Google Cloud

Feature Engineering

Quiz Question Answers

Module 2: Raw Data to Features

Raw Data to Features and Good vs Bad Features

Question 1: Before being input into an ML model, raw data must be turned into:

*A: Feature vectors
Feedback: This answer is correct.

B: Feature matrix
Feedback: This answer is incorrect, please review the module again.

C: Multidimensional vectors
Feedback: This answer is incorrect, please review the module again.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 2: A good feature has which of the following characteristics?

A: It should be related to the objective.
Feedback: This answer is partially correct, please review the module again.

B: It should be known at prediction time.
Feedback: This answer is partially correct, please review the module again.

C: It should be numeric with meaningful magnitude.
Feedback: This answer is partially correct, please review the module again.

*D: All of the above.
Feedback: This answer is correct.

Question 3: True or False: Training data sets require several example predictor variables to classify or predict a response. In machine learning, the predictor variables are called features and the responses are called labels.

A: False
Feedback: This answer is incorrect, please review the module again.

*B: True
Feedback: This answer is correct.

Question 4: True or False: Different problems in the same domain may need different features.

A: False
Feedback: This answer is incorrect, please review the module again.

*B: True
Feedback: This answer is correct.

Prediction time, Numeric, Enough Examples, Human Sight

Question 1: Fill in the blanks: A good feature should be _____ and have _____.

*A: Numeric, meaningful magnitude
Feedback: This answer is correct.

B: Row-based, polynomial attributes
Feedback: This answer is incorrect, please review the module again.

C: Row-based, loss attributes
Feedback: This answer is correct.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 2: True or False: As a best practice, it is recommended that you have at least five examples of any value before using it in your model.

*A: True
Feedback: This answer is correct.

B: False
Feedback: This answer is incorrect, please review the module again.

Question 3: Select which statement is true.

A: Feature engineering is the process of transforming data into features to act as outputs for machine learning models such that good quality features help in improving the overall model performance.
Feedback: This answer is incorrect, please review the module again.

*B: Feature engineering is the process of transforming data into features to act as inputs for machine learning models such that good quality features help in improving the overall model performance. Human insight
Feedback: This answer is correct.

Question 4: Which of the following operations can be performed on input variables?

A: Arithmetic operations
Feedback: This answer is partially correct, please review the module again.

B: Computing trigonometric functions
Feedback: This answer is partially correct, please review the module again.

C: Computing algebraic functions
Feedback: This answer is partially correct, please review the module again.

*D: All of the above
Feedback: This answer is correct.

Representing Features

Question 1: Select ALL true statements regarding the ML.EVALUATE function?

A: The ML.EVALUATE function can be used with linear regression, logistic regression, k-means, matrix factorization, and ARIMA-based time series models..
Feedback: This answer is partially correct, please review the module again.

B: The ML.EVALUATE function evaluates the predicted values against the actual data.
Feedback: This answer is partially correct, please review the module again.

C: You can use the ML.EVALUATE function to evaluate model metrics.
Feedback: This answer is partially correct, please review the module again.

*D: All of the above.
Feedback: This answer is correct.

Question 2: What is the significance of ML.FEATURE_CROSS?

*A: ML.FEATURE_CROSS generates a STRUCT feature with all combinations of crossed categorical features except for 1-degree items.
Feedback: This answer is correct.

B: ML.FEATURE_CROSS generates a STRUCT feature with few combinations of crossed categorical features except for 1-degree items.
Feedback: This answer is incorrect, please review the module again.

C: ML.FEATURE_CROSS generates a STRUCT feature with all combinations of crossed categorical features including 1-degree items.
Feedback: This answer is incorrect, please review the module again.

D: ML.FEATURE_CROSS generates a STRUCT feature with few combinations of crossed categorical features except for 1-degree items.
Feedback: This answer is incorrect, please review the module again.

Question 3: Select ALL true statements regarding the ML.BUCKETIZE function?

A: ML.BUCKETIZE is a pre-processing function that creates buckets by returning a STRING as the bucket name after numerical_expression is split into buckets by array_split_points..
Feedback: This answer is partially correct, please review the module again.

B: It bucketizes a continuous numerical feature into a string feature with bucket names as the value.
Feedback: This answer is partially correct, please review the module again.

