Congratulations! You passed!

Grade received 100% **To pass** 80% or higher

Go to next item

Issues in Training Data

Total points 4

1. What formula represents a *dataset shift?* **1 / 1 point**

- (●) $P_{train}(y, x) \neq P_{serve}(y, x)$
- ( ) $P_{train}(y|x) = P_{serve}(y|x)$ and $P_{train}(x) \neq P_{serve}(x)$
- ( ) $P_{train}(y|x) \neq P_{serve}(y|x)$ and $P_{train}(x) = P_{serve}(x)$

Correct
Well done! The most generic case of distribution skews is when the joint distribution of inputs and outputs differs between training and serving.

2. What measure is typically used to determine the degree of *data drift*? **1 / 1 point**

- (●) Chebyshev distance (L-infinity)
- ( ) Euclidean distance (L2)
- ( ) Manhattan distance (L1)
- ( ) Hamming distance

Correct
That's it! Chebyshev distance is defined as $\max_i(|x_i - y_i|)$

3. **Distribution skew** occurs when the distribution of the training dataset is significantly different from the distribution of the serving dataset, and is typically caused by: (check all that apply). **1 / 1 point**

- [x] Trend, seasonality, changes in data over time.

Correct
Keep it up! Data distributions between training and serving often change and so this is another case of distribution skew.

- [x] Different data sources for training and serving data.

Correct
Way to go! Data sources between training and serving often change and so this is another case of distribution skew.

- [ ] Occurs when serving and training data don't conform to the same schema. For example, int32 != float.
- [x] Faulty sampling method that selects a sample for training which is not representative of serving data distribution.

Correct
Spot on! A faulty sampling mechanism that chooses a non-representative subsample is an example of distribution skew.

- [ ] A data source that provides some feature values is modified between training and serving time.
- [ ] There is different logic for generating features between training and serving. For example, if you apply some transformation only in one of the two code paths.

4. TensorFlow Data Validation (TFDV) helps TFX users maintain the health of their ML pipelines. TFDV can analyze training and serves data to: **1 / 1 point**

- [ ] Perform feature selection.
- [ ] Deploy pipeline to a mobile application.
- [x] Compute descriptive statistics.

Correct
Perfect! TFDV goes beyond computing relevant statistics, it also has nice browser-based visualization tools.

- [x] Infer a schema.

Correct
Nice going! In short, schemas describe the expectations for "correct" data and can thus be used to detect errors in the data.

- [ ] Perform feature engineering.
- [x] Detect data anomalies.

Correct
That's the way! TFDV can check your data for error in the aggregate across an entire dataset or by checking for errors on a per-example basis.