Assignment 4

Foundations of Machine Learning
IIT-Hyderabad
Aug-Nov 2024

Max Marks: 30
Due: 24th Nov 2024 11:59 pm

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

- Learning Theory, Logistic Regression, Clustering, Dimensionality Reduction

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.
- Please use PYTHON for the programming questions.
- Answers for theoretical questions must be typed out in LaTeX.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. **We will use tools to check for plagiarism between yourselves or LLM sources, as well as conduct vivas for veracity of your submissions.** Please talk to instructor or TA if you have concerns.

Questions: Theory

1. **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 $\mathcal{H}$.

2. **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.

3. **Hierarchical Clustering** ***(4 marks)*****:** Given below is the distance matrix for 6 data points

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	0.12	0				
x_3	0.51	0.25	0			
x_4	0.84	0.16	0.14	0		
x_5	0.28	0.77	0.70	0.45	0	
x_6	0.34	0.61	0.93	0.20	0.67	0

(a) Draw a dendrogram for the final result of hierarchical clustering with single link. [*1 mark*]

(b) Draw a dendrogram for the final result of hierarchical clustering with complete link. [*1 mark*]

(c) Change two values from the matrix so that the answer to the last two questions is same. [*2 marks*]

4. **Principal Component Analysis** ***(4 marks)*****:** Suppose each data point $\mathbf{x}$ is an M-dimensional vector of the form $\mathbf{x} = a\delta_k = (0, ..., 0, a, 0, ...)^T$, where a is in the k^{th} slot, and k,a are random variables. k is uniformly distributed over $1, \cdots, M$ and $P(a)$ is arbitrary.

(a) Calculate the covariance matrix. [*1 mark*]

(b) Show that it has one eigenvector of form $(1, ..., 1)$ and that the other eigenvectors all have the same eigenvalue. [*2 marks*]

(c) Discuss whether PCA is a good way to select features for this problem. [*1 mark*]

(**Hint:** Use expectation to compute the covariance matrix: $C = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T]$. You should get C of the form $C_{ij} = \lambda + \mu\delta_{i,j}$ for some λ, μ.)

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.

   (b) Consider the training set and test set given in Tables 1 and 2. We use the linear

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

   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/House Price Prediction:** ***(9 marks)*** The next task of this assignment is to work on a (completed/ongoing) Kaggle challenge. You can do ANY ONE of the following: (i) Taxi fare prediction; OR (ii) House price prediction (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.)

   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.