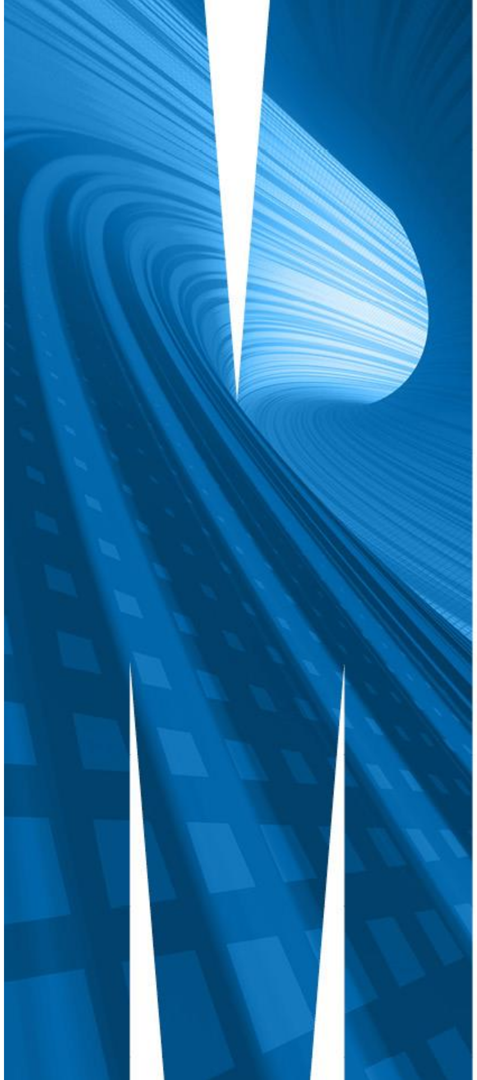
MONASH University

Session 7

FIT5202 Big Data Processing

K-Means Clustering

Model Selection



Week 7 Agenda

- **Part - A**
- Session 5 Review
- K-means Clustering
  - Shilouette Score
- Tutorial Instructions
  - Use case : Identify if 3 hackers were involved

- **Part - B**
- Model Selection
  - Hyperparameter Tuning
  - Cross Validation
    - K-fold Cross Validation
  - TrainValidationSplit
- Model Persistence
  - Saving and Loading a Model

K-Means Clustering

Finds groups (or clusters) of data

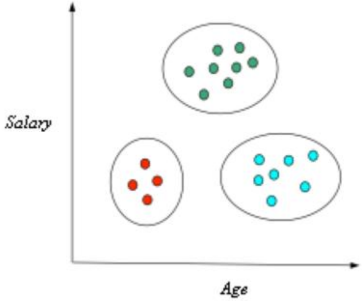
A cluster comprises a number of “similar” objects

A member is closer to another member within the same group than to a member of a different group

Groups have no category or label

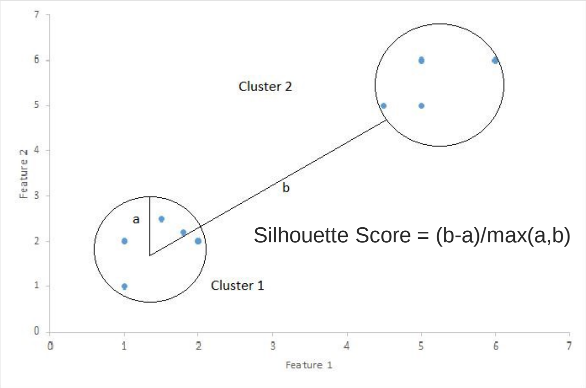
Unsupervised learning

Animation Demo (https://www.naftaliharris.com/blog/visualizing-k-means-clustering)



Silhouette Score (-1 to 1): calculates the goodness of a clustering technique

- **1** means Clusters are well apart from each other and clearly distinguishes
- **0** means clusters are not clearly distinguished, the distance between the clusters is not significant
- **-1** means clusters are assigned in the wrong way

