EXAM 2 PART 1 SOLUTION

Q1
What is label smoothing in logistic regression and neural networks? Why is it useful?
Ans1
The activation function used for the logistic regression is usually a sigmoid function which is given by the equation sigmoid(x) = 1/1+e^(-x)
In case of binary classification where the output label is either 0 or 1, we want the output from this function to map to either 0 or 1. In order to have output of 1, the value of x should be infinite. However if we use label smoothing, these binary labels are converted to some other values between 0 and 1, (lets say 0.1 and 0.9). Now in order to have output of 1, the value of x can be ln(9) which is much easier for the logistic regression model to converge to.

Q2
Suppose we have a neural network with a training set of size 100,000. If we use a batch size of 100 and train the network for 10 epochs, how many times will the weights of the network be updated during training?
Ans2
10,000
Explanation2
Train size = 100,000
Batch size = 100
In 1 epoch, weights will be updated (100,000/100 =) 1,000 times.
In 10 epochs, weights will be updated (1,000*10 =) 10,000 times.

Q3
How many trainable parameters (weights and biases) does a network that classifies MNIST (784 features, 10 classes) and has a single hidden dense layer with 100 units have?
Ans3
79510
Explanation3
Input size, i = 784
Size of hidden layer, h = 100
Output size, o = 10
Number of parameters = connections between layers + biases in every layer
= (i * h + h * o) + (h + o)
= (784 * 100 + 100 * 10) + (100 + 10)
= 79510

Q4
Consider the dataset in the table in which we have a 2 class problem with 8 training examples with 3 attributes. Recall the in decsion tree learning, we place at the root the attibute with hiighest information gain. If we sort the atttributes by increasing information gain, what ordering will we obtain? Hint - there is no need to perform any calculations.

A1	A2	A3	Class
A	C	E	0
B	C	F	0
A	C	F	0
B	C	F	0
A	D	F	1
B	D	E	1
A	D	E	1
B	D	E	1

Ans4
A1, A3, A2
Explanation4
The values of the attribute A2 are already split according to the classes. So the information gain will be the highest.
Almost all the values of attribute A3 are split class wise. So the information gain of A3 will be less than A2.
Finally the information gain of A1 will be the least.
If we sort the information gain in increasing order, we will get A1, A3, A2.

Q5
Why are decision and regression trees used more often than other models to build ensmbles?
Ans5
Ensembles tend to work where each individual classifier in the ensemble performs better than random guessing. Also, if all of the individual classifiers make errors in classifying the same type of input, there is not much point in using ensembles. In case of decision trees, due to the nature these trees are built, individual trees do not produce the same results like other algorithms (for eg. mlp) does. This improves the predictive power of ensembles built using decision and regression trees and hence are used more often than other models. The other advantage of using decision and regression trees is that each individual classifier can be trained relatively fast compared to some other algorithms.

Q6

Consider the following training dataset:

A1	A2	y
100	10	0
1000	9	1
0	4	2
5	0	3

If we apply min-max rescaling to every attribute, the attributes of the first training example, which originally are [100,10] will become:

Ans6

[0.1, 1]

Explanation6

The formula for min-max rescaling is (value – min(attribute)) / (max(attribute) – min (attribute)).
Given sample is [100,10]. So, A1 = 100 and A2 = 10
Min-max rescaling for A1:
Value = 100
Min(A1) = 0
Max(A1) = 1000
According to the formula, the value of A1 (=100) will be 0.1
Min-max rescaling for A2:
Value = 10
Min(A1) = 0
Max(A1) = 10
According to the formula, the value of A2 (=10) will be 1

Q7

The first layer of a convolutional neural network to classify color images has 32 3X3 filters. How many trainable parameters does the layer have?

Ans7

896

Explanation7

As the images are colored images, so there will be three channels.
Number of parameters = num_filters*filter_height*filter_width*num_channels + biases
= 32*3*3*3 + 32
= 896

Q8

If the input of the layer in the previous question (32 3x3 filters) is a 30x30 color image, the size of the resulting feature map will be:

Ans8

32x28x28

Explanation8

In the resulting feature map, the height is *img_height – filter_height + 1* and the width is *img_width – filter_width + 1*
The resulting feature map = 32 * (30-3+1) * (30-3+1)
= 32 * 28 * 28

Q9
The support vectors in an SVM are a subset of the original training examples
Ans9
True

Q10
When adding a model to an ensemble, bagging takes into consideration the performance of the rest of the models in the ensemble to choose the examples to prioritize (choosing with higher probability) in the new model
Ans10
False

Q11
The coefficients in logistic regression can be found analytically, without having to search.
Ans11
False

Q12
Increasing the number of trees in a random forest usually leads to overfitting
Ans12
False

Q13
Given a training set (X,y), k-means will ignore the target function y and build a model using only the attribute values X
Ans13
True

Q14
Principal component analysis finds a subset of the original features that minimizes information loss
Ans14
False, it finds a linear combination of the original features

Q15
What are the advantages of word embeddings over a one-hot word representation?
Ans15
One-hot word representation does not consider the relative similarity or dissimilarity between words while word embeddings does. In case of word embeddings, the Euclidean distance between two similar words is much smaller than the distance between two words not similar to each other. This improves the performance of the machine learning models on tasks such as classification where the similarity between words gives information to the model. Also, word embeddings are of much shorter length than one-hot word representation in situations where there are many different words in the training sample.

Q16
Explain two techniques for improving the results of an image classifier when the training set is small
Ans16
Any two techniques that can be used to improve the results of an image classifier when the training set is small are:

1. Data Augmentation: By implementing methods such as shifting, rotation, and flipping an image, training set can be augmented to make it larger. However, we must be careful while implementing

some of these techniques. For example: Flipping digits vertically in MNIST Handwritten Digit dataset may decrease the performance of model in distinguishing 6 from 9.

2. Semi supervised Learning: In case where we have training samples, but the labels are not available for these samples, we train a classifier with samples containing the labels. Then we predict the label for the samples that does not have labels. If the confidence of the samples belonging to one of the class is high, we add that sample/label pair to the training sample to increase the number of samples in training set.