Question 2

Complete

Marked out of 4.00

The following objective function is used during the construction of decision trees.

$$G(Q_m, \theta) = \sum_{i=1}^{b} \frac{|Q_{m,i}(\theta)|}{|Q_m|} H(Q_{m,i}(\theta))$$

Explain what the summation and the fraction are achieving. You do not have to explain every individual symbol; only explain the key parts and the meaning of the expression.

summation of all the impurities of every branch at node Qm. The fraction is the weight, ratio of the amount of data at branch Qmi compared to all the data at node Qm.

Question 3

Complete

Marked out of 4.00

In the decision tree learning algorithm, there is recursive call of the form `DTree(examples_i, features \ {F})`. Explain what `features \ {F})` means and why it is needed. Explain if the type of feature (numeric vs categorical) has any effect on this.

Given a binary decision tree, DTree(examples_i, features \ {F}) is all features excluding the ones already in set F. If set F was the left branch in node Qm, then features \ {F} is needed to find the right branch. Numeric vs Categorical would not matter if this was a binary decision tree then there can be a single numeric or categorical condition that only gives two possible outputs to split the data at Qm into left and right branches.

Question 4

Complete

Marked out of 4.00

One way of obtaining an optimal linear regression model is to use *normal equations*. Write one advantage and one disadvantage of using this method. In each case write one or two sentences.

Advantage: Can compute in one passing.

Disadvantage: Not efficient as taking O(n^3) because of matrix multiplication.

Question 6

Not answered

Marked out of 5.00

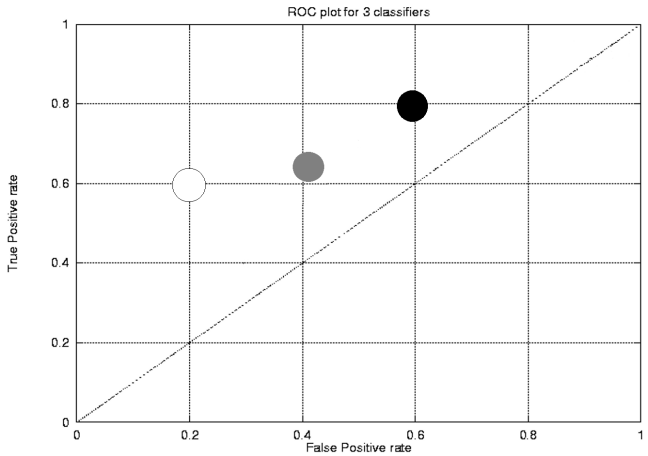
A common measure for evaluating learnt regression models is R^2. What range of values does this metric take and what do they mean?

Question 7

Complete

Marked out of 5.00

Consider the three classifiers in the ROC space pictured below:



The classifiers are marked with white, gray, and black circles. Even though the gray classifier is not dominated by the white or black classifiers, it is not part of the ROC convex hull. Explain why and what this implies in terms of finding a minimum-cost classifier.

grey is not part of the convex hull as it doesn't have a better TPR or FPR compared to white and black. this means the minimum cost classifier will be the classifiers white and black on the convex hull and not grey as grey is never better.

Question 8

Complete

Marked out of 4.00

Suppose we want to create a neural network library. Explain two main features (advantage) of using PyTorch (or a similar platform) as opposed to implementing the algorithms directly in Python. Very briefly explain how these features are achieved. [Do not use existing neural network functionalities in PyTorch as a reason in your argument.]

1. you only need to define the forward pass and can generate the graph on the fly and find the gradient for you.

2. manipulating data to leverage peak CPU and GPU performance with efficient PyTorch tensor operations that already exist.

Question 9

Complete

Marked out of 5.00

Explain the essential idea behind Echo State Networks and how they are trained.

The essential idea is that we fix each of the connections in a recurrent neural network, except for the hidden to output connections, to random values and only learn only the hidden to output connections.

The hidden to hidden weights are set so that the length of the activity vector stays about the same after each iteration. Sparse connectivity is used, so most of these weights are set to zero. This allows information to stay in one part of the network without being propagated to another part of the network too quickly. The scale of the input to hidden connections must be chosen carefully as they need to drive the oscillators in the hidden unit without wiping information from the network's recent history that they already contain.

Since the last layer is a linear model, training an echo state network is so fast that we can try out many different options for the weights and the sparseness of the hidden to hidden connectivity.

Question 10

Correct

Marked out of 3.00

Convolutional Neural Networks (CNNs) are commonly used for classification learning in images. In their most simple form, CNNs have a repeating pattern in their earlier layers. Complete the following template.

Input layer

Convolution ✔

Relu ✔

Pooling ✔

More hidden layers ...

Output layer

Drag and drop the following layers in the right order:

Your answer is correct.

Question 11

Complete

Marked out of 6.00

Enumerate three advantages of SVMs over using feedforward neural networks for classification. For each advantage write one or two sentences.

1.
SVMs can be Effective in High-Dimensional Spaces, even when the number of features is larger than the number of training examples (this is because the kernel trick transforms the input data into higher dimensions where the data can become separable)

2.
SVM is more robust to outliers, as data points that are far away from the decision boundary have minimal impact on the model

3.
SVM can be more interpretable, as a subset of the training data (the support vectors) are used to define the decision boundary

Question 14

Complete

Marked out of 4.00

In the context of Bagging, answer the following questions:

a) Why is Bagging more suitable than Boosting for parallelisation (using multiple machines/CPUs)?

b) In the Bagging algorithm presented in the course, the sampling is "with replacement". What happens if we use sampling "without replacement"?

a) Bagging has its weak learners trained in parallel, they are independent of each other so suitable for parallelisation compared to Boosting. Boosting has its weak models trained sequentially because it takes into account (and is dependent on) misclassifications in previous weak models.

b) If n was the size of the whole dataset, bagging needs to sample with a replacement for a sample dataset of size n too. If we use sampling "without replacement", then will take all n examples in one sample, and cannot produce more samples, so only have one weak mod el.

