Mid-term 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** If data frame A and B both have no missing data, which of the following operations will definitely **NOT** produce missing data?

- ◯ Right join
- ◯ Outer join
- ◯ Left join
- ◯ Inner join

1.0p **2** Which of the following pairs is used to calculate precision?

- ◯ True Positive and False Negative
- ◯ True Negative and False Negative
- ◯ True Negative and False Positive
- ◯ True Positive and False Positive

1.0p **3** Suppose we fit a binary classification model to predict if a bad smell event will happen in the city for the next 8 hours. The positive label means that there will be smell events, and the negative label means no events. Which of the following statements is **TRUE**?

- ◯ If the precision of the model is very high, it means that the model catches almost all the smell events and does not miss them.
- ◯ If the recall of the model is very low, it means that the model is not very reliable when it predicts that there would be smell events in the future.

- If the model predicts that there would be NO event, but it turns out that the model made a wrong prediction, this is called a True Positive.
- If the model predicts that there would be an event, but it turns out there is nothing happening, this is called a False Positive.

1.0p **4** What is the main purpose of computing feature importance?

- To improve the accuracy of the model on the training set
- To prevent overfitting of the model
- To increase the number of features used in the model
- To determine which features are most important in making predictions

1.0p **5** Which of the following practices is **recommended** in the data science pipeline?

- Assuming that someone else has already framed the data science problem
- Using the same modeling technique to deal with all different types of data
- Sticking to a linear data science pipeline that starts from problem framing and data preparation to model building
- Visualizing and exploring data in addition to using descriptive statistics

1.0p **6** Suppose we flip a coin (with two sides) many times and we compute the entropy. Which of the following statements is **TRUE**?

- If we change the probability of one side of the coin (to make it appear more frequently or less frequently), entropy is not sensitive to this change in probabilities.
- Entropy is always one in this case because the coin has only two sides.
- Entropy reaches the minimum when the coin is fair, meaning two sides have equal probability.
- Entropy intuitively means the averaged surprise when we flip the coin.

Decision Tree and Random Forest Questions

1.0p **7** Which of the following is a common metric that is used to evaluate the quality of a node split in a Decision Tree model?

- Mean squared error

- ◯ F1 Score
- ◯ R-squared
- ◯ Entropy

1.0p **8** Which of the following statements about Random Forest is **TRUE**?

- ◯ Random Forest contains multiple Decision Tree models, and the best tree is used for performing the task.
- ◯ Random Forest contains multiple Decision Tree models that are trained identically using the same set of features.
- ◯ Random Forest is more likely to overfit the data than the Decision Tree model.
- ◯ Random Forest uses randomly selected features and bootstrapped samples (i.e., sample with replacement).

1.0p **9** What is the main **advantage** of using an ensemble of decision trees, such as a Random Forest, over a single decision tree in a classification or regression problem? Recall that errors of the model that we trained can be decomposed into bias, variance, and noise.

- ◯ There is not really an advantage
- ◯ Reduced noise in the error
- ◯ Reduced bias in the error
- ◯ Reduced variance in the error

1.0p **10** Suppose we want to train a Decision Tree based on the following dataset to predict whether Alex will go out or not. We use the misclassification error as the strategy when splitting a node. Which feature will the Decision Tree pick to split the first node?

Weather	Feeling	Wind	Time	Going out?
sunny	cold	calm	daytime	yes
rainy	warm	calm	nighttime	no
rainy	warm	windy	daytime	yes

rainy	cold	windy	daytime	no
rainy	warm	calm	daytime	no

- ◯ Wind
- ◯ Time
- ◯ Feeling
- ◯ Weather

Neural Networks and Deep Learning Questions

1.0p **11** Which of the following about the Gradient Descent algorithm is **TRUE**?

- ◯ Gradient Descent is guaranteed to find the global minimum.
- ◯ The mini-batch size determines the number of iterations required to find the optimal parameters in the neural network.
- ◯ Gradient Descent is used to minimize the error/cost in regression problems only.
- ◯ The learning rate determines how large or small the step will be when updating model parameters.

1.0p **12** What is the purpose of regularization when performing Gradient Descent?

- ◯ To reduce the number of features in the dataset
- ◯ To schedule the learning rate appropriately
- ◯ To speed up the algorithm and save computer memory
- ◯ To prevent the algorithm from overfitting the training data

1.0p **13** What is the main purpose of using an activation function in a neural network?

- ◯ To reduce the dimensionality of the input data
- ◯ To regularize the model parameters
- ◯ To scale the input features to a consistent range

- ◯ To introduce non-linearity into the model

1.0p **14** Which of the following statements about the backpropagation algorithm when training neural networks is **TRUE**?

- ◯ Backpropagation is used to transform the weighted sum of the input non-linearly.
- ◯ Backpropagation is used to add new layers to a neural network.
- ◯ Backpropagation is only used to train the final layer of a neural network.
- ◯ Backpropagation iteratively updates the weights in previous layers in the neural network.

