Introduction to Clustering in Apache Beam

Jasper Van den Bossche
ML6
Agenda

- What is clustering?
  - Online vs offline clustering
  - What are the applications?
- How does clustering in Apache Beam work
  - High level overview of the transform
- Example pipeline
What is clustering?
What is clustering?

Clustering is an *unsupervised* technique used to *group similar data points* together based on their *characteristics* or *patterns*.
What is Unsupervised Training?

Supervised

Unsupervised
How are datapoints grouped together?
How are datapoints grouped together?

Distance Matrix

Spectral Clustering

DBSCAN
K-means clustering

Step 1

Step 2

Step 3
Online vs offline clustering
Offline Clustering
Online Clustering
What are the applications of clustering?
What are the applications of clustering?
Anomaly detection

- Detect fraudulent transactions
- Detect diseases
- Quality control
- Spam filters
Personalisation

- Personalised ads
- Movie/music recommendations
Grouping documents

1. Use a language model to calculate embeddings
2. Group together points in the embedding space close to each other
How does clustering in Apache Beam work
A High Level Look Behind the Scenes

Step 1
- Preprocess Data
- Calculate Clusters
- Save Model

Step 2
- Load Trained Model
- Calculate Predictions
Preprocessing

1. Create Batches of Datapoints
2. Convert to Numpy and Reshape
Calculate Cluster Centers

Process batch by batch to calculate cluster centers
Clustering is a stateful transform
Clustering is a stateful transform
Save the trained model to persistent storage

Save Model

model.checkpoint
Assign Labels

Assign all datapoints a label using the trained model.
Let’s look at an example!
Example: Clustering California Houses

Group similar houses based on location and income of the owner

<table>
<thead>
<tr>
<th>longitude</th>
<th>latitude</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>-122.23</td>
<td>37.83</td>
<td>52.000</td>
</tr>
<tr>
<td>-122.28</td>
<td>37.81</td>
<td>152.000</td>
</tr>
<tr>
<td>-122.17</td>
<td>37.82</td>
<td>48.000</td>
</tr>
<tr>
<td>-122.26</td>
<td>37.79</td>
<td>56.000</td>
</tr>
<tr>
<td>-122.23</td>
<td>37.84</td>
<td>72.000</td>
</tr>
</tbody>
</table>
# 1. Calculate clustering centers and save model to persistent storage
model = ( 
    housing_features
    | "Train clustering model" >> OnlineClustering(
        OnlineKMeans,
        n_clusters=6,
        batch_size=256,
        cluster_args={},
        checkpoints_path=known_args.checkpoints_path))
# 2. Calculate labels for all records in the dataset using the trained clustering model using in memory model

```python
_ = (housing_features
     | "RunInference" >> AssignClusterLabelsInMemoryModel(
         model=pvalue.AsSingleton(model),
         model_id="kmeans",
         n_clusters=6,
         batch_size=512)
     | beam.Map(print)))
```
Calculating Predictions

```python
pipeline = test_pipeline
if not test_pipeline:
    pipeline = beam.Pipeline(options=pipeline_options)

data = pipeline | read_csv(known_args.input)

features = ['longitude', 'latitude', 'median_income']

housing_features = to_pcollection(data[features])
```
Summary

- Clustering is a technique to group similar datapoints based on their characteristics.
- Many applications ranging from anomaly detection to document grouping.
- Clustering is a twofold transform in Apache Beam:
  - Data preprocessing and model training.
  - Assigning cluster labels to datapoints.
QUESTIONS?