Question 15

Complete

Marked out of 4.00

In the context of the AdaBoost algorithm, answer the following questions:

a) During model generation, some instances are multiplied by e/(1-e). What range of values does this expression take and what does it achieve?

b) In prediction, class weights are updated by - log e/(1-e). Why is there a negative sign? What does this achieve?

a) e/(1-e) is a range of values between 0 and 1. Instances in the dataset that were classified correctly have their weights multiplied by this value which helps reduce its rate as don't need the next model to focus on this correct classification as much as compared to the misclassified. Higher weight for misclassification to give more focus to improve on, and lower weights to correct classified.

b) negative sign as [- log e/(1-e)] returns a positive value and adds more positive values to help achieve more confident class weighting.

Question 16

Correct

Marked out of 4.00

Complete the pseudo-code of the k-means algorithm by dragging and dropping the statements in the correct order:

```
randomly pick k centroids;
while convergence is not achieved:
    compute the distance of each point to each centroid;
    assign points to centroids;
    compute the mean of each cluster;
```

Drag and drop these items:

Your answer is correct.

Question 17

Complete

Marked out of 4.00

Suppose we have a collection of clustering algorithms and we want to evaluate their performance. Each clustering algorithm is implemented as a function of the form `cluster(dataset, k)` which takes a dataset that is an *n by m* array (n data points in an m-dimensional numeric space) and a number *k* which is the desired number of clusters. The function returns a list of length *n* where each element is a number between 1 and *k* inclusive and assigns a cluster number to the corresponding point in the input dataset.

We evaluate each algorithm using a classification dataset as the following:

- Given a labelled dataset of size *n by (m + 1)* where the last column is the class label, we call the last column *y* and remove it to obtain an unlabelled *n by m* dataset;
- The unlabelled dataset is fed to the clustering algorithm with k set to the number of classes, to obtain an *n by 1* assignment vector. We call this vector c.
- We put the vectors *y* and *c* side by side and compare the corresponding elements. Each pair of elements that match is considered a +1 score. By looking at all *n* pairs of elements, the clustering algorithm gets an score out of *n*. Dividing the score by *n* gives us a normalised score between 0 and 1.

Identify two issues with this way of evaluating clustering algorithms. Explain each issue in a few sentences.

Issue 1: double up of the error. if a y1-c2 doesn't match, that means there must be another y2-c1 that also doesn't match. So error can be disproportional in the evaluation.

Issue 2: clustering is an unsupervised learning method to find patterns in data when there are no class labels. In this case, the algorithm has access to label classes so it is not trying to find new patterns in the data (knowledge discovery), but instead comparing if the results from the learning match what is already labelled, which is not a useful evalution to leverage clusterings unsupervised learning strength/application. Instead should be measuring the quality of the clustering, and wanting tight compactness between datapoints in a cluster and clear separation of clusters.

Question 18

Partially correct

Marked out of 5.00

Consider a binary classification problem with four input features. Two of these features are binary and the other two are: Colour (with a domain of "red", "green", and "blue") and Size (with a domain of "small", "medium", and "large"). Suppose we use *H* and *R* to denote the set of all hypotheses and the set of all codes for *H* respectively. Answer the following questions with integers or by selecting an item from the drop-down list. Do not use words or expressions.

- The input space has a total of
  36
  ✔ elements.
- If we use conjunction of constraints, the size (cardinality) of *R* is
  41
  ✖ and the size of *H* is
  26
  ✖ .
- For each hypothesis represented in the form of conjunction of constraints there is one or more ✔ equivalent representation(s) in the form of a decision tree.
- If we use decision trees and we have a training dataset with 24 examples, there are
  4096
  ✔ hypotheses that are consistent with the training set.

Question 19

Complete

Marked out of 5.00

Describe two ways in which learning the version space is different from learning a model in supervised learning algorithms such as decision trees and neural networks.

set S and G implicitly define the version space.

Question 20

Complete

Marked out of 5.00

What does "no free lunch" theorem in machine learning sate? Answer in one sentence.

that there is no best machine learning algorithm as they all have their pros and cons and different applicability

Question 21

Complete

Marked out of 6.00

Throughout the course you have seen a number of ways to prevent overfitting in certain learning algorithms (or family of algorithms). Name two of these methods/ways. For each algorithm briefly explain how it can overfit the data and how the said method can prevent it. Your two solutions should not be regarding the same algorithm.

1. decision tree can overfit as it has no representation bias so it can take a lot of parameters into consideration that may be irrelevant. The method to prevent is pruning where branches that don't have improved impurity can be removed, decreasing the number of params and preventing overfitting.

2. convolution and pooling, where there can be heaps of data and param so can be overfitting - very specified to training example. Pooling can prevent overfitting as its reduction in spatial dimensions reduces sensitivity/specificity to small variances in the training input by taking the max or average of a local area instead, decreasing the number of parameters and preventing overfitting.

Question 22

Complete

Marked out of 6.00

Name two learning schemes (algorithms, models, methods, architectures) that are considered both supervised and unsupervised. For each case explain why.

1. GAN. unsupervised as discriminator trained a label data.

2. Autoencoders

Question 23

Complete

Marked out of 4.00

Name two learning schemes (algorithms/models) that are universal function approximators. For each case briefly explain why.

1.

2.