Midterm Examination Results

ⓘ Correct answers are no longer available.

Score for this quiz: **49.5** out of 45
Submitted Oct 13, 2021 at 3:56pm
This attempt took 106 minutes.

Partial

Question 1

6 / 6 pts

The following pandas dataframe *df* consists of information about product sales of a particular company. There is one row per customer per product. Items with the same *ITEM_ID* have the same price. The first five rows of *df* are shown to you.

Note: The size of the blank is smaller than we like it to be. Feel free to write in another text editor and paste it in here

CUST_ID	ITEM_ID	UNIT_PRICE	QTY
62	10089	10	37
62	10029	32.7	11
19	83421	15	15
12	95571	23.11	200
19	10089	10	50

(a) Write the python code which will give you the total number of customers this company has. ***(3 points)***

```
len(df['CUST_ID'].u
```

(b) Write the python code which will identify the top five items (in a dataframe) with highest gross earnings. ***(3 points)***

df['EARN'] = df['UN

Answer 1:

len(df['CUST_ID'].unique())

Answer 2:

df['EARN'] = df['UNIT_PRICE'] * df['QTY'];
df.groupby('ITEM_ID').sum().reset_index().sort_values('EARN', ascending = False)['ITEM_ID'][:5]

Partial

Question 2

6 / 6 pts

You are running k-means clustering for k=2 on a dataset with a single feature. Using an initial Cluster A centroid of 4 and Cluster B centroid of 11, do two iterations of k-means clustering on a single feature dataset with features values = [1,4,6,11,21] to answer the following questions. (***6 points total***)

(a) Which feature values belong to Cluster A after one iteration?

1,4,6

(b) Which feature values belong to Cluster A after a second iteration?

1,4,6

(c) Would you expect your assigned clusters to change if you were to do a third iteration of k-means? Why or why not?

No; they would not

(d) What is the silhouette coefficient for the data point with feature value ‘6’ after the 2nd iteration?

0.65

Answer 1:

1,4,6

Answer 2:

1,4,6

Answer 3:

No; they would not change because there was no change between the first and second iteration which means that the clusters were already stable.

Answer 4:

0.65

Question 3

3 / 3 pts

You are using a decision tree classifier and one of the features is the GPA of the student. If there are 511 possible distinct splits of this feature (according to Gini Index), how many distinct GPA values are present in the dataset? (***3 points)***

512

Incorrect

Question 4

0 / 2 pts

You are training a decision tree on the below dataset consisting of two classes (True and False) and one continuously valued numeric feature (A). Using the Gini Index, determine the best split point of Feature A. ***(2 points)***

Feature A	Label
1	True
2	False
8	True
4	True
10	False

8

Best split is at 9, with a Gini of 0.3, which is the lowest of the 4 possible split points.

We wanted you to show work for at least some of the Gini score calculation.

Something like 4/5 * (1- (3/4)^2 - (1/4)^2) + 1/5 * 0 = 0.3.

Question 5

2 / 0 pts

[EXTRA CREDIT] In a decision tree, is it possible for overall entropy to increase after a split? Justify your answer. (***2 points)***

Your Answer:

No; It is not possible to have overall entropy increase after split because each split of a decision tree occurs only if there is a decrease in data impurity.

No, it's impossible. The purpose of entropy is to improve the purity of the data, with each
split making the data more orderly.

Partial

Question 6

4 / 4 pts

You have a dataset which contains images of cats, dogs, monkeys and rabbits. The images are 32 x 32 pixels and serialized into a vector of 1024 features per image. You have to build a neural classifier for identifying which of the four animals are present in an image. Answer the following four questions. (***4 points + 1 extra credit point)***

(a) How many nodes are in the input layer of the neural network?

1024

(b) How many nodes are in the output layer of the neural network?

4

(c) What is a reasonable activation function for the output nodes?

Softmax

(d) What is a reasonable loss function for this classification task?

