Final Exam

This is a partial exam of the Data Science course. There are 25 multiple-choice questions. All the questions have the same weight. There is only one correct or best answer for each question. Below are the general instructions:

- You will have a maximum of 2 hours to complete the exam. Students who have applied for special accommodation receive 30 minutes extra time.
- You are not allowed to leave the exam room during the first 30 minutes and last 15 minutes of the normal exam schedule.
- You can use the Jupyter Notebook to perform computations.
- You can use the “?” feature in the Jupyter Notebook to check the documentation of functions.
- You cannot use any reference materials (e.g., smartphone, books) except your one-page double-sided A4 cheatsheet.
- Read each question and all options very carefully.
- Double-check your answers before submission. We suggest answering all questions. We also suggest that you prioritize the questions that you are confident in and then strategically guess the answers for others that you are not confident in.
- Avoid spending too much time on a difficult question. It is better to complete other questions first and then go back to finish the difficult ones.
- You must keep the exam content confidential and are not allowed to copy or distribute the content in any form.

We have defined several terms in the table below. You can refer back to them when the questions mention these terms. The feature (i.e., the input) is x. The ground truth (i.e., the true output) is y. The predicted output is z.

Term	Equation
Identity activation function $f(x)$	$f(x) = x$
Sigmoid activation function $f(x)$	$f(x) = 1/(1+e^{-x})$
Tanh activation function $f(x)$	$f(x) = (e^x - e^{-x})/(e^x + e^{-x})$
Hinge loss function $f(y,z)$	$f(y,z) = \max(1-yz, 0)$
Squared error loss function $f(y,z)$	$f(y,z) = (y-z)^2$
Perceptron loss function $f(y,z)$	$f(y,z) = \max(-yz, 0)$
Logistic loss function $f(y,z)$	$f(y,z) = \log(1+e^{-yz})$

Binary cross-entropy loss function $f(y, z)$	$f(y,z) = (1-y)\log_2(1-z) - y\log_2(z)$
Softmax activation function $f(x)$, where x_i means the i^{th} element in array x, and n means the total number of elements in array x	$f(x_i) = e^{x_i} / \sum_{j=1}^{n} e^{x_j}$
Entropy H for a coin with two sides (one side has probability p_1, and another side has probability p_2)	$H = p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2)$

General Data Science Questions

1.0p 1 Which of the following is a commonly used loss function for regression?

○ Perceptron loss

○ Logistic loss

○ Cross-entropy loss

○ Squared error loss

1.0p 2 Which of the following best describes the purpose of data augmentation?

○ To perform feature selection

○ To reduce the model training time

○ To reduce the size of input (e.g., images)

○ To combat overfitting by increasing dataset diversity

Structured Data Processing Questions

1.0p 3 Suppose we flipped a coin 100 times, and we got 50 heads and 50 tails. What is the entropy of the result in this coin-flipping experiment?

○ 0.25

○ 0.5

○ 0

○ 1

1.0p **4** Which of the following statements about the Decision Tree model is **TRUE**? Notice that we refer to the general Decision Tree, not a specific implementation.

○ Decision Tree can only be used for the classification task.

○ Decision Tree cannot handle continuous features.

○ Decision Tree can be trained well without problems using the misclassification error as the node-splitting strategy.

○ When splitting nodes, Decision Tree can use the same feature multiple times.

1.0p **5** Suppose that we fit a binary classification model in identifying spam and ham (i.e., non-spam). Spam is the positive label, and ham is the negative label. The following shows the evaluation result of the model. What is the f-score of the model based on the evaluation result?

- 30 samples are predicted as spam, and they are indeed spam in reality
- 20 samples are predicted as spam, but it turns out that they are not spam in reality
- 70 samples are predicted as ham, but it turns out that they are spam in reality
- 80 samples are predicted as ham, and they are indeed ham in reality

○ 0.3

○ 0.55

○ 0.6

○ 0.4

1.0p **6** Suppose that we are performing a prediction task using artificial neurons. Which of the following statements about artificial neurons is **TRUE**?

○ Support Vector Machine uses the sigmoid activation function with the perceptron loss function.

○ The perceptron classifier uses the identity activation function with the hinge loss function.

○ Logistic regression uses the identity activation function with the binary cross entropy loss function.

○ Linear regression uses the identity activation function with the squared error loss function.

Text Data Processing Questions

1.0p 7 Which of the following about word embeddings in natural language processing is **TRUE**?

○ Word embeddings require manual feature engineering to capture the meaning of words effectively.

○ Word embeddings can only be learned using supervised methods, which means we need labels for the sentences in the training data.

○ Word embeddings typically represent each word as one integer value to capture semantic relationships between words.

○ Word embeddings can be trained by iteratively using one word in a sentence to predict nearby words as accurately as possible.

