ECE 449 Assignment 1 (Conceptual Questions)

Ruiqi Li <3180111638>

March 12, 2021

1 Problem 1

For each of the following tasks, identify which type of learning is involved (supervised, reinforcement, or unsupervised) and the training data to be used. If a task can fit more than one type, explain how and describe the training data for each type.

- **(a) Recommending a book to a user in an online bookstore**

  1. Supervised learning can be involved. One approach of the training set can be: the characteristics of a user (e.g., the age, gender, browsing history, etc) are input vectors (features), the possible category of that user is the label. So it is a classification problem. Once we have the order of different categories of a user, we can recommend some popular books from those most possible categories (for example, the 5 most possible categories) to the user.

  2. Reinforcement learning can be involved. If we construct a user-bookstore interact environment, we can define the reward and penalty based on the choice the user makes. We can firstly categorize the user randomly, then reward the model if the user confirms the assumption. Therefore there is no training set and all the parameters are learned from the immediate interaction.

- **(b) Playing tic tac toe**

  Reinforcement learning can be involved. A chess game is basically a state machine. If we build a player-player interact environment and define different reward and penalty for different states, we can train the model based on each steps and condition. Therefore, there is no training set and all the parameters are learned from the immediate interaction.

- **(c) Categorizing movies into different types**

  1. Supervised learning can be involved. One approach of the training set can be: given the artificially defined categories of a bunch of movies, we let the basic characteristics of the movies (e.g., duration, shooting time, language, etc) be the input features and the categories are the labels. Then we run the classification models.

  2. Unsupervised learning can be involved. If we don't know the predetermined labels of each movie, then we can only run a cluster algorithm to see the structure of the data. Movies of different categories may automatically form a cluster and therefore intuitively belong to one type. The training set can be the basic characteristics of a movie (e.g., duration, shooting time, language, etc).

- **(d) Learning to play music**

  Supervised learning can be involved.

  If "play music" stands for recommending a preferred music to a user, it's similar to question (a). The input features can be the basic characteristics of a user and the label is the category.

  If "play music" stands for writing music, then the problem is more complicated. For one example, if the "music" stands for a song with sounds (meaning it's a audio file and can be played) and we want to generate this "sound" from a text of music score, then we can let a music score (or sequence of pitches or lyrics) as one sample of input feature, which is well annotated (the time stamp, the duration, etc.). The label is then a waveform of the music (or the MEL frequency spectrum, specifically). We can train the model by feeding in small sequences of these data to learn the generater.

- **(e) Credit limit: Deciding the maximum allowed debt for each bank customer**

  Supervised learning can be involved. Basically, it's a regression problem. The input feature can be the basic characteristics of a user, like the age, gender, credit record, current income, etc. The label is therefore the objective debt limit. We want to fit the input data to these debt limits.

2 Problem 2

State the key difference between linear regression and logistic regression. Provide one example of problem linear regression and logistic regression can solve respectively. The key difference is that the logistic regression applys the sigmoid function to a linear regression output, so it can deal with non-linear problem and give a categorized output. Also, the linear regression makes the assumption that the input data confirms to the Gaussian distribution and compute the error in a Gaussian way, while the logistic regression assume that the input data confirms to the Bernoulli distribution.

Example: The linear regression can predict the house price in a certain area given the house floor area and crime rate; The logistic regression can classify if a client is high-risk or not given the credit record, investment amount and so on.

3 Problem 3

Assume we have 3 points, i.e., (1,2), (2,1), (3,2), in a 2-D Euclidean space. We want to fit a line which is y = w0 + w1x with respect to these 3 points. Derive the optimal solution for w0 and w1 with mean square error (MSE) loss. Show your steps for full score.