Cross Entropy

(e) Assuming a single hidden layer with 10 nodes, how many total weights are in this neural network, including bias weights? (**extra credit**)

10294

Answer 1:

1024

Answer 2:

4

Answer 3:

softmax

Answer 4:

Cross Entropy

Answer 5:

10294

Question 7

2 / 2 pts

What are *two* typical criteria for stopping the training process in a neural network? (***2 points)***

Your Answer:

The two criteria are:

1. A specified maximum number of epochs has occurred

2. The difference in the errors or the difference between the expected and desired output between epochs is below some threshold

The main answers we were looking for were

(1) There is some minmum decrease in loss
(2) A pre-specified number of epochs has expired

but we accepted some others.

Question 8

2 / 2 pts

What is an *epoch when training a* neural network? (***2 points)***

Your Answer:

Epoch: a measure of the number of times the training vectors have been used one time all the way through through to update weights

In terms of artificial neural networks, an epoch refers to one cycle through the full training dataset.

Question 9

4 / 4 pts

The following 10 row table shows the ground truth and predicted class labels for a sentiment classification task. Calculate the F-1 score of the classifier. Provide your answer up to the 4th decimal place (***4 points total)***

Sample #	Ground Truth	Prediction
1	Positive	Positive
2	Negative	Negative
3	Positive	Positive
4	Positive	Negative
5	Positive	Negative

6	Negative	Positive
7	Positive	Negative
8	Negative	Negative
9	Negative	Negative
10	Positive	Negative

0.4444

Question 10

2 / 0 pts

[EXTRA CREDIT] In an application for predicting whether a person has a rare disease, it is known that about 1 of every 10^3 patients tested are infected with the disease. If we were to build a classifier for the prediction, which metric would be a better measure of performance: Accuracy or F-1 score? Justify your answer. (***2 points)***

Your Answer:

The F-1 score would be a better measure than the accuracy because the F-1 score equation is (2 * Precision * Recall) / (Precision + Recall). Accuracy is instead (True Positive + True Negative) / (Positive + Negative). The F-1 Score values false positives and false negatives more than accuracy does. With a rare disease, it is more important that people who have the disease are not getting a negative result. It would be better to have more people falsely scored as having the rare disease than those with it to be marked as not having the disease.

Accuracy = (TP + TN) / (P + N) = (TP + TN) / (TP + TN + FP + FN)
F-1 = 2 * precision * recall / (precision + recall) = 2 * TP ^ 2 / [(TP + FP) * (TP + FN)]
F-1 Score would be better.
As there is about 1 of every 10^3 patients is infected with the disease, which means that TP would be really small. Thus, Accuracy is approximate to TN / (TN + FN) or we could say Accuracy is always approximate to 1, which is nonsense for the task.
F-1 takes both precision and recall into account. And we want the model have the high
precision and recall in the same time as we only tell people who are infected by the disease.

Incorrect

Question 11

4 / 4 pts

For each of the following scenarios, explain which algorithm you would employ for the task and why. In some of these scenarios, more than one algorithm may be applicable. You are expected to list *any one*. You will be given credit as long as you are able to correctly justify your choice. (***4 points)***

Note: The size of the blank is smaller than we like it to be. Feel free to write in another text editor and paste it in here

(a) Learning the operation of addition from examples of FeatureA and FeatureB and label = FeatureA + FeatureB

Neural Network: Be

(b) You are part of a committee tasked with investigating the admissions policies of a University. The data you have at your disposal are the application materials for applicants over the past five years and the admissions decision that was made.

Decision Tree: The

(c) Predicting the best next move in chess given millions of board states and winner information from chess games played online.

Neural Network: We

(d) Predicting the average weather in Florida two weeks from today given a dataset of average temperatures for every state, every day, for the last two years.

Linear Regression:

Answer 1:

Neural Network: Because we are learning patterns with labelled data.

Answer 2:

Decision Tree: The data has decisions at our disposal, so predictions can be made with a training set of data that develops decision rules.

