ECON425: Machine Learning
Winter 2022

Question 1. Typical machine learning algorithms can be used to address both supervised or unsupervised proble these algorithms are either continuous values or discrete labels. There pes of machine learning algorithms, as summarized in the following

1. Supervisec
2. Supervisec
3. Unsupervi
4. Unsupervi

Please read the following problems that can be potentially solved by the above machine learning algorithms.

- **Financial forecasting**: to predict the stock value of a company based on the historical stock records and the company's sale records.
- **Company Analysis:** to group a set of companies into multiple clusters so that the companies in the same cluster share similar properties or features.
- **Social network relationships**: to group social media users so that the users in the same group have similar social or network properties (e.g., geo-location, log on time, frequencies of posting).
- **Product recommendation**: to predict the interestingness (real-valued scores between 0 and 1) of a customer for a product based on using users' purchase records.
- **Face recognition**: to recognize the identity of a person from his/her facial photo and retrieve his/her profile from a customer database.
- **Global seismic monitoring**: to predict if earthquake will happen for a certain area and a time-period. The inputs include sensory data and history seismic data.

For each problem, please select the appropriate algorithm type, and write its index (1-4) after the problem index, in the following format:

A:____B:____C: ____D:____E:____F:____

Question 2. Linear Regression is one of the most popular supervised learning methods. It aims to learn a linear function from training samples through minimizing a cost function.

Let denote a single feature (e.g., area) for the i-th sample (e.g., a house), denote its output label (e.g., price). Let denote the parameters of the linear regression model, returns the regressed label of the training sample . To find the optimal model parameters , we might develop methods to minimize the cost function or the cost function , where returns the absolute value of x . Please discuss the differences between the two cost functions: F() and G(), and analyze their advantages or disadvantages while being used as cost functions.

Question 3. Gradient descent algorithm is one of the most widely used optimization framework in machine learning. It usually starts with initializing model parameters and then updates these parameters following their gradient directions. Taking the least square loss as an example, the update equation for is,

where m is the number of training samples used for estimating the gradient of , is the learning rate. The implementation of gradient descent methods might be varying with specific choices over the hyper-parameters, e.g., , or the number of training samples to use. **Please explain the differences** between three variants of gradient descent method, i.e., full-batch, mini-batch and online learning. Tips: [illegible] this question by analyzing how these variants might change the above up[illegible]

Question 4 In Linea[illegible] use the polynomial terms of existing features. Taking the following objecti[illegible]ce, the last three terms are polynomial.

Please discuss when [illegible] polynomial terms and how these terms affect the learning and testing [illegible]

Question 5 Suppose you are applying the Logistic Regression method for fish classification. Each fish sample is represented using a single feature, denoted as . m is the number of testing samples. Consider two classes: 1, salmon; 0, otherwise. The prediction function is defined as:), where , as shown in the following figure. For the feature, outputs its confidence of belonging to the class 1.

Let . The table below shows the features of 10 testing samples and their ground-truth classes.

(a) For each sample , please classify it to be class 1 if is larger than 0.5; class 0, otherwise. To do so, you need to calculate for each sample. Note that . Please write down the binary class label of each sample.

(b) Please further calculate the accuracy, and per-class recall rate, and per-class precision rate. (Do not miss any metrics.)

	3.2	2.5	2.0	0.5	0.6	0.2	0.3	0.7	3.5	4.2
	1	1	1	1	1	0	0	0	0	0

: ground-truth class; : predicated class.

TIP: you might first calculate the confusion matrix and then report the above metrics. No need to include the confusion matrix in your answer.

Question 6. One might choose ROC curves to evaluate a classifier's performance. A ROC curve essentially plots how the false positive rates of a classifier change along with the true positive rates. The following figure shows the ROC curves for three binary classifiers over the same testing dataset, respectively. Please answer the following two questions.

Question 6.1. Which classifier (A, B, or C) achieve the overall best performance?

Question 6.2. Consider an application where the risk of having false negative predictions is extremely high. False negatives are also called miss detections in different scenarios. One might want to avoid false negative predictions to lower the risk. Please discuss whether classifier B is a better choice than classifier C for this application. You might need to check the definitions of true positive rate/false positive rate as introduced in the lecture slides.

Question 7. To apply hods to train deep neural networks, one need to run forward propagation ation with one or multiple training samples. Modern neural networks usu reds of thousands of neurons and many deep learning platforms (e.g., tens the so-called *computational graphs* to implement forward propagation and background propagation. For a neuron of the network, the partial derivatives of the loss function with respect to its neuron parameters are calculated as the product of *local gradients* and *upper-level gradient* received by this neuron.

Consider a network for binary classification tasks. The network takes three features as its inputs and outputs a binary class label (positive: +1, negative: -1). The network includes two neurons: **neuron M and neuron S**. The neuron M's output is defined as:

where are the network parameters, and is the 3-dimensional feature vector of a sample. No bias term nor activation function is used in this neuron. Please note that

The neuron M's output is used as the input to the Neuron S. The neuron S's output is defined as:

where is the sigmoid function. Note that the output of neuron S h(x) is used as the input to the neuron M and the output of M is used as the prediction for the same x.

Consider a training sample with the feature vector: $\mathbf{x}$=(3, -1, 1) whose class label is: y=+1. The current network parameters $\mathbf{w}$=(1.0, -2.0, -4.5). The following figure illustrates the computational graph for the above neural network. Please answer the following two questions.

Question 7.1. For the training sample $\mathbf{x}$=(3, -1, 1), suppose the upper level gradient received by the neuron S is 0.25, please calculate upper-level gradient received by the neuron M (illustrated as B in red)

Note that:

- The derivative of a sigmoid function with respect to its input variable is:

- Sigmoid(0.5)=1/(1+exp(-0.5))=0.62

Question 7.2. Following the above question, please further calculate the gradient of the loss function with respect to (illustrated as C in red). Note that, like the previous question, there is only one training sample involved in the calculation of gradients. Please note that the neuron M does not employ any activation function.

Question 8: Recurrent neural networks (RNNs) or its variant Long Short-Term memory (LSTM) are widely applied in financial forecasting. Consider a sequence of yearly inflation rates for USA between 2010 and 2020:

(1.64, 3.14, 2.07, 1.47, 1.62, 0.12, 1.26, 2.14, 2.44, 1.81, 0.62).
where the inflation rates of 2010 and 2020 are 1.64 & 0.62 respectively. The task is to forecast the inflation rate of USA in 2021.
The following sample codes are used to train a LSTM model using the TensorFlow module.

Please explain what [illegible]cript**. Please also discuss, given the original sequence of inflatio**[illegible]re the training data X and y, which are needed for training the LSTM n[illegible]n_len=4 for this task, **please list two pairs of training samples (x, y)** where [illegible] and y is the target label.

Question 9: Cross-validation is considered as an effective way to select the most appropriate model for a given machine learning problem. Please answer the following problems.
(a) Please discuss what model overfitting and model under-fitting mean in model selection.
(b) How could you determine if a model is overfitting to the training samples?
(c) Please provide three possible ways to address the overfitting issue if any. You might consider model complexity, dataset, and use of regularization terms.

Question 10: For binary classification problems, we could apply logistic regression to regress the target class label of a given sample, or directly find a separating hyperplane in the feature space.

- Please explain the concept of max-margin learning which can be used for finding separating hyper-planes. (1-2 sentences)
- Please explain how slack variables function to enhance the robustness of a max-margin classifier. (1-2 sentences)
- Please discuss the differences and similarities between logistic regression and support vector machine (2-4 sentences)