Hong Hanh

Practical Data Science Specialization

Week 1 quiz

Practice Quiz • 20 min

1. You are training an NLP model to classify product reviews on a very large dataset, but training is taking a long time. How can you use feature engineering to reduce training time and possibly increase model performance? (choose all that apply) — 1 / 1 point

- [ ] Randomly delete instances in the dataset to reduce the dataset size.
- [x] Filter irrelevant and redundant attributes, then retrain the model.

> ✓ **Correct**
> Correct! This is a feature selection approach that ensures that you are only keeping relevant attributes which is likely to reduce training time and possibly increase model accuracy..

- [x] Normalize or Standardize your data before training

> ✓ **Correct**
> Correct! Normalizing or standardizing data is the process of converting data to a common format/scale and can therefore reduce training time and sometimes increase model accuracy..

- [ ] None of the above

2. You perform correlation analysis on your feature set and discover that some features are highly correlated to each other. How can you take advantage of this information to improve your model's performance? (choose all that apply) — 1 / 1 point

- [x] Combine the correlated features together.

> ✓ **Correct**
> Correct! Combining highly correlated features together can increase train speed and model performance as this will prevent the duplication of information.

- [ ] Ignore this information as highly correlated features always improve model performance.
- [ ] Apply different methods to increase the number of highly correlated features.
- [x] Eliminate one of the correlated features.

3. You have a task to train a text classifier on a customer product reviews dataset.. You decide to use the "star rating" to create 3 sentiments. — 1 / 1 point

Rating 1 & 2 = Negative

Rating 3 = Neural

Rating 4 & 5 = Positive

What feature engineering method did you use in this scenario?

- ( ) Feature Selection
- (•) Feature Creation
- ( ) Feature Transformation
- ( ) All of the above

> ✓ **Correct**
> That's right! In feature creation we can combine existing features into new features or even create new attributes from existing ones. This is exactly what was done here as new sentiment features have been created from the "star rating" attribute.

Hong Hanh

Practical Data Science Specialization

4. BERT, which stands for Bidirectional Encoder Representations from Transformers, and Amazon SageMaker BlazingText are 2 popular natural language processing (NLP) algorithms. What are some characteristics of these algorithms? **1 / 1 point**

- [ ] They both take into account the word position when generating the embedding.
- [x] Both models can generate embeddings which can be used as input to a neural network.

✓ **Correct**
Correct! BlazingText and BERT both generate embeddings of the text, which are vector representations of the text.

- [ ] Unlike BERT, BlazingText cannot generate vectors for words encountered outside its vocabulary space i.e it does not support out-of-vocabulary words.
- [x] BERT is based on the transformer architecture while BlazingText is based on the Word2Vec architecture.

✓ **Correct**
That's right! BERT uses a bidirectional transformer architecture which utilizes a self-attention mechanism to capture the semantic relationships in the input data. BlazingText on the other hand uses Word2Vec which uses a shallow neural network to group similar words together in a vector space.

5. Consider the following two sentences as input to either BERT or Blazing Text: **1 / 1 point**

Sentence 1: You have to take a **right** turn to reach that supermarket

Sentence 2: Her decisions are often **right**.

Which of the following statements are true about these examples?

- [ ] BERT will use the same vector representation for the word "right" for both sentences.
- [ ] BlazingText will use different vector representations for the word "right" for both sentences.
- [x] BlazingText will use the same vector representation for the word "right" for both sentences.

✓ **Correct**
Correct! BlazingText does not account for context when generating embeddings. So, no matter the meaning of the word in the sentence, BlazingText will use the same vector representation.

- [x] BERT will use different vector representations for the word "right" for both sentences.

✓ **Correct**
Correct! BERT uses a bidirectional transformer architecture and therefore generates contextual embeddings. If the same word is used in different ways, BERT will capture different meanings and therefore produce 2 different vectors to represent the different meanings.

Hong Hanh

Practical Data Science Specialization

6. Amazon SageMaker Processing lets you perform preprocessing, post processing, and model evaluation on your data. 1 / 1 point

Is it true that SageMaker Processing Jobs:

- accept data as input from S3
- can work on any large scale data
- use a built-in container for scikit-learn.

- ○ No
- ◉ Yes

✓ **Correct**

Amazon SageMaker Processing Jobs allow you to perform feature engineering at scale. These jobs retrieve your data from S3 and perform preprocessing using popular frameworks such as scikit-learn, Apache Spark, TensorFlow, and PyTorch. You can also use a custom Docker container for your Processing Job.

7. You can think of a feature store as a central store for curated features. What are the key advantages of using a feature store such as Amazon SageMaker Feature Store? 1 / 1 point

- ☑ It provides consistency between features used for training and inference and helps to reduce skew.

✓ **Correct**

Correct! Training and serving skew could happen because of discrepancies between the data used for training and the data used for inference. Amazon SageMaker Feature Store helps reduce training and serving skew.

- ☐ It provides access to various feature engineering algorithms.
- ☐ It contains original datasets from which features can be extracted.
- ☑ It can provide online and/or offline access to curated feature data.

✓ **Correct**

That's right! A feature store holds the results of the feature engineering pipeline so that you don't have to re-execute the feature engineering pipeline again and again. Amazon SageMaker Feature Store provides access to features for both online real-time inference and offline batch inference.