*C: Both A and B
Feedback: This answer is correct.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 4: Which of the following is true about Feature Cross?

A: It is a process of combining features into a single feature.
Feedback: This answer is partially correct, please review the module again.

B: Feature Cross enables a model to learn separate weights for each combination of features.
Feedback: This answer is partially correct, please review the module again.

*C: Both A and B
Feedback: This answer is correct.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 5: Which of the following statements is incorrect?

A: When we do feature crosses, we run into the risk of overfitting
Feedback: This answer is incorrect, please review the module again.

B: We use the Regularization process in order to prevent overfitting.
Feedback: This answer is incorrect, please review the module again.

C: BQML by default assumes that numbers are numeric features and strings are categorical features.
Feedback: This answer is incorrect, please review the module again.

*D: None of the above
Feedback: This answer is correct.

Module 3: Preprocessing and feature creation

Preprocessing and Feature Creation

Question 1: You are training a model to predict how long it will take to sell a house. The list price of the house, with numeric $20,000 to $500,000 values, is one of the inputs to the model. Which of these is a good practice?

*A: Rescale the real valued feature like a price to a range from 0 to 1
Feedback: This answer is correct.

B: Rescale the real valued feature like a price to a range from 0 to $100,000
Feedback: This is incorrect, please review the module again.

C: Rescale the real valued feature like a price to a categorical range from low, medium, high
Feedback: This is incorrect, please review the module again.

Question 2: Which of these tools are commonly used for data pre-processing? (Select 3 correct responses)

* BigQuery
Feedback: Correct, this is one tool used for data pre-processing.

*TensorFlow
Feedback: Correct, this is one tool used for data pre-processing.

*C: Apache Beam

Feedback: Correct, this is one tool used for data pre-processing.

Google Cloud Storage
Feedback: This answer is incorrect, please review the module again.

Bigtable
Feedback: This answer is incorrect, please review the module again.

Question 3: Which one of these is NOT something you would commonly do in data preprocessing?

*Tune your ML model hyperparameters
Feedback: Correct - this will come later

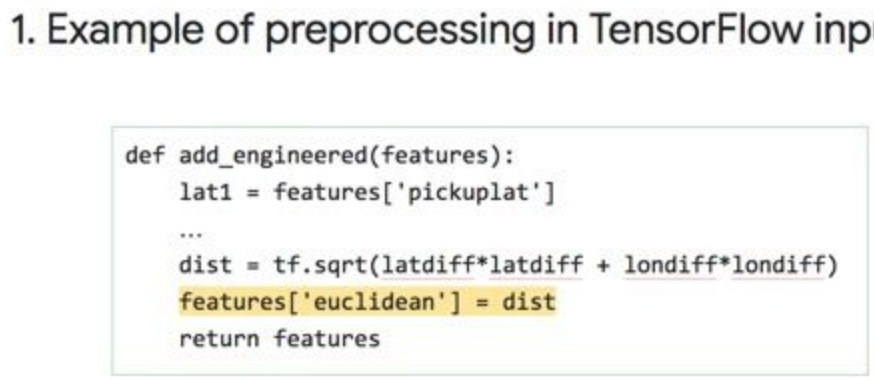
Remove examples that you don't want to train on
Feedback: This answer is incorrect, as this is a common task in data preprocessing. Please review the module again.

Compute vocabularies for categorical columns
Feedback: This answer is incorrect, as this is a common task in data preprocessing. Please review the module again.

Compute aggregate statistics for numeric columns
Feedback: This answer is incorrect, as this is a common task in data preprocessing. Please review the module again.

Compute time-windowed statistics (e.g. number of products sold in previous hour) for use as input features.
Feedback: This answer is incorrect, as this is a common task in data preprocessing. Please review the module again.

Questions 4: In your TensorFlow model you are calculating the distance between two points on a map as a new feature. How do you ensure the preprocessing you're doing for model training is also done the exact same way in prediction?



A: Wrap features in training/evaluation input function:

```
def input_fn():
  features = ...
  label = ...
  return add_engineered(features), label
```

Feedback: This is incorrect, please review the module again.

