EE 526X Deep Machine Learning: Theory and Practice — Homework 1
Assigned: 09/14. Due: 09/23

Problem 1.

There is a ground-truth model

$$y = f(x) + \epsilon, \quad 0 \leq x \leq 1, \tag{1}$$

where

$$f(x) = 1 + 2\sin(5x) - \sin(15x) \tag{2}$$

and ϵ is additive noise that is Gaussian distributed with zero mean and unit variance. The added noises for different x values are independent.

Using this model, we would like to test polynomial fitting. Specifically,

(a) Generate 51 equally spaced x values between 0 and 1:

$$x_i = \frac{i}{50}, \quad i = 0, 1, \ldots, 50. \tag{3}$$

(b) Generate y_i values for these x_i values:

$$y_i = f(x_i) + \epsilon_i, \quad i = 0, 1, \ldots, 50. \tag{4}$$

where ϵ_i are independently and identically distributed standard Gaussian random variables. The set of generated values $(x_i, y_i), i = 0, \ldots, 50$, will serve as the training data.

(c) Fit a polynomial of order $k = 1$ to the dataset, minimizing the residual sum of squares. Plot the fitted polynomial using black color.

(d) Repeat the steps (b) to (c) 30 times. Keep all the plotted polynomials.

(e) Repeat the experiment for $k = 3, 5, 7, 9, 11$. For each k, use a new figure. Also, on each figure, show the function $f(x)$ using red color.

You need to write the code using Python. You should implement the polynomial fitting function (see the "Linear Regression" lecture notes). You will be graded based on two artifacts: 1) the figures generated (need to be included in the submitted homework report), and 2) the source code submitted.

Problem 2. Use `Python` to implement the perceptron algorithm and test it on the following data:

	x_1	x_2	y
1	1	2	1
2	1	4	1
3	2	2	1
4	4	2	-1
5	3	4	-1
6	2	3	-1

where there are 6 points, (x_1, x_2) is the input, and y is the binary output label. Initialize your algorithm with the vector

$$\theta = [b, w_1, w_2]^T = [0, 0, 0]^T. \tag{5}$$

Plot the points and the final hyperplane (a line) on the same graph.

Problem 3. The Spambase Data Set contains email spam data for 4601 email messages. Download the data from `https://archive.ics.uci.edu/ml/datasets/spambase` and divide the data into training set and test set. The training set should contain the first 2/3 of spam messages and first 2/3 of ham (i.e., non-spam) messages. The test set should contain the last 1/3 spam messages and last 1/3 ham messages.

(a) Write a logistic regression program (function) using gradient descent algorithm. And train the weights using training set and then test the result on the test set. Experiment with the step size (learning rate).

(b) Next, normalize the features, so that each feature in the training data has mean 0 and variance 1. Then run logistic regression on the normalized data.

You should perform the simulation using `Python`.

You need to submit in the homework report a summary of the results you obtained, and the results on the learning rate used, and training and test errors. Source code in `Python` should also be submitted.

END OF ASSIGNMENT