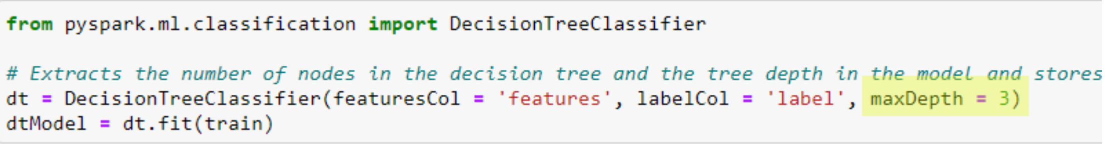


https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

Model Selection

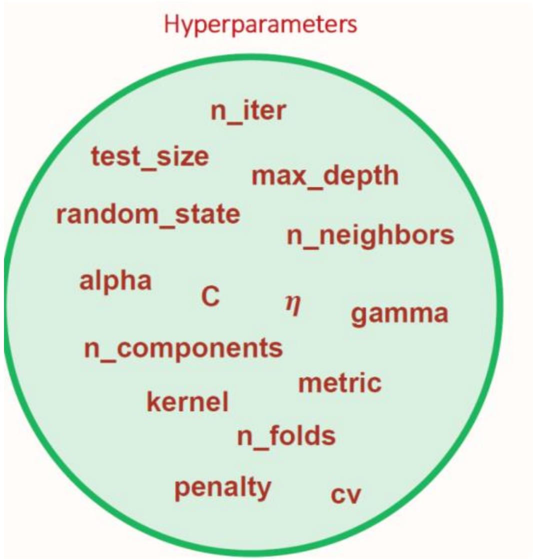
Hyperparameter Tuning

- Hyper-parameters are not model parameters : they cannot be trained from the data
- Hyperparameter tuning : choosing a set of optimal hyperparameters for a learning algorithm
- model.getParamMap() to get the list of hyperparameters for the model



```
from pyspark.ml.classification import DecisionTreeClassifier

# Extracts the number of nodes in the decision tree and the tree depth in the model and stores
dt = DecisionTreeClassifier(featuresCol = 'features', labelCol = 'label', maxDepth = 3)
dtModel = dt.fit(train)
```

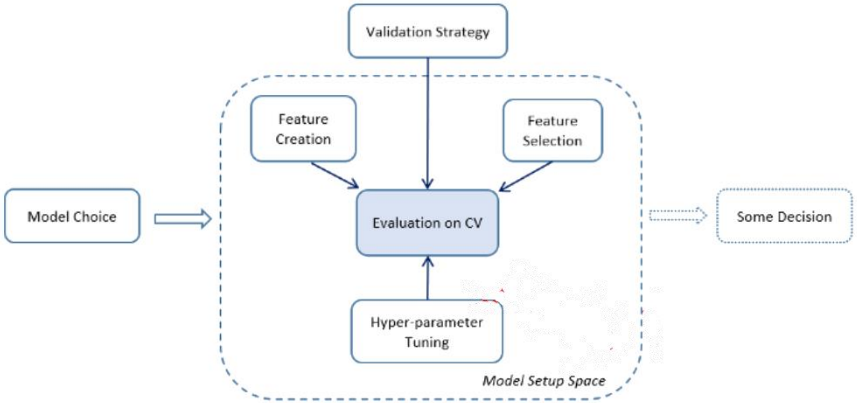


https://towardsdatascience.com/model-parameters-and-hyperparameters-in-machine-learning-what-is-the-difference-702d30970f6

Model Selection

All models are wrong; some are useful (George E.P. Box)

- Finding the best model or parameters
- Tuning can be done for individual Estimators or the entire Pipeline



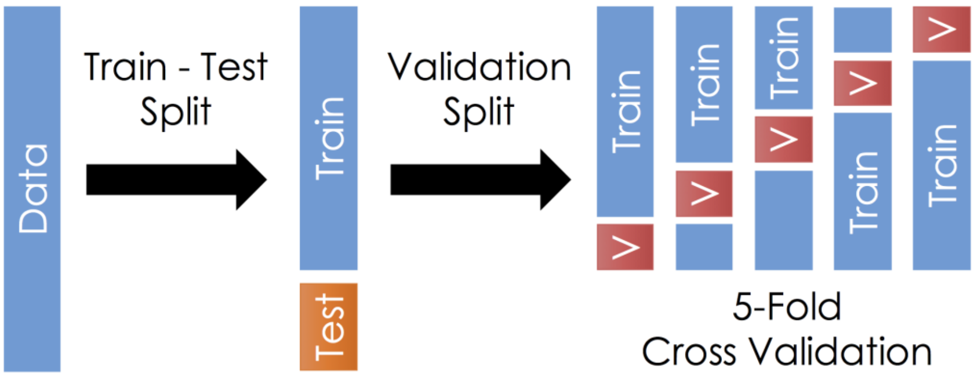
Model selection for Mlib has the following tools:

