Assignment 4

Foundations of Machine Learning
IIT-Hyderabad
Aug-Dec 2021

Max Marks:
Due: 11th Nov 2021 11:59 pm

This homework is intended to cover theory and programming exercises in the following topics:

- Linear regression, Optimal bayes classifier, VC dimension, Regularizers

Instructions

- Please upload your submission on Google Classroom by the deadline mentioned above. Your submission should comprise of a single file (PDF/ZIP), named `<Your_Roll_No>_Assign4`, with all your solutions.
- For late submissions, 10% is deducted for each day (including weekend) late after an assignment is due. Note that each student begins the course with 7 grace days for late submission of assignments (of which atmost 4 can be used for a given submission). Late submissions will automatically use your grace days balance, if you have any left. You can see your balance on the FoML Marks and Grace Days document (will be shared on Piazza).
- Please use PYTHON for the programming questions.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. Please talk to instructor or TA if you have concerns.

Questions: Theory

1. **Non-Uniform Weights in Linear Regression:** ***(6 marks)*** You are given a dataset in which the data points are denoted by $(\mathbf{x}_n, t_n), n = 1, \cdots, N$. Each data point is associated with a non-negative weighting factor $g_n > 0$. The error function is thus modified to:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} g_n\Big(t_n - \mathbf{w}^T\Phi(\mathbf{x}_n)\Big)^2$$

where $\Phi(\cdot)$ is any representation of the data.

(a) ***(3 marks)*** Find an expression for the solution $\mathbf{w}^*$ that minimizes the above error function.

(b) ***(3 marks)*** Give two alternative interpretations of the above weighted sum-of-squares error function in terms of: (i) data-dependent noise variance and (ii) replicated data points.

2. **Bayes Optimal Classifier:** ***(2 marks)*** Let there be 5 hypotheses h_1 through h_5 that could guide a robot to move either Forward(F) or Left(L) or Right(R):

$P(h_i\|D)$	$P(F\|h_i)$	$P(L\|h_i)$	$P(R\|h_i)$
0.4	1	0	0
0.2	0	1	0
0.1	0	0	1
0.1	0	1	0
0.2	0	1	0

Compute the MAP estimate and Bayes optimal estimate using the data provided in the table. Are they the same? Justify your answer.

3. **VC-Dimension:** ***(2 marks)*** Consider a data setup of one-dimensional data $\in \mathbb{R}^1$, where the hypothesis space $\mathcal{H}$ is parametrized by $\{p, q\}$ where x is classified as 1 iff $p < x < q$. Find the VC-dimension of $\mathcal{H}$.

4. **Regularizer:** ***(4 marks)*** Given D-dimensional data $\mathbf{x} = [x_1, x_2, \cdots, x_D]$, consider a linear model of the form:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{k=1}^{D} w_k x_k$$

Now, for N such data samples with their corresponding labels $(\mathbf{x}_i, t_i), i = 1, 2, \cdots, N$, the sum-of-squares error (or mean-squared-error) function is given by:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \Big(y(\mathbf{x}_i, \mathbf{w}) - t_i \Big)^2$$

Now, suppose that Gaussian noise $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ (i.e. zero mean and variance σ^2) is added independently to each of the input variables x_k. Find a relation between: minimizing the above sum-of-squares error averaged over the noisy data, and minimizing the standard sum-of-squares error (averaged over noise-free input data) with a $\mathcal{L}_2$ weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Questions: Programming

5. **Logistic Regression:** ***(7 marks)***

(a) ***(3 marks)*** Implement your own code for a logistic regression classifier, which is trained using gradient descent and cross-entropy error as the error function.

Index	x_1	x_2	y
1	0.346	0.780	0
2	0.303	0.439	0
3	0.358	0.729	0
4	0.602	0.863	1
5	0.790	0.753	1
6	0.611	0.965	1

Table 1: Train Set

Index	x_1	x_2	y
1	0.959	0.382	0
2	0.750	0.306	0
3	0.395	0.760	0
4	0.823	0.764	1
5	0.761	0.874	1
6	0.844	0.435	1

Table 2: Test Set

(b) Consider the training set and test set given in Tables 1 and 2. We use the linear model $f_\theta(x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ and the logistic regression function $\sigma(f_\theta(x_1, x_2)) = \frac{1}{1+\exp^{-f_\theta(x_1,x_2)}}$. Consider the initial weights as $\theta_0 = -1$, $\theta_1 = 1.5$, $\theta_2 = 0.5$, and learning rate as 0.1 (for gradient descent).

i. ***(1 mark)*** What is the logistic model $P(\hat{y} = 1|x_1, x_2)$ and its cross-entropy error function?

ii. ***(1 mark)*** Use gradient descent to update θ_0, θ_1, θ_2 for one iteration. Write down the updated logistic regression model.

iii. ***(2 mark)*** At convergence of gradient descent, use the model to make predictions for all the samples in the test dataset. Calculate and report the accuracy, precision and recall to evaluate this model.

Deliverables:

- Code
- Brief report with answers to above questions.

6. **Kaggle - Taxi Fare Prediction:** ***(9 marks)*** The next task of this assignment is to work on a (completed) Kaggle challenge on taxi fare prediction. As part of this task, please visit https://www.kaggle.com/c/new-york-city-taxi-fare-prediction to know more about this problem, and download the data. (You now know how to download data from Kaggle.)

   You are allowed to use any machine learning library of your choice: scikitlearn, pandas, Weka (we recommend `scikitlearn`), and any regression method too. Use `train.csv` to train your classifier. Predict the fares on the data in `test.csv`, and report your best 2 scores in your report. (We will also upload your codes randomly to confirm the scores.). Your model should achieve at least RMSE < 4.
   Deliverables:

   - Code
   - Brief report with top-2 scores of your methods, and a brief description of the methods that resulted in the top 2 scores.
   - Your report should also include your analysis of why your best 2 methods performed better than others you tried.