the test data. Comment on the choice of regularized parameters.

b) Plot the decision boundary of your best model along with test set.