1. CrossValidator
2. TrainValidationSplit

https://spark.apache.org/docs/latest/ml-tuning.html
https://towardsdatascience.com/hyperparameter-tuning-explained-d0ebb2ba1d35

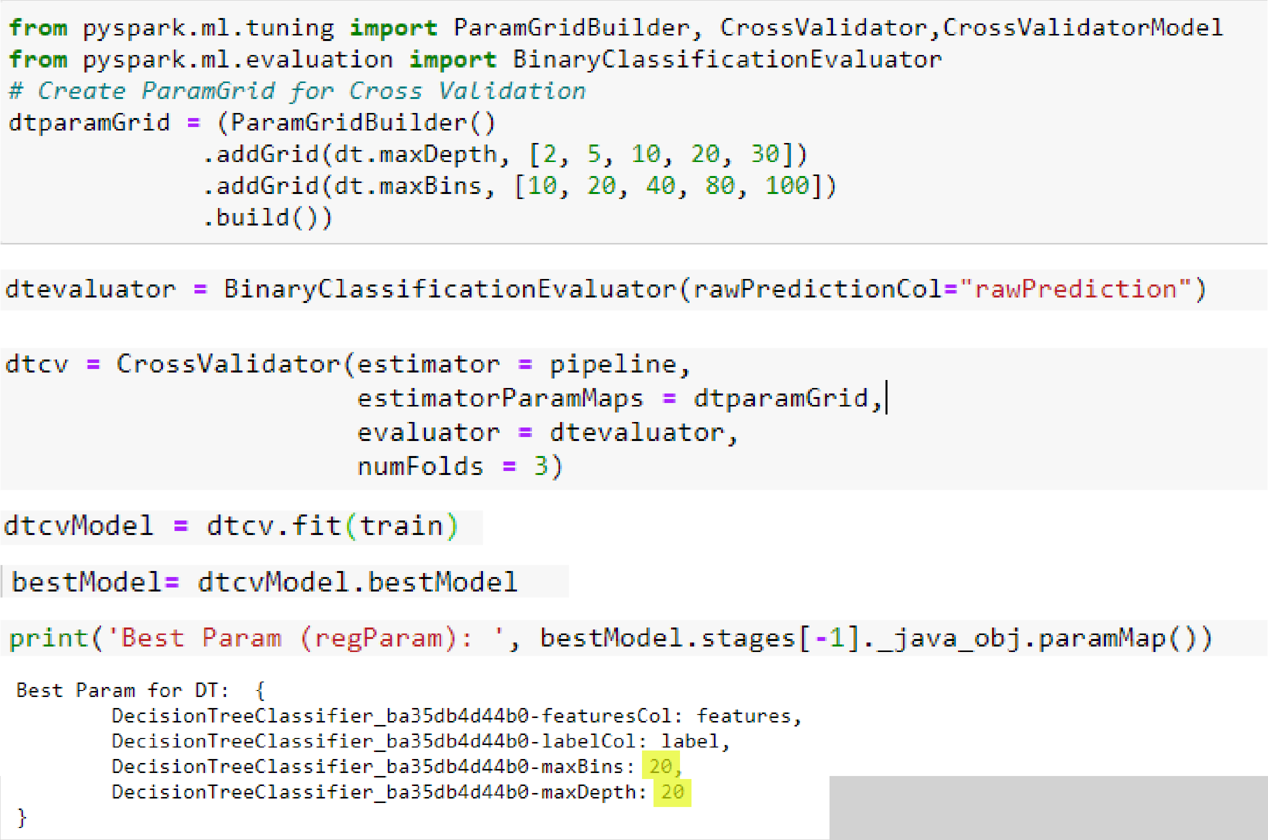
Cross Validation (K-Fold)

- Splitting dataset into a set of folds, which are used as separate training and test datasets.



http://www.ds100.org/sp17/assets/notebooks/linear_regression/Cross_Validation_and_the_Bias_Variance_Tradeoff.html