1.0p 8 Which of the following about text processing is **TRUE**?

○ Lemmatization simply chops word tails to obtain the word's base form.

○ Tokenization represents each word as a data point in a high-dimensional space.

○ Stemming can correctly identify the original form of each word.

○ POS tagging labels the role of each word in a particular part of speech, such as verb.

1.0p 9 What is the primary goal of using dot product in the context of word embeddings?

○ To normalize word vectors so that the output is a probability distribution.

○ To find the frequency of words in a text document.

○ To predict the probability of a word belonging to a certain topic.

○ To measure the similarity between two words.

1.0p 10 What is the output of topic modeling using Latent Dirichlet Allocation?

○ A set of clusters, where each cluster represents a group of similar documents.

○ A set of word vectors, where each word is represented as a data point in a high-dimensional space.

○ A set of document vectors, where each document is represented as a bag-of-words vector.

○ A set of topic vectors, where each vector is represented as a probability distribution over the words in the text corpus.

1.0p **11** Which of the following about TF-IDF (term frequency-inverse document frequency) is **TRUE**?

○ The goal of using TF-IDF is to transform each word into a vector so that we can use the vector for further tasks, such as classification.

○ IDF weights each word by considering how frequently it shows in different documents. IDF is lower when the word appears in fewer documents.

○ If a word appears in only a few documents (and frequently in these documents), TF-IDF can show that the word contains less information and should be less important.

○ TF-IDF can be seen as a weighted Bag of Words approach, which means TF-IDF weights the term frequency based on the inverse document frequency.

Image Data Processing Questions

1.0p **12** Given an image, which of the following kernels will blur the image after performing the typical convolution operation (no padding, stride 1) using the kernel? Assume that numpy is imported.

○ numpy.array([[0, 0, 0],[1, 0, 0],[0, 0, 0]])

○ numpy.array([[1, 0, -1],[2, 0, -2],[1, 0, -1]])

○ numpy.array([[0, 0, 0],[0, 2, 0],[0, 0, 0]])

○ numpy.array([[0.11, 0.11, 0.11],[0.11, 0.11, 0.11],[0.11, 0.11, 0.11]])

1.0p **13** Which of the following about Convolutional Neural Network (CNNs) is **TRUE**?

○ The activation functions in CNNs are used to reduce the dimensions of the features.

○ The number of trainable parameters in the convolution layers is roughly the same as the parameters in the fully connected layers.

○ The max pooling layer is used to help the network pay attention to the most important information, and there are trainable parameters.

○ The normalization layer acts as regularization during training and can make deep neural networks much easier to train.

1.0p **14** You are training a deep neural network for image classification. Which of the following is likely to happen if you initialize all the weights (i.e., the model parameters) in the network to zero? And what is the most likely reason for this to happen?

○ The loss and model performance metrics are decreasing slowly. This is because initializing the weights to a constant number makes the gradient smaller.

○ The loss and model performance metrics are good in both the training and validation set. This is because weights are initialized to the same value and do not have any biases during model training.

○ The loss and model performance metrics converge quickly, as the weights are not too far from their optimal values.

○ The loss and model performance metrics almost do not change. This is because the neurons learn the same features, and the network cannot tell the difference between different inputs.

1.0p **15** You are training a deep feedforward neural network for classifying images with clothes using a dataset with 10 classes. Your network structure has four hidden layers with sizes 512, 256, 256, and 128, respectively. The outputs from every hidden layer are passed to sigmoid activation functions. You use a learning rate 0.01 with the stochastic gradient descent optimizer. After trying hard to get the best possible performance, you found that the network can achieve only about 10% accuracy on the training and test set. Which of the following best describes what happened in this situation, and also what action to take?

○ The learning rate is too high, causing the network training process to be unstable. Decreasing the learning rate after some epochs may fix the problem.

○ The network underfits the training data. Use regularization techniques such as dropout or data augmentation will improve generalization performance.

○ The network is too shallow to capture the complexity of the image classification task. Increasing the depth of the network would likely lead to better performance.

○ The sigmoid activation function saturates when the input is large or small, causing the gradients to vanish. We can change the activation function to leaky ReLU.

1.0p **16** Given an input RGB image (3 channels) of size 33 by 33, a convolutional layer with 6 image filters (each filter has size 5 by 5), a stride of 2, and padding of 1. What is the dimensionality of the output of the layer? The following options are in the format of width x height x channel.

○ 32 x 32 x 6

○ 32 x 32 x 3

○ 16 x 16 x 3

○ 16 x 16 x 6

Deep Learning Questions

1.0p **17** Which of the following about Recurrent Neural Networks (RNNs) is **TRUE**?

○ RNNs are mainly designed to handle image data.

