

Algorithm Cluster

ALGORITHM CLUSTER



Algorithm Cluster

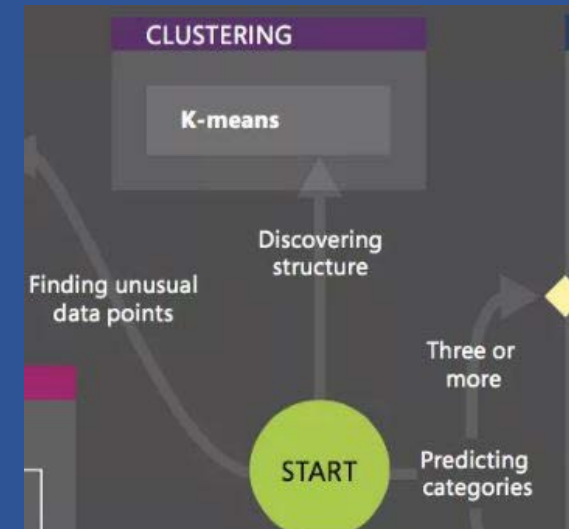
In this session

- Cluster Algorithms in Azure ML
- Model overview
- Dataset
- Feature Hashing module
- Train
- Edit Metadata

Algorithm Cluster

Cluster Algorithms in Azure ML

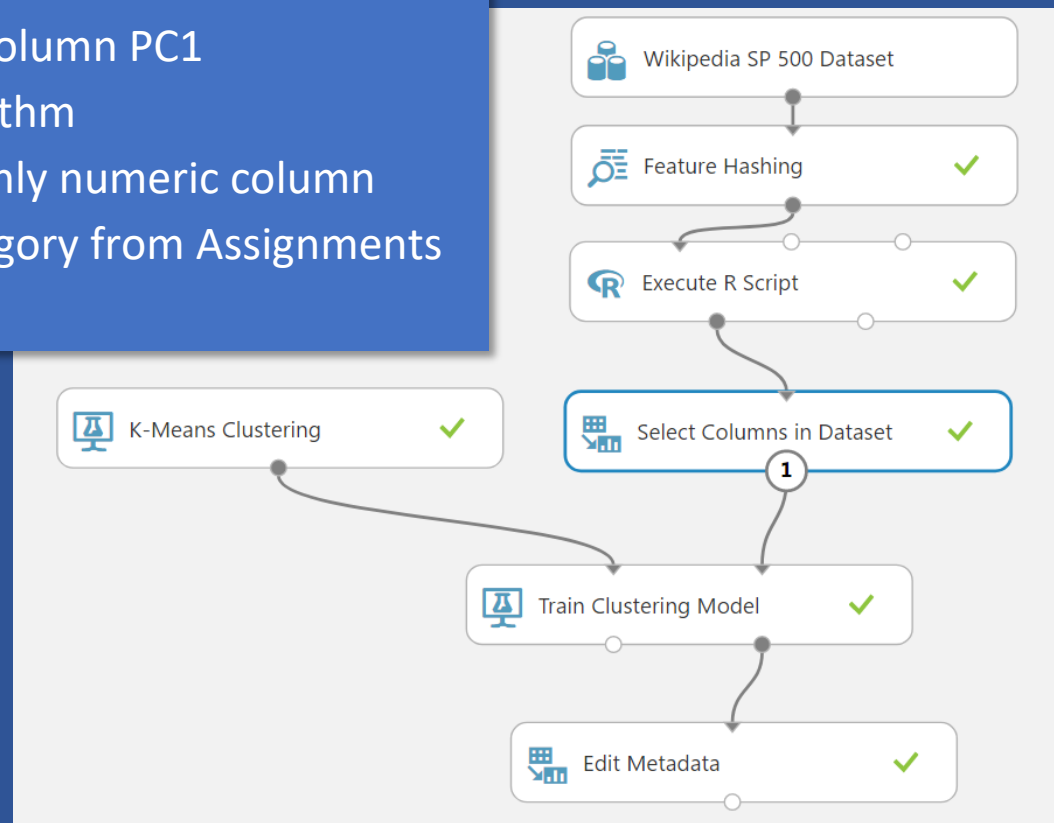
- Uses iterative techniques
- Group cases in a dataset into clusters
- Contain similar characteristics
- Useful for exploring data
- Identifying anomalies in the data
- Making predictions
- Identify relationships in a dataset
- Not logically derive by browsing or simple observation
- Used in the early phases of machine learning tasks
- Explore the data and discover unexpected correlations
- Only algorithm in AML that is **Unsupervised**