Regression and Classification Questions

1.0p **15** What is the purpose of using a loss function in classification or regression?

- ◯ To reduce the number of features
- ◯ To increase the accuracy of the model on the training set
- ◯ To regularize the model parameters
- ◯ To measure the error between the prediction and the ground truth

1.0p **16** What is the goal of minimizing the sum of squared errors in linear regression?

- ◯ To reduce the variance in the model and prevent overfitting.
- ◯ To ensure that the residuals (the differences between predicted and true values) are normally distributed.
- ◯ To find the optimal values of the input variables (features) that minimize the output variable (responses).
- ◯ To find the line that best fits the data by minimizing the distance between the predicted and true values.

1.0p **17** Which of the following statements about overfitting is **TRUE**?

- ◯ We can combat overfitting by increasing the size of the test set.
- ◯ Overfitting means fitting the test data extremely well.

- ◯ Most of the time, we can deal with overfitting by removing outliers.
- ◯ Overfitting usually happens when using a very complex model.

1.0p **18** Suppose we fit a simple linear regression model F using the training data with one feature X and one true response Y. We then use the model to output prediction Z=F(X). The coefficient of determination (R-squared) is equal to the square of:

- ◯ Pearson correlation coefficient between the predictions Z and the feature X.
- ◯ The coefficient of feature X.
- ◯ Pearson correlation coefficient between the true response Y and the feature X.
- ◯ Pearson correlation coefficient between the true response Y and the predictions Z.

1.0p **19** In the setting of fitting a model Y=F(X), Which of the following statements about the coefficient of determination (R-squared) is **TRUE**?

- ◯ R-squared intuitively means the unexplained variation for the response variable Y.
- ◯ A bad R-squared means that there is no pattern in the data.
- ◯ R-squared is a good evaluation metric for classification.
- ◯ R-squared increases as we add more features X.

1.0p **20** Which of the following about data modeling is **TRUE**?

- ◯ We can just use the training data to evaluate if the model will work well.
- ◯ When using Random Forest, we can usually put raw data into the model without feature engineering.
- ◯ R-squared is a common evaluation metric for classification models.
- ◯ Classification is used to predict categorical labels, and regression is used to predict continuous values.

Calculation Related Questions

1.0p **21** Suppose we want to classify if an image contains a banana or mango by using a Decision Tree model. We have 1 green banana image, 1 yellow banana image, 3 green mango images, and 3 yellow mango images. The Decision Tree uses entropy as the node-splitting strategy. Recall

that Information Gain is the difference between the parent node's entropy and the leaf nodes' averaged entropy. What is the Information Gain after we ask the question, "is the fruit color yellow or not"?

The formula of entropy H is given below:
$H = p_1 * log_2(1/p_1) + p_2 * log_2(1/p_2)$

- ◯ 0.5
- ◯ 1
- ◯ 0.33
- ◯ 0

1.0p 22 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.

- 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

What is the precision of the model based on the evaluation result?

- ◯ 0.4
- ◯ 0.55
- ◯ 0.3
- ◯ 0.6

Coding Questions

1.0p 23 Assume you have a pandas DataFrame named D containing time series data with inconsistent frequency. Which of the following code resamples the data to a frequency of one hour?

- ◯ D.resample("1Min", '1H')
- ◯ D.resample("1Min").asfreq()
- ◯ D.resample("60H")
- ◯ D.resample("60Min")

1.0p **24** Suppose we have a pandas data frame D with 100 rows and two columns (C1 and C2). Column C1 has no missing data, and column C2 has 25% missing data. We want to sum up all valid items in column C2 and get a single integer (not a pandas.Series or pandas.DataFrame). Which of the following code produces the desired output? For example, if D looks like the table below, the code should output only one single integer 27, which is a sum of 3, 4, and 20.

C1	C2
1	NaN
2	3
2	4
10	20

- ◯ D["C1"].dropna().sum()
- ◯ D.groupby("C2").sum()
- ◯ D.drop("C2", axis=1).sum()
- ◯ D.dropna().sum()["C2"]

1.0p **25** Suppose we have a pandas data frame D with two columns (Smell, and Zipcode). The data frame contains the complaints of bad smell that citizens submitted in Pittsburgh. The “Smell” column contains ratings of how bad the smell is at the corresponding timestamp. The “Zipcode” column means the zip code where the citizen submitted the report. We want to know the average number of smell ratings for ONLY the zip code 15213. Which of the following code produces the desired output? For example, if D looks like the table below, the code should output 4, which is the average of 3 and 5.

Smell	Zipcode
1	15208
2	15202
3	15213
4	15222

5	15213

- D.groupby("Zipcode").mean()["Smell"].iloc("15213")
- D.mean()["Smell"].groupby("Zipcode")["15213"]
- D["Smell"].groupby("Zipcode").mean()["15213"]
- D[D["Zipcode"]=="15213"].mean()["Smell"]