COMP 432 Intro. to Machine Learning (Fall 2024)

Major Assignment #1

Due: 11:59PM, September 30th, 2024

Note Your will be submitting two separate files from this assignment as follows:

(a) **One(1) .pdf file:** containing answers to Question as well as reported results from coding you develop. Include snapshots of the pieces of code you developed in the appendix.

(b) **One(1) .zip folder:** containing all developed Python codes including a README.txt file on explaining how to run your code.

Theoretical Questions

Question 1 Answer, in a detailed manner, each of the following questions:

(a) Define the Turing Test.

(b) What is the difference between Classification and Regression in Machine Learning?

(c) What are the basic components of Machine Learning? Give a clear explanation for each component.

(d) What is the difference between Supervised and Unsupervised learning? Give examples for each type.

(e) What is the difference between Overfitting and Underfitting?

(f) What is the learning rate when training a ML model? and how does it affect the learning process?

(g) What is the difference between Gradient Descent and Stochastic Gradient Descent?

Question 2 (a) Consider a linear regression problem with the absolute error (or L1 error) function. The error associated with a single training sample with input $\mathbf{x}$ and target value y is given as:

$$J(\mathbf{w}) = |y - \mathbf{w}^{\mathbf{T}}\mathbf{x}| = |y - (w_0 x_0 + ... + w_i x_i + ... + w_n x_n)| \tag{1}$$

You are tasked with developing a gradient-descent learning rule for the above objective function. Your rule should be in the form:

$$w_i \leftarrow w_i - \eta??? \tag{2}$$

(b) Assume you have a problem with a dataset of two data points $\mathbf{x_1}$ = [0.9, 0.8, 0.4, 0.1] and $\mathbf{x_2}$ = [0.5, 0.2, 0.9, 0.8] and targets y_1 = 0.2 and y_2 = 0.9. You aim to train a linear regression model using the absolute error function, with the initial weights $\mathbf{w}$ =[0.5, 0.5, 0.5, 0.5] and a learning rate of 0.1 (assume no bias weight). Assume that the weights are updated after processing each data point, and that your model is trained with two epochs (meaning that your weights are updated 4 times). In each step of the training process, compute the output, the error, and the updated weights. Show your work.

Question 3 Consider a classification problem with three possible classes: A, B, and C. For a batch of 5 sample data points, your neural network produces the following outputs. Each output represents a probability distribution over the 3 classes.

(a) Using categorical cross-entropy, calculate the **average** loss for the batch of five samples. Show your work.

Sample	Output			True Label
	Class A	Class B	Class C	
1	0.7	0.2	0.1	B
2	0.1	0.2	0.7	A
3	0.3	0.6	0.1	B
4	0.8	0.1	0.1	C
5	0.2	0.3	0.5	A

Table 1: Table of Predicted Probabilities (Output) and True Labels

(b) Assuming the model chooses the class of the highest probability as the output class, what is the classification accuracy of the classification model represented in the table? Show your work.

Question 4 You are given with a dataset of the following eight (x, y) data points:

[(0.5,0.57), (0.98,0.5), (1,0.98), (1.4,0.89), (5,-0.75), (6.1,0), (5.5,-0.45), (6.5,0.2)]

(a) Using Mean Squared Error (MSE) as your criteria, determine which of the following candidate functions best fits the given data: $f_1(x) = e^x$, $f_2(x) = x^2$, $f_3(x) = cos(x)$, and $sin(x)$. (note: for cos(x) and sin(x), the inputs are in radians). Show your work.

(b) Inspect your results visually. Plot the data points and show how they fit the curves of the different candidate functions. Elaborate on your results.

Question 5 Consider a min-dataset of two data points $\mathbf{x_1}$ = [5, 3, 1, 3] and $\mathbf{x_2}$ = [1, 5, 1, 20] and targets $y_1 = 1$ and $y_2 = 0$. Your goal is to train a logistic regression model with the binary cross-entropy (BCE) loss function, a learning rate of 0.2, and all initial weights having a value of 0.5. You are to consider biases in your calculations.

(a) Given input $\mathbf{x_i}$, what is the output function of a logistic regression model? What is the BCE loss function? What is the gradient of the BCE function?

(b) Train a logistic regression model using the provided dataset. Use a batch of size 2 to train your model in 2 epochs (i.e. each weight is updated twice). Show your work.

Implementation Questions

Question 1 You are tired of paying exorbitant health insurance premiums every year. Your goal is to train a machine learning model that can accurately predict health insurance prices for individuals based on attributes such as age, sex, region, etc. You can use this dataset to train your model.

(a) Use statistical methods and graphs/plots to describe your daataset.

(b) Split your dataset into train and test sets with a 7:3 ratio. Use the train_test_split tool from scikit-learn.

(c) Build and train a Linear Regression model using scikit-learn. Explore the parameters of the model in scikit-learn, and aim for higher accuracies.

(d) Evaluate the performance of your model on both the train and test sets (separately). You can use scikit-learn's mean squared error tool.

Question 2 Your tasked with developing a ML model for lung cancer prediction. Given information about the patient, such as their sex, age, allergies, etc, your model should predict whether or not they have lung cancer. You can use this dataset to train your model.

(a) Use statistical methods and graphs/plots to describe your daataset.

(b) Split your dataset into train and test sets with a 7:3 ratio. Use the train_test_split tool from scikit-learn.

(c) Build and train a Logistic Regression model using scikit-learn. Explore the parameters of the model in scikit-learn, and aim for higher classification accuracies.

(d) Report and discuss the performance of your developed model on both the train and test sets (separately). You can use scikit-learn's classification report tool.