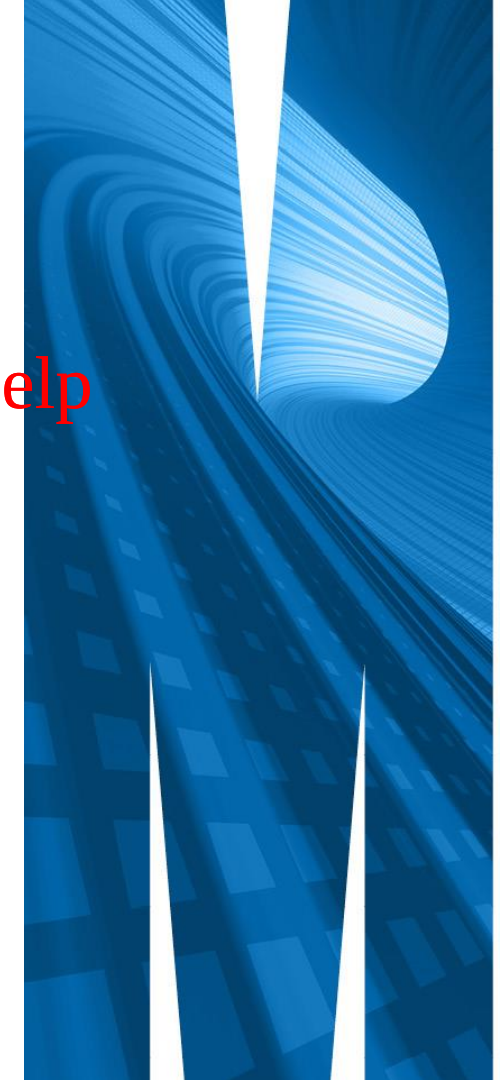
MONASH University

Week 7

FIT5202 Big Data Processing

K-Means Clustering

Model Selection



Week 7 Agenda

- **Part - A**
- Week 6 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 Persistance
  - Saving and Loading a Model

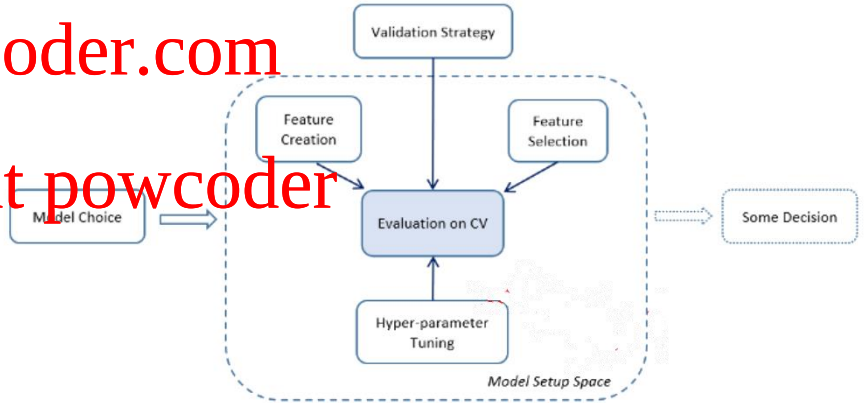
Model Selection

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

- **HyperParameter Tuning**
- 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:

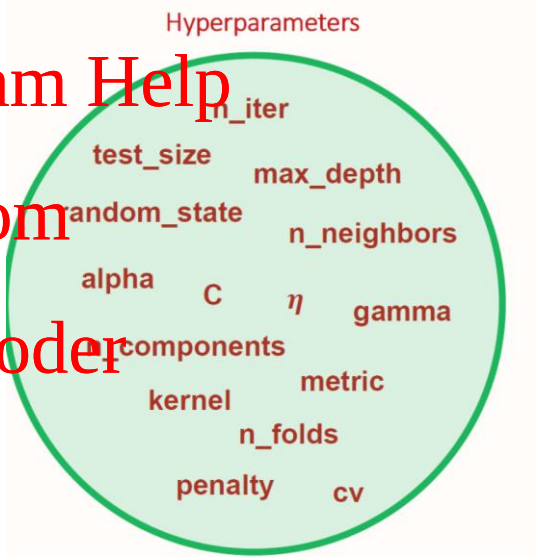
1. CrossValidator
2. TrainValidationSplit