B:Wrap features in serving input function:

```
def serving_input_fn():
    feature_placeholders = ...
    features = ...
    return tf.estimator.export.ServingInputReceiver(
      add_engineered(features), feature_placeholders)
```

Feedback: This is incorrect, please review the module again.

*C: Wrap features in training/evaluation input function AND wrap features in serving input function:

```
def input_fn():
  features = ...
  label = ...
  return add_engineered(features), label
```

```
def serving_input_fn():
    feature_placeholders = ...
    features = ...
    return tf.estimator.export.ServingInputReceiver(
      add_engineered(features), feature_placeholders)
```

Feedback: This answer is correct.

Question 5: The below code preprocesses the latitude and longitude using feature columns. What is the point of the 38.0 and 42.0 in the column buckets?

```
def build_estimator(model_dir, nbuckets):
  latbuckets = np.linspace(38.0, 42.0, nbuckets).tolist()
  b_plat = tf.feature_column.bucketized_column(plat, latbuckets)
  b_dlat = tf.feature_column.bucketized_column(dlat, latbuckets)

  return tf.estimator.LinearRegressor(
      model_dir=model_dir,
      feature_columns=[..., b_plat, b_dlat, ...])
```

These define how many samples to put into each bucket (at least 38 but no more than 42 in each small bucket)
Feedback: This is incorrect, please review the module again.

*Latitudes must be between 38 and 42 will be discretized into the specified number of bins.
Feedback: This answer is correct.

These parameters ensure all latitudes in the raw dataset do not include 38 and 42 which we want to exclude for this dataset.
Feedback: This is incorrect, please review the module again.

Question 6: What are two advantages of using TensorFlow to preprocess your code instead of building an Apache Beam pipeline? (Select two correct responses)

*In TensorFlow you will have access to helper APIs to help automatically bucketize and process features instead of writing your own java or python code.
Feedback: Correct, this is one of the advantages, please review the module again.

In TensorFlow you will have access to helper APIs to help automatically bucketize and process features instead of writing your own java or python code.
Feedback: Not correct, this is not one of the advantages.

*In TensorFlow the same pipelines can be used in both training and serving.
Feedback: Correct, this is one of the advantages, please review the module again.

Question 7: What is one key advantage of preprocessing your features using Apache Beam?

*A: The same code you use to preprocess features in training and evaluation can also be used in serving.
Feedback: This answer is correct.

B: Apache Beam transformations are written in Standard SQL which is scalable and easy to author.
Feedback: This is incorrect, please review the module again.

C: Apache Beam code is often harder to maintain and run at scale than BigQuery preprocessing pipelines.
Feedback: This is incorrect, please review the module again.

Apache Beam and Cloud Dataflow

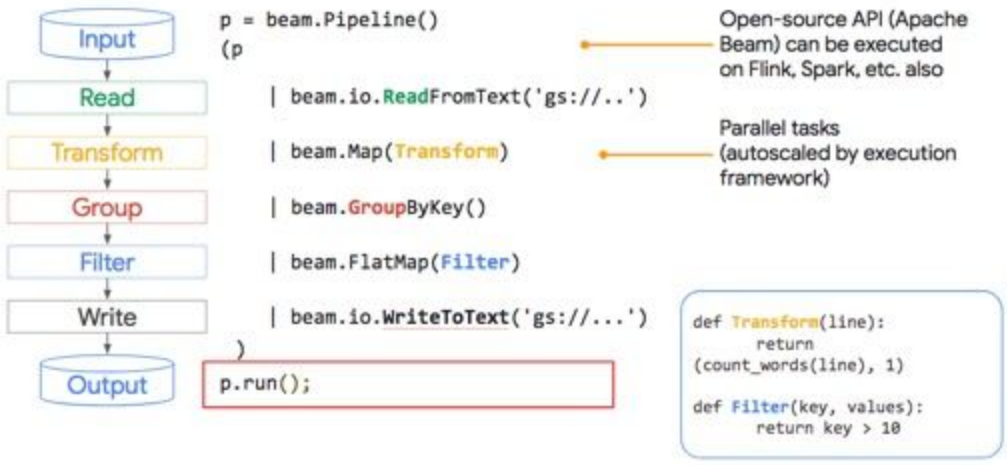
Question 1: Which of these accurately describes the relationship between Apache Beam and Cloud Dataflow?