Answer 3:

Neural Network: We are working with a giant set of data which neural networks work well on for classification

---

Answer 4:

Linear Regression: We are working with continuous data so a linear regression would be good for prediction

The most appropriate answers are

(a) linear regression / neural network
(b) decision trees
(c) neural network
(d) linear regression / neural network

Some other answers were accepted.

Incorrect

Question 12

2.5 / 3 pts

For each of the following algorithms, **describe** any two hyperparameters. (Don't just simply list variable names like *K* or *c*). ***(3 points)***

(a) K-Means

n_clusters is the nu

(b) Random Forest

n_estimators is the

(c) Neural Network

hidden_layer_sizes

Answer 1:

n_clusters is the number of clusters that we want to form; random_state is used to set a random seed that will result in reproducible results.

Answer 2:

n_estimators is the number of trees in the forest; max_depth is the maximum depth of the tree.

Answer 3:

hidden_layer_sizes has each element corresponding to its index number hidden layer with each index of hidden_layer_sizes indicating how many neurons are in that index number hidden layer; solver is the solver for weight optimization.

random_state is not a hyprparam

Question 13

1 / 1 pts

State TRUE or FALSE: ***(1 points)***

Stochastic Gradient Decent is a way to fit a neural network, represented by its weights, to data.

- (●) True
- ( ) False

Question 14

1 / 1 pts

State TRUE or FALSE: ***(1 points)***

The number of hidden nodes in a neural network cannot be more than the number of data points in the training set.

- ( ) True
- (●) False

Question 15

1 / 1 pts

State TRUE or FALSE: ***(1 points)***

Overfitting is when an algorithm predicts the test set better than the training set.

- ○ True
- ◉ False

Question 16

2 / 2 pts

How is bagging different from boosting? Mention *at least two* ways in which they are different. ***(2 points)***

Your Answer:

- The boosting method allows for parallel training of the models in the ensemble while bagging does not.

- Boosting pays closer attention to misclassified samples while bagging does not and rather has an equal weighted collection of models.

- Boosting trains sequentially while bagging does not.

Training: For bagging, each model can be trained in parallel. For boosting, models will be
trained one after another.
Prediction: All prediction functions in bagging are equally weighted, while boosting has
different weights for different classifier.

Question 17

2 / 2 pts

Match each of the terms on the left with the appropriate definition on the right ***(2 points)***

Boosting	An ensemble method
Bagging	An ensemble method
Bootstrapping	Sampling a dataset w
Blending	An ensemble method

Question 18

2 / 2 pts

Explain why an ensemble created with blending has more opportunity for overfit than an ensemble created using a simple combiner. Please keep your answer to fewer than five sentences. (**2 points**)

Your Answer:

Blending has more opportunity for overfit than an ensemble created using a simple combiner. Blending can overfit to the validation set which may cause an overfitted overall model on the test set. In contrast, a simple combiner does not restrict to a specific model and doesn't require learning as it combines predictions through simple averaging or other non-trainable combiner and is therefore less prone to overfit.

Blending enables more complexity and flexibility in how the models' predictions are combined. It involves more learned parameters and complexity than simple combination, which is not parametric. With additional complexity and training always comes the possibility of overfitting. Blending also might have the tendency to rely on a particular model more (which might be overfit). But a simple combiner will average all the models with equal weights, hence not introducing additional possibilities of overfitting.

Question 19

3 / 0 pts

[**EXTRA CREDIT]** Describe the procedure for training a Stacking ensemble model and using it to make predictions on a test set. (***3 points)***

Your Answer:

- First, the stacking trains and predicts using cross-validation for the training set

- Secondly, the same models are trained on the entire training set

- Next, predictions are made on the test set using these models

- Then, a combiner is trained on the cross-validated training set predictions

- Lastly, the combiner is applied to the test set and predictions are made

(1) Train and predict using an ensemble of models using cross-validation on the training set
(2) Train those same models on the entire training set, make predictions on test set.
(3) Train a combiner on the cross-validated ensemble model predictions of the training set.
(4) Apply trained combiner to the test set predictions.

Quiz Score: **49.5** out of 45