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

Hyperparameter Tuning

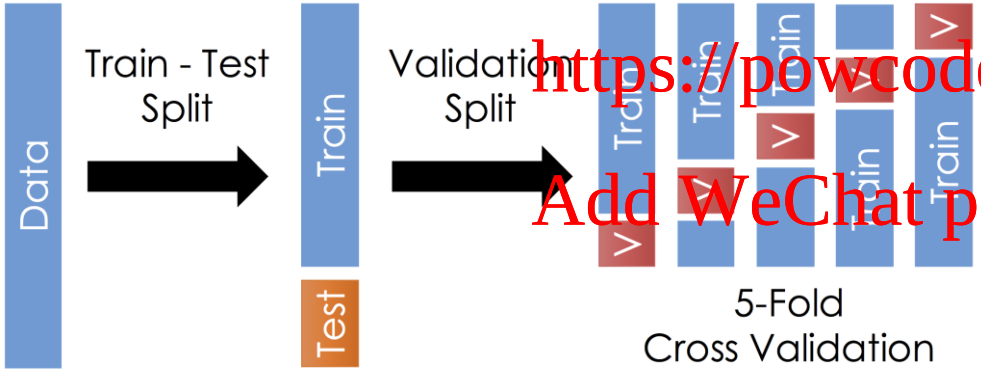
- Hyper-parameters are not model parameters : they cannot be trained from the data
- Hyperparameter tuning : choosing a set of optimal hyperparameters for a learning algorithm
- model.getParamMap() to get the list of hyperparameters for the model



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

Cross Validation (K-Fold)

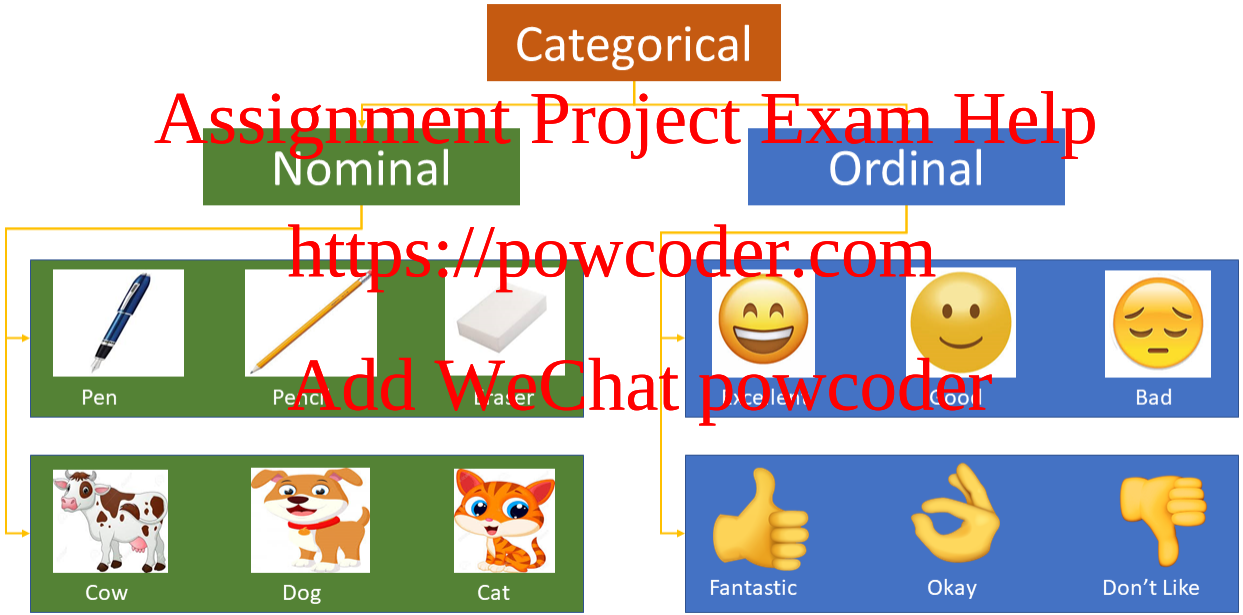
- Splitting dataset into a set of folds, which are used as separate training and test datasets.



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

Categorical features

Categorical variables represent types of data which may be divided into groups.



No ordering

The variables have natural, ordered categories

Hyperparameter: maxBins

Continuous features

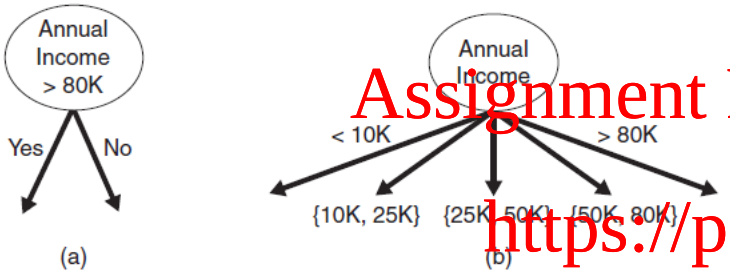


Figure 4.11. Test condition for continuous attributes.