Cross Validation (Decision Tree)



```
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator,CrossValidatorModel
from pyspark.ml.evaluation import BinaryClassificationEvaluator
# Create ParamGrid for Cross Validation
dtparamGrid = (ParamGridBuilder()
             .addGrid(dt.maxDepth, [2, 5, 10, 20, 30])
             .addGrid(dt.maxBins, [10, 20, 40, 80, 100])
             .build())
```

```
dtevaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
```

```
dtcv = CrossValidator(estimator = pipeline,
                      estimatorParamMaps = dtparamGrid,
                      evaluator = dtevaluator,
                      numFolds = 3)
```

```
dtcvModel = dtcv.fit(train)
```

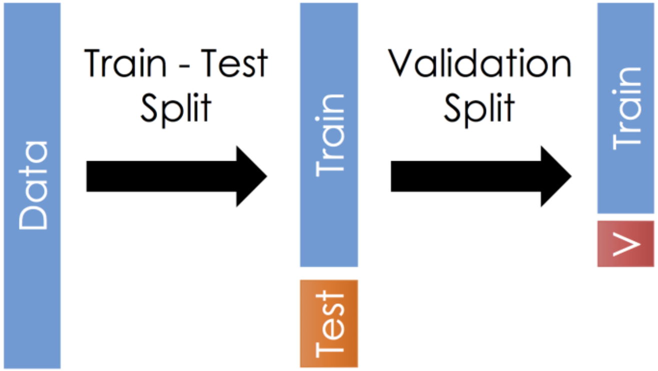
```
bestModel= dtcvModel.bestModel
```

```
print('Best Param (regParam): ', bestModel.stages[-1]._java_obj.paramMap())
```

```
Best Param for DT:  {
	DecisionTreeClassifier_ba35db4d44b0-featuresCol: features,
	DecisionTreeClassifier_ba35db4d44b0-labelCol: label,
	DecisionTreeClassifier_ba35db4d44b0-maxBins: 20,
	DecisionTreeClassifier_ba35db4d44b0-maxDepth: 20
}
```

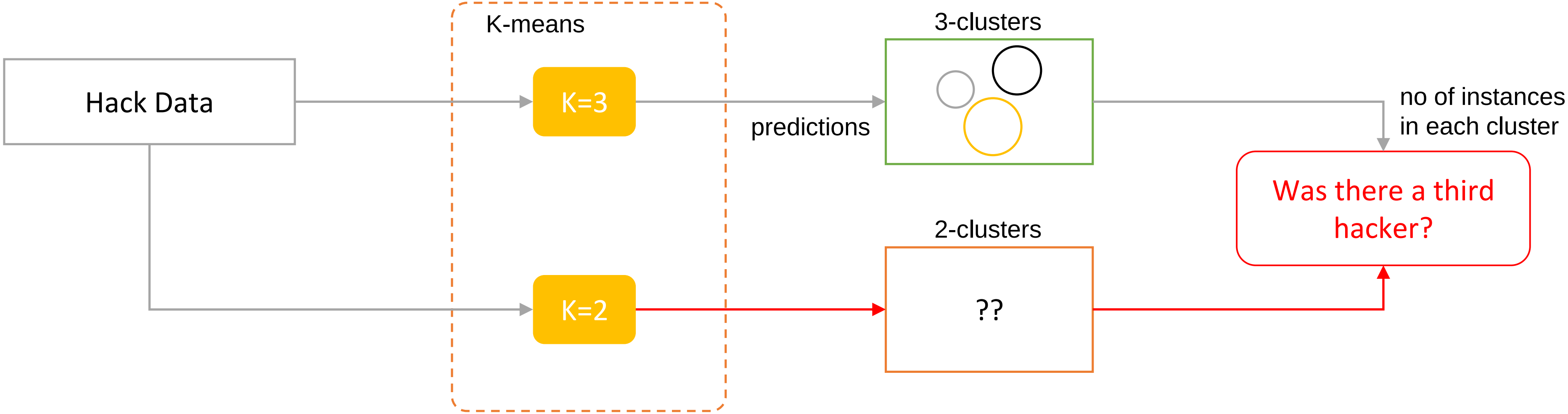
TrainValidationSplit

- Creates a single dataset pair
- Only evaluates each combination of parameter once as opposed to k-times in case of CrossValidator
- Less expensive but not reliable if the training dataset is not large enough



Use case : Was there a third hacker?

Assumption : Hackers trade off attacks equally



Thank You!

See you next week.