CS 5824/ECE 5424: Advanced Machine Learning Assignment 1

Fall 2022

Due Date: Sep 29, 2022, 11:59pm EST

Introduction

This homework will cover Maximum Likelihood Estimators (MLE), Linear Regression, Logistic Regression, and Evaluation (cross-validation, confusion matrices, bias-variance trade-off, etc.). The objective of this homework is to enable you to understand these concepts in more detail and help you draw connections between many important mathematical concepts and their utility in ML.

The homework consists of two sections, Section A and Section B. Only Section A questions are given in this document. Section B is a Python Notebook that you will have received along with this document. You are expected to solve both the questions in the notebook and in this problem set, i.e., complete all questions from both sections A and B.

- The homework is due at 11:59 PM on the due date. We will be using Canvas for collecting homework assignments. For section A, feel free to submit a PDF, either of a scanned copy of your handwritten solution or a typed solution (using Microsoft Word, Latex, etc). Contact the TAs if you have technical difficulties in submitting the assignment. Section A homework should be submitted as a single pdf using the name convention **yourFirstName-yourLastName.pdf**.
- We encourage you to not submit late so that you don't accrue late days. Please declare the number of late days used for this homework at the top of your report. **Refer to the late days policy on the canvas page for more information.** *TLDR: HWs are to be done individually, and each student gets up to 5 late days to use as per their discretion. No more than up to 3 late days can be allocated to a specific HW.*
- For each question, please include all necessary calculation steps. If the result is not an integer, round your result to 3 decimal places.
- Please use the discussion section on Piazza (`https://piazza.com/class/l74tedj5el86tp`) to ask questions about the homework. Also, feel free to e-mail us at `cs5824-g@cs.vt.edu` and come to office hours.

Section A [60 pts]

You are expected to submit your solutions to this section as a PDF file. LaTeX is strongly preferred, but is not mandatory. The solutions need to be neat and legible. You should provide reasoning for your solution and show all your work. Except for obvious solutions, points will be deducted if there is no reasoning provided.

Problem 1. MLE (10 points total)

Q1. (5 points) Assume that there are N observations $x_1, x_2, \ldots, x_N$, which are i.i.d sampled from the same underlying distribution. Suppose that the underlying distribution is Bernoulli $Ber(p)$. Derive the MLE estimator of p with details.

Q2. (5 points) Assume that there are N observations $x_1, x_2, ..., x_N$, which are i.i.d sampled from the same underlying distribution. Suppose that the underlying distribution is Poisson distribution with PMF

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

1. (3 points) Derive the MLE estimator of λ with details.
2. (2 points) Let Y be a discrete random variable drawn from a Poisson distribution. Derive the expectation of Y.

Problem 2. Linear Regression (30 points total)

Table 1: A small data set about vehicle speed

No	1	2	3	4	5	6
number of wheels	4	4	2	8	4	3
cost (dollars)	15000	25000	5000	40000	22000	17000
max mph	160	150	70	100	200	150

Consider the dataset shown in Table 1. We want to train a Linear Regression model using the given dataset as training data, where the first two properties (e.g., number of wheels and vehicle cost) are treated as features and the third property (e.g., max miles per hour, mph) is treated as the target which we want to predict. Here, we use $x^{(i)}$ to denote the input features of the i-th sample in the dataset (e.g., ${x_j}^{(i)}$ is the j-th feature of the i-th sample) and $y^{(i)}$ to denote the label of the corresponding sample. The linear model $h_w(x)$ can be viewed as a trainable function of the input feature x of the form

$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2,$$

where the w_i's are the parameters of the linear model.

Q1. (5 points) First, we want to standardize the features by removing the mean and scaling to unit variance. The standardized feature $\hat{x}$ of a feature x can be calculated as

$$\hat{x} = (x - \mu)/\sigma,$$

where μ is the mean of the feature scores and σ is the standard deviation of the feature. Calculate the standardized features of the given samples in the dataset.

Q2. (5 points) Suppose we want to use a quadratic cost function $J(w)$ for training the linear model (least squares). Write down $J(w)$ as a function of the features x_i, linear model h_w, and labels y_i.

Q3. (5 points) Why is least squares potentially suitable for this problem? You may answer this question either intuitively or theoretically, or in whatever ways make sense.

Q4. (5 points) We will now use gradient descent to optimize the linear model. Calculate the partial derivative $\frac{\partial}{\partial w_i} J(w)$ of the cost function $J(w)$ w.r.t. the parameter w_i.

Q5. (5 points, bias and variance) Suppose that we only use the number of wheels as the input feature to fit the following two models on the given dataset:

1. $h_w(x) = w_0$
2. $h_w(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$

What kind of problems do you expect to encounter while fitting the dataset using these two models? Explain your statement accordingly in terms of **bias and variance**.

Q6. (5 points) What could you do to improve both models?

Problem 3. Logistic Regression (20 points total)

Q1. (5 points) You are a data scientist that wants to develop a binary classifier to classify a patient based on the features which include the observed symptoms captured in the data vector x. A certain disease can be labeled as infectious (class C_0) or non-infectious (class C1). In this context, describe what each of $P(C_0|\mathrm{x}), P(\mathrm{x}|C_0), P(C_0)$ means w.r.t. your patients.
(Refer to the Bayes rule $P(C_k|x) = P(x|C_k)P(C_k)/P(x)$ as discussed in class.)

Q2. (5 points) Assume a binary classification model (i.e., examples are represented as feature vectors x and are classified as either C_0 or C_1) of the following form

$$P(C_1|\mathrm{x}) = \frac{1}{1 + \exp(-\mathrm{w}^\top \mathrm{x})},$$

where w is a vector of weights. Find $P(C_0|\mathrm{x})$ and $\log\left(\frac{P(C_1|\mathrm{x})}{P(C_0|\mathrm{x})}\right)$.

Q3. (10 points) Consider another binary classification problem. Now we have N pairs of data points $\{(\mathrm{x}_1, C_1), ..., (\mathrm{x}_N, C_N)\}$, where $\mathrm{x}_i \in \mathbb{R}^d$ is a feature vector and C_i is its binary class label (i.e., C_i is either 0 or 1). We will consider using Logistic Regression for this problem. Suppose

$$y_i = \mathrm{w}^\top \mathrm{x}_i + w_0$$

and the prediction is based on the Sigmoid function $\sigma(\cdot)$

$$\sigma(y_i) = \frac{1}{1 + \exp(-y_i)}.$$

If $\sigma(y_i) \geq 0.5$, x_i is predicted as a data point from class 1; otherwise class 0.

1. (3 points) What is the likelihood function for this problem ?
2. (4 points) Prove that the log-likelihood function can be rewritten as

$$log(\mathcal{L}(\mathrm{w}, w_0)) = \sum_{i=1}^{N} C_i(\mathrm{w}^\top \mathrm{x}_i + w_0) - \log[1 + \exp(\mathrm{w}^\top \mathrm{x}_i + w_0)]$$

3. (3 points) Compute the derivatives of the log-likelihood function w.r.t. w and w_0, respectively. *(you may refer to the Matrix Cookbook for concepts on matrix computation).*