- The test condition can be expressed as a comparison test ($A < v$) and ($A > v$) with binary outcome, or a range of a range of outcomes $v_i < A < v_{i+1}$ for i=1,…, *k*
- For binary tree, algorithm will consider all split position *v* (splitting point / threshold)

Example

Consider variable *X* with instances [1,3,4,6,2,5,18,10,-3,-5]

We can sort data, and cluster data into bins to choose splitting point (e.g., -1,2.5,4.5, and 8)

[-5,-3,1,2,3,4,5,6,10,18]

[-5,-3],[1,2],[3,4],[5,6],[10,18]

Maximum number of bins can be specified using maxBins.

https://spark.apache.org/docs/1.1.0/mllib-decision-tree.html

https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf

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)
```



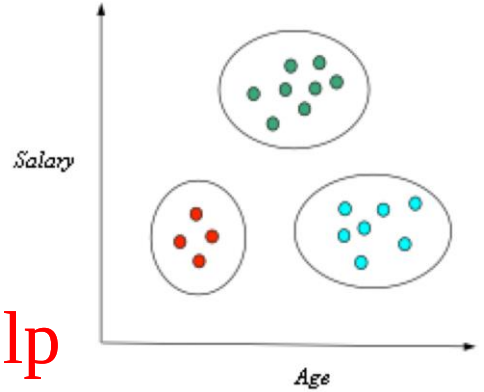
```
dtcvModel = dtcv.fit(train)
```

```
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
}
```

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

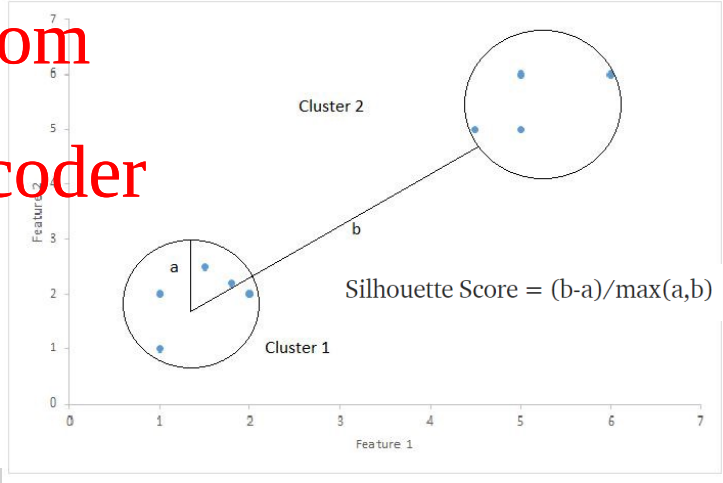
Groups have no category or label

Unsupervised learning

Animation Demo , DEMO 2

Silhouette Score [-1 1] : calculates the goodness of a clustering technique

- **1** - Clusters are well apart from each other and clearly distinguishes
- **0** - Clusters are not clearly distinguished, the distance between the clusters is not significant (overlapping cluster)
- **-1** – Clusters assigned wrongly



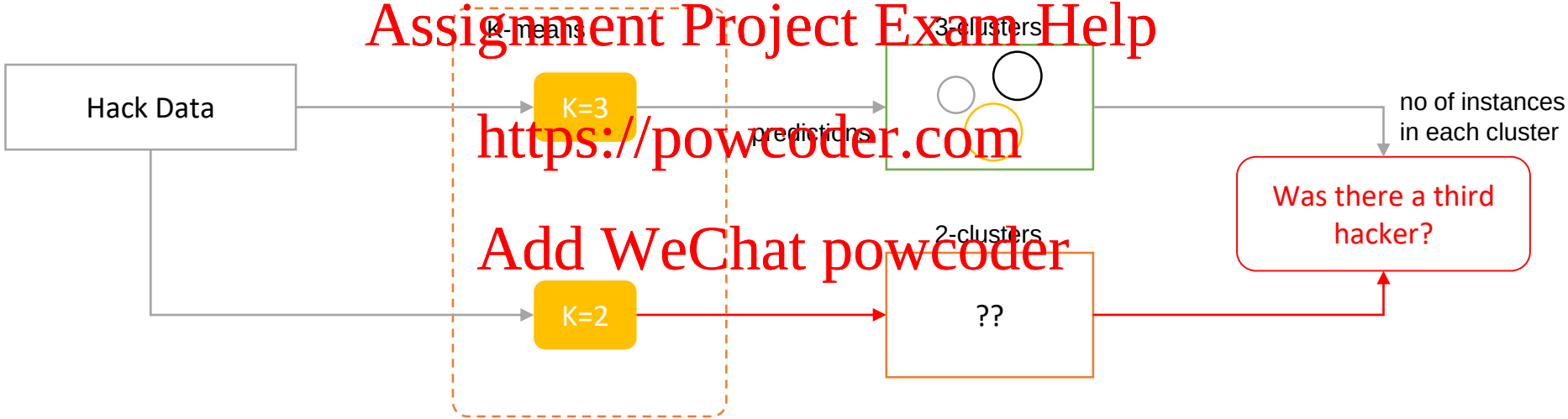
https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

a= average intra-cluster distance

b= average inter-cluster distanc

Use case : Was there a third hacker?

Assumption : Hackers trade off attacks equally



Thank You!

See you next week.