*Cloud Dataflow is the API for data pipeline building in java or python and Apache Beam is the implementation and execution framework.
Feedback: This answer is correct.

They are the same.
Feedback: This is incorrect, please review the module again.

Cloud Dataflow is the proprietary version of the Apache Beam API and the two are not compatible.
Feedback: This is incorrect, please review the module again.

Question 2: TRUE or FALSE: The Filter method can be carried out in parallel and autoscaled by the execution framework:



A: True: Anything in Map or FlatMap can be parallelized by the Beam execution framework.
Feedback: This answer is correct.

B: False: Anything in Map or FlatMap can be parallelized by the Beam execution framework.
Feedback: This is incorrect, please review the module again.

Question 3: What is the purpose of a Cloud Dataflow connector?

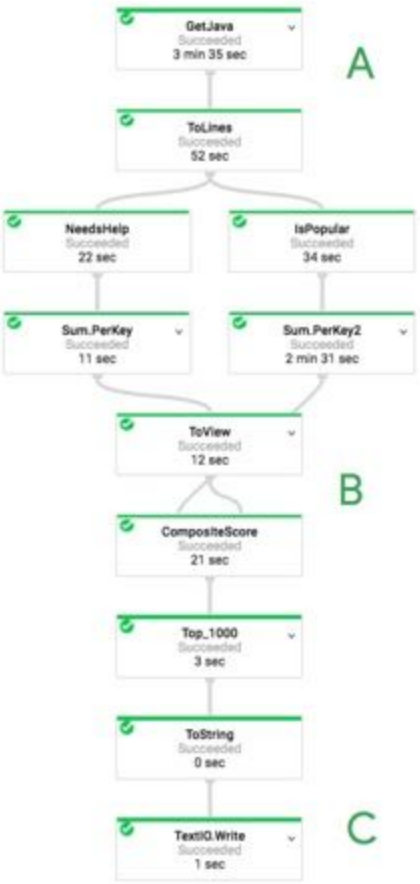
```
.apply(TextIO.write().to("gs://…"));
```

*Connectors allow you to output the results of a pipeline to a specific data sink like Bigtable, Google Cloud Storage, flat file, BigQuery, and more...
Feedback: This answer is correct.

Connectors allow you to chain multiple data-processing steps together automatically so they process in parallel.
Feedback: This is incorrect, please review the module again.

Connectors allow you to authenticate your pipeline as specific users who may have greater access to datasets.
Feedback: This is incorrect, please review the module again.

Question 4: Below you'll find a Cloud Dataflow preprocessing graph. Correctly identify the terms for A, B, and C.



*A is a data source, B are transformation steps, and C is a data sink.
Feedback: This answer is correct.

A is a data stream,
B are transformation steps, and
C is a data sink
Feedback: This is incorrect, please review the module again.

A is a data stream,
B are transformation steps, and
C is a data source
Feedback: This is incorrect, please review the module again.

Question 5: To run a pipeline you need something called a _____.

*runner
Feedback: This answer is correct.

executor
Feedback: This is incorrect, please review the module again.

pipeline
Feedback: This is incorrect, please review the module again.

Apache Beam
Feedback: This is incorrect, please review the module again.

Question 6: Your development team is about to execute this code block. What is your team about to do?

```
mvn compile -e exec:java \
      -Dexec.mainClass=$MAIN \
      -Dexec.args="--project=$PROJECT \
      --stagingLocation=gs://$BUCKET/staging/ \
      --tempLocation=gs://$BUCKET/staging/ \
      --runner=DataflowRunner"
```

*We are compiling our Cloud Dataflow pipeline written in Java and are submitting it to the cloud for execution.
Notice that we are calling mvn compile and passing in --runner=DataflowRunner.
Feedback: This answer is correct.

We are compiling our Cloud Dataflow pipeline written in Python and are loading the outputs of the executed pipeline inside of Google Cloud Storage (gs://)
Feedback: This is incorrect, please review the module again.

We are preparing a staging area in Google Cloud Storage for the output of our Cloud Dataflow pipeline and will be submitting our BigQuery job with a later command.
Feedback: This is incorrect, please review the module again.