The model is

$$y = w_0 + w_1 x$$

and the predictions should be

$$\hat{y}_1 = w_0 + w_1$$
$$\hat{y}_2 = w_0 + 2w_1$$
$$\hat{y}_3 = w_0 + 3w_1$$

The overall cost is

$$J = \frac{1}{6}((w_0 + w_1 - 2)^2 + (w_0 + 2w_1 - 1)^2 + (w_0 + 3w_1 - 2)^2)$$

where

$$\frac{\partial J}{\partial w_0} = \frac{1}{6}(6w_0 + 12w_1 - 10)$$
$$\frac{\partial J}{\partial w_1} = \frac{1}{6}(12w_0 + 28w_1 - 20)$$

let the both be 0, then we have

$$w_0 = \frac{5}{3}$$
$$w_1 = 0$$

4 Problem 4

Assume we have 4 points, i.e., (-2,1), (-1,0), (1,0), (2,2), in a 2-D Euclidean space. We want to fit a 2nd order polynomial function y = w0 + w1x + w2x2 2with respect to these 4 points. Remember the closed form solution would be , where w = [w0; w1; w2]T . Write down what would be H and y.

We can augment the data by adding a additional column of x^2. The final result is

$$H = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

$$y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

5 Problem 5

Explain the process of gradient descent algorithm. What is the difference between gradient descent algorithm and gradient ascent algorithm?

Process:

- Initialize the parameters.
- Compute the prediction.
- Compute the loss function, which is usually the dissimilarity between the prediction and the ground truth.
- Compute the gradient of the loss function corresponding to the parameters.
- Update the parameters by subtracting the gradient multiplied by a factor (learning rate)
- Repeat the above steps for a certain number of times, a local minimum should be reached.

In the gradient descent algorithm, the parameters are updated by subtracting the gradient multiplied by the learning rate, while in the ascent algorithm it's addition. The gradient descent algorithm finds the local minimum, while the gradie ascent algorithm finds the local maximum.

6 Problem 6

Describe one advantage and one disadvantage respectively for a large learning rate and small learning rate in gradient descent algorithm.

- One advantage for a large learning rate: the model will converge fast.
- One disadvantage for a large learning rate: the model may get trouble in converging. It may not be able to find the minimum.
- One advantage for a small learning rate: The computation is therefore very subtle, resulting in a higher possibility to find the minimum.
- One disadvantage for a small learning rate: the model will converge slowly.

7 Problem 7

What is the entropy of a fair four-sided die?

$$H(X) = -4 \times \frac{1}{4} \times \log_2 \frac{1}{4} = 2$$

8 Problem 8

	A	B	C	D
UIUC	1/8	1/16	1/32	1/32
MIT	1/16	1/16	1/16	1/16
Colombia	1/8	1/8	1/16	1/16
Harvard	1/8	0	0	0

Table 1: Feature Importance

ZJUI institute has collected previous records of scores in course ECE449 and their postgraduate admission and made a prediction on the probability. The prediction is shown in the table below. Base on the table, calculate the conditional entropy H(X|Y), H(Y|X), and mutual entropy I(X; Y). X denotes ECE449 grades and Y denotes admission university.

Answer:

For each grade:

$$P(X) = \{\frac{7}{16}, \frac{1}{4}, \frac{5}{32}, \frac{5}{32}\}$$

For each university:

$$P(Y) = \{\frac{1}{4}, \frac{1}{4}, \frac{3}{8}, \frac{1}{8}\}$$

Therefore, we have

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \approx 1.859$$

$$H(Y) = -\sum_{j=1}^{n} p_j \log_2 p_j \approx 1.906$$

For mutual information, we have

$$H(X,Y) = -\sum_{i}\sum_{j} p_{ij} \log_2 p_{ij} \approx 3.563$$

Therefore,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \approx 0.202$$

$$H(X|Y) = H(X) - I(X;Y) \approx 1.66$$

$$H(Y|X) = H(Y) - I(X;Y) \approx 1.704$$

For conditional entropies for each individual universities and grades, the computation is similar and the results are

$$H(X|Y = \text{UIUC}) = 1.75$$

$$H(X|Y = \text{MIT}) = 2$$

$$H(X|Y = \text{Colombia}) \approx 1.918$$

$$H(X|Y = \text{Harvard}) = 0$$

$$H(Y|X = \text{A}) \approx 1.95$$

$$H(Y|X = \text{B}) = 1.5$$

$$H(Y|X = \text{C}) \approx 1.522$$

$$H(Y|X = \text{D}) \approx 1.522$$

9 Problem 9

Please see the attached another file, thank you!