Algorithm Cluster




Model overview

- **Dataset:** Wikipedia SP 500 Dataset
- **Feature Hashing:** create feature from column Text
- **R Script:** reduce feature to 10 columns
- **Select Columns:** exclude column PC1
- **K-Means:** clustering algorithm
- **Train:** train model using only numeric column
- **Edit Metadata:** make category from Assignments column



Algorithm Cluster

Dataset

rows	columns			
466	3			
	Title	Category	Text	
view as				
	Apple Inc.	Information Technology	nasdaq 100 component s p 500 component foundation founder location city apple campus 1 infinite loop street infinite loop cupertino california cupertino california location country united	

Pre-processed outside Azure ML Studio

- Removing wiki formatting
- Removing non-alphanumeric characters
- Converting all text to lowercase
- Adding company categories, where known

Algorithm Cluster

Feature Hashing module

Feature Hashing module


- Tokenizes the text string
- Transforms the data into a series of numbers
- Based on the hash value of each token

Feature Hashing


Target column(s)

Selected columns:
Column names: Text








Launch column selector

Hashing bitsize 

12

N-grams 

1

rows	columns				
466	4099				
	Title	Category	Text	Text_HashingFeature_1	Text_HashingFeature_2
view as  			 nasdaq 100 component s p 500 component component		

Algorithm Cluster

R Script

R Script

```
1 dataset1 <- mam1.mapInputPort(1)
2 titles_categories = dataset1[,1:2]
3 pca = prcomp(dataset1[,4:4099])
4 top_pca_scores = data.frame(pca$x[,1:10])
5 data.set = cbind(titles_categories,top_pca_scores)
6 mam1.mapOutputPort("data.set");
```

Execute R Script

R Script

```
1 dataset1 <- mam1.mapInputPo
2 titles_categories = dataset
3 pca = prcomp(dataset1[,4:40
4 top_pca_scores = data.frame
5 data.set = cbind(titles_cat
6 mam1.mapOutputPort("data.se
```

Random Seed

R Version

- Dimensionality of the data from hashing is too high (4K)
- Cannot be used by the K-Means clustering algorithm directly
- Principal Component Analysis (PCA) was applied using a custom R script
- Reduce the dimensionality to 10 variables
- View the result = double-clicking the right-hand output of the Execute R Script

Algorithm Cluster

- Select Columns in Dataset
- K-Means Clustering

<https://msdn.microsoft.com/en-us/library/azure/dn905944.aspx>

▲ Select Columns in Dataset

Select columns

Selected columns:
All columns
Exclude column names: PC1

Launch column selector

▲ K-Means Clustering

Create trainer mode

Single Parameter ▼

Number of Centroids

3

Initialization

K-Means++ Fast ▼

Random number seed

7654

Metric

Cosine ▼

Iterations

100

Assign Label Mode

Ignore label column ▼

Algorithm Cluster


- Train Clustering Model
- Edit Metadata

▲ Train Clustering Model

Column Set

Selected columns:
Column type: Numeric, All

Launch column selector

☒ Check for Append or Uncheck f... 

▲ Edit Metadata


Column

Selected columns:
Column names: Assignments


Launch column selector

Data type


Unchanged ▼

Categorical 

Make categorical ▼

Fields 

Unchanged ▼

New column names 

Algorithm Cluster

More information

K-Means Clustering

<https://msdn.microsoft.com/en-us/library/azure/dn905944.aspx>

This Experiment

<https://gallery.cortanaintelligence.com/Experiment/Clustering-K-Means-basic>