Question 7: TRUE or FALSE: A ParDo acts on all items at once (like a Map in MapReduce).

A: True
Feedback: This is incorrect, please review the module again.

*B: False. A ParDo acts on one item at a time (like a Map in MapReduce)
Feedback: This answer is correct.

Preprocessing with Cloud Dataprep

Question 1: What are some of the advantages to exploring datasets with a UI tool like Cloud Dataprep?

*A: Dataprep uses Dataflow behind-the-scenes and you can create your transformations in a UI tool instead of writing Java or Python.
Feedback: This is one of the correct answers.

B: Dataprep allows for quick visual inspection of your dataset with accurate frequency histograms for **every** value in a column.
Feedback: This answer is not correct. Dataprep only loads a sample of your dataset into the viewer for inspection. While you can control what type of sample to load, you will need to run the actual pipeline to view and analyze all of your data.

*C: Dataprep has a number of transformation steps available that you can chain together as part of a recipe.
Feedback: This is one of the correct answers.

*D: Dataprep supports outputting your data into BigQuery, Google Cloud Storage, or flat files.
Feedback: This is one of the correct answers.

Question 2: TRUE or FALSE: You can automatically setup pipelines to run at defined intervals with Cloud Dataprep.

A: True
Feedback: This is one of the correct answers. Follow this guide here.

B: False
Feedback: This answer is not correct. Actually you can!

https://cloud.google.com/blog/big-data/2017/11/scheduling-and-sampling-arrive-for-google-cloud-dataprep

Module 4: Feature Crosses

Feature Crosses

Question 1: You are building a model to predict the number of points ("margin") by which Team A will beat Team B in a basketball game. Your input features are (1) whether or not it is a home game for Team A (2) average number of points Team A scored in its past 7 games and (3) average number of points Team B scored in its past 7 games. Which of these is a linear model suitable for machine learning?

*margin = b + w1 * is_home_game + w2 * avg_points_A + w3 * avg_points_B
Feedback: Correct. The output is a weighted sum of the input features

margin = (avg_points_A - avg_points_B)
Feedback: This answer is not correct.

*margin = w1 * is_home + w2 * (avg_points_A - avg_points_B)^3
Feedback: This is also a linear model. The first feature is a raw input, but the second feature is an engineered feature, created by combining two of the inputs (this is a feature cross).

margin = w1 * is_home + w1^2 * avg_points_A + w1^3 * avg_points_B
Feedback: This is not a linear model. Notice that we have w1, w1^2 and w1^3 -- i.e. powers of the weights. So, this is a non-linear ("polynomial") model.

Question 2: Feature crosses are more common in modern machine learning because:

* Feature crosses memorize, and that is okay only if you have extremely large datasets.
Feedback: This answer is correct.

People didn't know about feature crosses 10 years ago.
Feedback: This is incorrect, please review the module again.

Feature crosses work only with neural networks.
Feedback: This is incorrect, please review the module again.

Feature crosses require GPUs in order to compute efficiently.
Feedback: Not quite. In fact, GPUs don't help for sparse inputs like feature crosses. a model with a lot of feature crosses might be more effectively trained on a distributed CPU cluster.

Question 3: The function tf.feature_column.crossed_column requires:

*A list of categorical or bucketized features
Feedback: This answer is correct.

A list of categorical features
Feedback: This is incorrect, please review the module again.

A list of categorical or bucketized features
Feedback: This is incorrect, please review the module again.

A list of numeric features
Feedback: This is incorrect, please review the module again.

Question 4: You might create an embedding of a feature cross in order to:

*Create a lower-dimensional representation of the input space
Feedback: This is one of the correct answers.

*Identify similar sets of inputs for clustering
Feedback: This is one of the correct answers.

*Reuse weights learned in one problem in another problem
Feedback: This is one of the correct answers.

Module Quiz

Question 1: What is a feature cross?

A: A feature cross is a synthetic feature formed by adding (crossing) two or more features. Crossing combinations of features can provide predictive abilities beyond what those features can provide individually.
Feedback: This answer is incorrect, please review the module again.

*B: A feature cross is a synthetic feature formed by multiplying (crossing) two or more features. Crossing combinations of features can provide predictive abilities beyond what those features can provide individually.
Feedback: This answer is correct.

