Assignment 1

CS 5824: Advanced Machine Learning
Spring 2022

Due Date: Feb 15th, 2022

- Feel free to talk to other members of the class when doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution only yourself. Please try to keep the solution brief and clear.
- Please use the discussion section on Canvas (`https://canvas.vt.edu/courses/145285/discussion_topics`) first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- The homework is due at 11:59 PM on the due date. We will be using Canvas (`https://canvas.vt.edu/courses/145285`) for collecting homework assignments. Please do NOT hand in a scan of your handwritten solution, only the typed solution (e.g., Microsoft Word, Latex, etc) will be accepted for grading. Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework!
- The homework should be submitted as a **single** pdf using the name convention `yourFirstName-yourLastName.pdf`.
- For each question, you will NOT get full credit if you only give out a final result. Necessary calculation steps are required. If the result is not an integer, round your result to 3 decimal places.

Problem 1. Probability Basics (**20** points total)

(a) (5 points) Given that X is a discrete random variable with PMF $p_X(x)$, and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function, write down the expectation $E[g(X)]$ of $g(X)$.

(b) (5 points) Formally, the variance of a random variable X can be defined as:

$$Var[X] \triangleq E[(X - E(X))^2]$$

Derive the following alternative form for the variance from the above definition:

$$Var[X] = E[X^2] - E[X]^2$$

Write down your solution with details.
Hint:

(1) $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$

(2) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$

(c) (5 points) The covariance of two random variables X and Y can be defined as the following form:

$$Cov[X, Y] \triangleq E[(X - E[X])(Y - E[Y])] \tag{1}$$

It can also be rewritten in the form such that

$$Cov[X, Y] = E[XY] - E[X]E[Y] \tag{2}$$

Write down your steps to derive $Cov[X, Y]$ from form 1 to form 2.

(d) (5 points) Show that $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.

Problem 2. Probability Basics & Data Statistics (**25** points total)

(a) (5 points) Consider the following dataset shown in Table 1.

Table 1: The apple dataset.

No.	Size	Color	Shape
1	Small	Green	Irregular
2	Large	Red	Circle
3	Large	Red	Irregular
4	Small	Green	Circle
5	Large	Green	Circle
6	Small	Red	Circle
7	Small	Red	Irregular

Estimate the conditional probabilities for $P(\text{Size} = \text{Small}|\text{Color} = \text{Green})$, $P(\text{Color} = \text{Red}|\text{Shape} = \text{Circle})$, and $P(\text{Shape} = \text{Irregular}|\text{Size} = \text{Large})$ using the data from Table 1.

(b) (5 points) Given the samples in Table 1, what is the probability of $P(\text{Size} = \text{Small}|\text{Color} = \text{Red})$? Calculate the conditional probability in terms of $P(\text{Color} = \text{Red}|\text{Size} = \text{Small})$, $P(\text{size} = \text{Small})$ and $P(\text{Color} = \text{Red})$ using Bayes' theorem.

Table 2: The housing dataset.

No.	1	2	3	4	5
living area (feet^2)	2104	1600	2400	1416	3000
# bedrooms	3	3	3	2	4
price (1000$s)	400	330	369	232	540

(c) (10 points) Considering the housing dataset shown in Table 2, compute the following statistics properties for both living area and price.

(1) (6 points) Mean, mode, and median.

(2) (4 points) Variance (both sample and population).

(d) (5 points) Find the covariance (population) between living area and price.

Problem 3. MLE (**25** points total)

(a) (15 points) Assume that there are N integers $k_1, k_2, ..., k_N \in \mathbb{Z}$, which are i.i.d sampled from the same underlying distribution. Suppose that the underlying distribution is a Poisson distribution with PMF

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

(1) (10 points) Derive the MLE estimator of λ with details.

(2) (5 points) Let X be a discrete random variable with the Poisson distribution. What is the expectation of X?

(b) (10 points) Assume that there are N integers $k_1, k_2, ..., k_N \in \mathbb{Z}$, which are i.i.d sampled from the same underlying distribution. Suppose that the underlying distribution is a Bernoulli distribution $Bernoulli(p)$. Derive the MLE estimator of p with details.

Problem 4. Linear Regression (**30** points total)

Considering the dataset shown in Table 2. Suppose that we want to train a Linear Regression model using the given dataset as training data, where the first two properties (e.g., living area and the number of bedrooms) are treated as features and the third property (e.g., price) 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 $h_\theta(x)$ can be viewed as a trainable function of the input feature x such that:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

where the θ_i's are the parameters parameterizing the linear model.

(a) (5 points) Suppose that 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:

$$\hat{x} = (x - u)/s$$

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

(b) (5 points) Suppose that we are using the ordinary least squares as our cost function $J(\theta)$ for training the linear model. Write down $J(\theta)$ as a function of the features $x^{(i)}$, linear model h_θ, and labels $y^{(i)}$.

(c) (5 points) Why is ordinary least squares potentially suitable as the cost function? You can answer this question either intuitively or theoretically, or in whatever ways that make sense.

(d) (5 points) Following previous questions, we additionally use gradient descent to optimize the linear model. Calculate the partial derivative $\frac{\partial}{\partial \theta_i} J(\theta)$ of the cost function $J(\theta)$ regarding the parameter θ_i.

(e) (5 points) Following previous questions, suppose the learning rate is α, use pseudocode to describe how gradient descent works to update the model parameters.

(f) (5 points) Suppose we only use the living area as the input feature and we are using the following two models to fit the given dataset:

(1) $h_\theta(x) = \theta_0$

(2) $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$

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.