Congratulations! You passed!
TO PASS 70% or higher

Keep Learning

GRADE
100%

Week 2

TOTAL POINTS 7

1. A movie streaming system provides personalised recommendations. The system implements a feedback loop that considers the users choices to make it more robust. However, the system presents to the user a limited set of movies. More importantly, some movie options are never presented to the user. This is a case of a **biased system.** What kind of bias does this system introduce? — 1 / 1 point

- ○ Societal Bias
- ○ Activity Bias
- ◉ Selection Bias
- ○ Data Drift

Correct
Correct! The movie recommender system creates a feedback loop and only presents a limited number of options to the user which introduces selection bias.

2. A financial institution uses a loan approval machine learning system that was deployed 5 years ago to grant loans. In these 5 years, the economic conditions have changed drastically due to a recession. Interest rates have risen significantly from when the system was originally deployed. Which of the following problems that a typical ML system faces best describes what is happening with this loan approval process? — 1 / 1 point

- ○ Societal Bias
- ◉ Data Drift
- ○ None of the given options applies to the loan approval process
- ○ Activity Bias

Correct
Correct! Data drift occurs when the distribution of data in a deployed machine learning system changes statistically quickly enough to be very different from the initial set of data. In the loan approval system, the interest rates have changed, hence the system has to be updated to reflect these changes.

3. Suppose you are building a sentiment classifier to determine whether product reviews have positive, neutral or negative sentiments. From the star rating column, you realize that a disproportionate amount of the ratings are five stars (50% of the total ratings). What is the most likely metric you can use to measure the statistical bias in this scenario? — 1 / 1 point

- ○ Kullback-Leibler Divergence (KL)
- ○ Total Variation Distance (TVD)
- ○ Difference in proportions of labels (DPL)
- ◉ Class Imbalance (CI)

Correct
Correct! Class Imbalance measures the imbalance in the number of members between different attribute values. Models trained on an unequal class distribution like this one tend to be biased towards the majority class which can be harmful when the classification of minority classes are more valued than that of the majority class. This article provides more details and examples.

4. As a data scientist who works in a sales company, you are asked to predict the sales for December 2020. Your dataset contains daily sales data from January 2020 to September 2020. You successfully train your model on this data, but the actual sales measured in December are completely different from what your model predicted. These types of events are called **data drift.** — 1 / 1 point

After some research, you realize that this change occurred because there is always a sharp increase in sales due to the holiday season.

What specific kind of data drift best describes this change in sales?

- ○ Covariate shift
- ◉ Prior probability shift
- ○ Concept jump
- ○ None of the above

Correct

Correct! A prior probability shift occurs when there is a shift in the target variable.

5. Properly measuring statistical data bias is key to build fair and successful ML models. Select the most appropriate metric to detect an imbalance of positive outcomes between different facet values. **1 / 1 point**

- ○ Class Imbalance(CI)
- ○ Total Variation Distance (TVD)
- ○ Kullback-Leibler Divergence (KL)
- ◉ Difference in proportions of labels (DPL)

✓ **Correct**
Correct! This metric measures the imbalance of positive outcomes between different facet values. This article provides more details and examples.

6. You are tasked with creating a model that predicts the popularity of a new pair of socks. You suspect the presence of statistical bias in your dataset due to the product category. **1 / 1 point**

Which of the following tools from the AWS toolkit best provides the ability to detect statistical bias in the dataset?

- ○ AWS Glue
- ○ Amazon Athena
- ◉ Amazon SageMaker Clarify
- ○ All of the above

✓ **Correct**
Correct! Amazon SageMaker Clarify can be used to identify statistical bias during data preparation without having to write your own code.

7. Feature importance refers to a metric that assigns a score to input features based on how useful they are at predicting a target variable. **1 / 1 point**

True or false: applying feature importance always improves model performance.

- ○ True
- ◉ False

✓ **Correct**
That's right! While feature importance may lead to an improvement in model performance, sometimes removing features can reduce model performance.