C: A feature cross is a synthetic feature formed by dividing (crossing) two or more features. Crossing combinations of features can provide predictive abilities beyond what those features can provide individually.
Feedback: This answer is incorrect, please review the module again.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 2: True or False: We can create many different kinds of feature crosses. For example:

- **[A X B]: a feature cross formed by multiplying the values of two features.**
- **[A x B x C x D x E]: a feature cross formed by multiplying the values of five features.**
- **[A x A]: a feature cross formed by squaring a single feature.**

*A: True
Feedback: This answer is correct.

B: False
Feedback: This answer is incorrect, please review the module again.

Question 3: True or False: Feature Engineering is often one of the most valuable tasks a data scientist can do to improve model performance, for three main reasons:

1. **You can isolate and highlight key information, which helps your algorithms "focus" on what's important.**
2. **You can bring in your own domain expertise.**
3. **Once you understand the "vocabulary" of feature engineering, you can bring in other people's domain expertise.**

*A: True
Feedback: This answer is correct.

B: False
Feedback: This answer is incorrect, please review the module again.

Module 5: TensorFlow Transform

tf.transform

Question 1: During the training and serving phase, tf.Transform:

*A: Provides a TensorFlow graph for preprocessing
Feedback: This answer is correct.

B: Provides computation over the entire dataset, including on both internal and external data sources.
Feedback: This answer is incorrect, please review the module again.

C: Provides a transformation polynomial to train the data.
Feedback: This answer is incorrect, please review the module again.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 2: What is Tensorflow transform is a hybrid of?

*A: Apache and TensorFlow
Feedback: This answer is correct.

B: Dataflow and Tensorflow
Feedback: This answer is incorrect, please review the module again.

C: Both a & b
Feedback: This answer is incorrect, please review the module again.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 3: True or False: One of the goals of tf.Transform is to provide a TensorFlow graph for preprocessing that can be incorporated into the serving graph (and, optionally, the training graph).

*A: True
Feedback: This answer is correct.

B: False
Feedback: This answer is incorrect, please review the module again.

Question 3: Fill in the blank:
The _____ _____ is the most important concept of tf.Transform. The _____ _____ is a logical description of a transformation of the dataset. The _____ _____ accepts and returns a dictionary of tensors, where a tensor means Tensor or 2D SparseTensor.

*A: Preprocessing function
Feedback: This answer is correct.

B: Preprocessing variable
Feedback: This answer is incorrect, please review the module again.

C: Preprocessing method
Feedback: This answer is incorrect, please review the module again.

Module 6: Summary

Course Quizz

Question 1: Which of the following process steps are considered a best practice in predictive modeling?

A: Feature engineering > Data Cleaning > Model Building
Feedback: This answer is incorrect, please review the module again.

*B: Data Cleaning > Feature engineering > Model Building
Feedback: This answer is correct.

C: Model building > Feature engineering > Data cleaning
Feedback: This answer is incorrect, please review the module again.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 2: Feature engineering can include:

A: Using indicator variables to isolate key information.
Feedback: This answer is partially correct, please review the module again.

B: Highlighting interactions between two or more features.
Feedback: This answer is partially correct, please review the module again.

C: Representing the same feature in a different way.
Feedback: This answer is partially correct, please review the module again.

*D: All of the above
Feedback: This answer is correct.

Question 3: A good feature typically ....

A: Is related to the objective
Feedback: This answer is partially correct, please review the module again.

B: Is known at prediction time
Feedback: This answer is partially correct, please review the module again.

*C: Both a & b
Feedback: This answer is correct.

D: None of the above
Feedback: This answer is incorrect, please review the module again.

Question 4: An example of preprocessing a date feature is ...

A: Extracting the parts of the date into different columns: Year, month, day, etc.
Feedback: This answer is partially correct, please review the module again.

B: Extracting the time period between the current date and columns in terms of years, months, days, etc.
Feedback: This answer is partially correct, please review the module again.

C: Extracting some specific features from the date: Name of the weekday, weekend or not, holiday or not, etc.
Feedback: This answer is partially correct, please review the module again.

*D: All of the above
Feedback: This answer is incorrect, please review the module again.