○ We can create an encoder-decoder structure using RNNs, and using the final encoder output is easy for the model to remember information from a long time ago.

○ RNNs cannot handle sequences longer than a fixed number of time steps.

○ RNNs contain hidden layers that can capture information from previous time steps.

1.0p **18** Which of the following about the PyTorch deep learning framework is **TRUE**?

○ PyTorch has no objects or functions to support automatic batching, data shuffling, or multi-process data loading. Users need to handle these by themselves.

○ PyTorch sets the gradients of all parameters to zero after each backpropagation step, which prevents the new gradients from being added to the previous ones. Users do not need to handle this by themselves.

○ PyTorch automatically uses GPU (Graphics Processing Unit) by default if the machine has GPUs, which means that users do not need to specify the device to use.

○ PyTorch can automatically compute gradients for tensors, which enables users to easily implement backpropagation and other optimization algorithms.

1.0p **19** Suppose we have a deep feedforward neural network with two layers. The first layer has four artificial neurons, and the second layer has one neuron. We want to use the network to perform binary classification on the data that is not linearly separable. This means that the dataset cannot be separated with a reasonable performance by just using a linear classifier. Which of the following settings can help us achieve our task well?

○ Use the tanh activation function for all neurons and use the squared error loss

○ Use the identity activation function for all neurons and use the squared error loss

○ Use the identity activation function for all neurons and use the cross-entropy loss

○ Use the tanh activation function for all neurons and use the cross-entropy loss

1.0p **20** Which of the following best describes the goal of using a softmax activation function?

○ To ensure that the gradients cannot be negative and the model converges faster.

○ To produce a binary output.

○ To ensure that the output is always between -1 and 1.

○ To map arbitrary input values to a probability distribution output.

1.0p **21** Which of the following is the order that best describes the attention mechanism?

- A: Compute the attention distribution using softmax
- B: Compute attention-weighted sum of encoder output
- C: Get the encoder output values (from the RNN)
- D: Compute attention scores (dot product similarity)
- E: Transform encoder outputs (dimension reduction)

○ ECDBA

○ CBDAE

○ CABDE

○ CEDAB

Coding Questions

1.0p **22** Given a 2-dimensional numpy array A with integers or floats. All values in the array are different. Assume that numpy is imported. What is the following code doing?

- numpy.argmax(A, axis=1)

○ Return an array containing the indices of the minimum value in each row of array A.

○ Return an array containing the minimum value in each column of array A.

○ Return an array containing the maximum value in each column of array A.

○ Return an array containing the indices of the maximum value in each row of array A.

1.0p **23** Given a pandas series object S with integers or floats. The index of each row means the time steps. What is the following code doing?

- S.rolling(3, min_periods=1, closed="right").sum()

○ Replace each value in series S with the sum of the current value and the maximum value in a window of size 3. The window is moved one position ahead at a time.

○ For each row in series S, compute the difference between each value and the average of the previous two values from the two preceding time steps.

○ Compute the cumulative sum of series S, where the value in each row is the sum of all values up to and including that value.

○ For each row in series S, compute the sum of the current value and the previous two values from the two preceding time steps. Then, store the sum in the current row.

1.0p **24** You are preprocessing text data for a document classification task. Given a pandas dataframe object D with two columns “c” and “t”. Each row means a document. The “c” column contains information about the class that the document belongs to, such as “Business”. The “t” column contains a list of tokens, such as ['Wall, St.', 'Bears', 'Claw', 'Back', 'Into']. What is the following code doing?

- D.explode("t").groupby(["c", "t"]).size().reset_index(name="n").sort_values(["c", "n"], ascending=[True, False]).groupby("c").head(5)

○ Remove rows with duplicate tokens in the “t” column. Return a new dataframe with a new column that counts the frequency of each token among all documents per class.

○ Compute the average frequency of each token in each document per class. Return a new dataframe with a new column that shows the 5 tokens with the highest average frequency.

○ Return a new dataframe with a new column that shows the 5 least used words per class, and their frequency.

○ Return a new dataframe with a new column that shows the 5 most used words per class, and their frequency.

1.0p **25** Given a pandas dataframe object D, where column “w” in each row contains a list of words, such as ['Wall, St.', 'Bears', 'Claw', 'Back', 'Into']. What is the following code doing? Variable “S” in the code below contains a list of words, such as ['the', 'of', 'a', 'on'].

- D["w"].apply(lambda x: [word for word in x if word.lower() not in S])

○ Select only the rows in the dataframe where the word list in the “w” column contains any word in list S.

○ Count the frequency of each word in the “w” column of the dataframe, and then return the words that occur less frequently than a threshold defined in list S.

○ For each word list in each row in the “w” column of the dataframe, add words that are in list S.

○ For each word list in each row in the “w” column of the dataframe, remove words that are in list S.