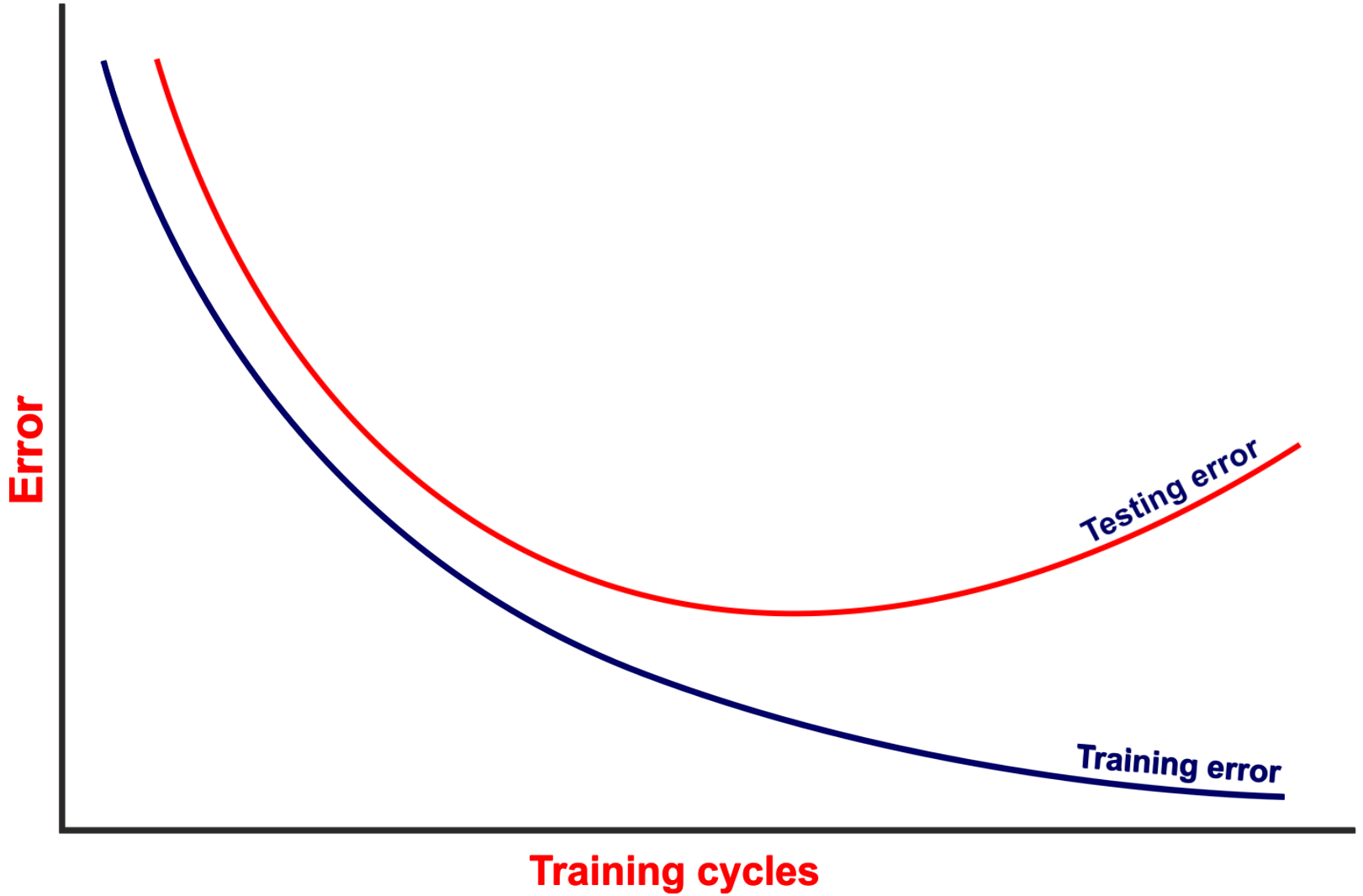
COMP307/AIML420 Week 3 (Tutorial)

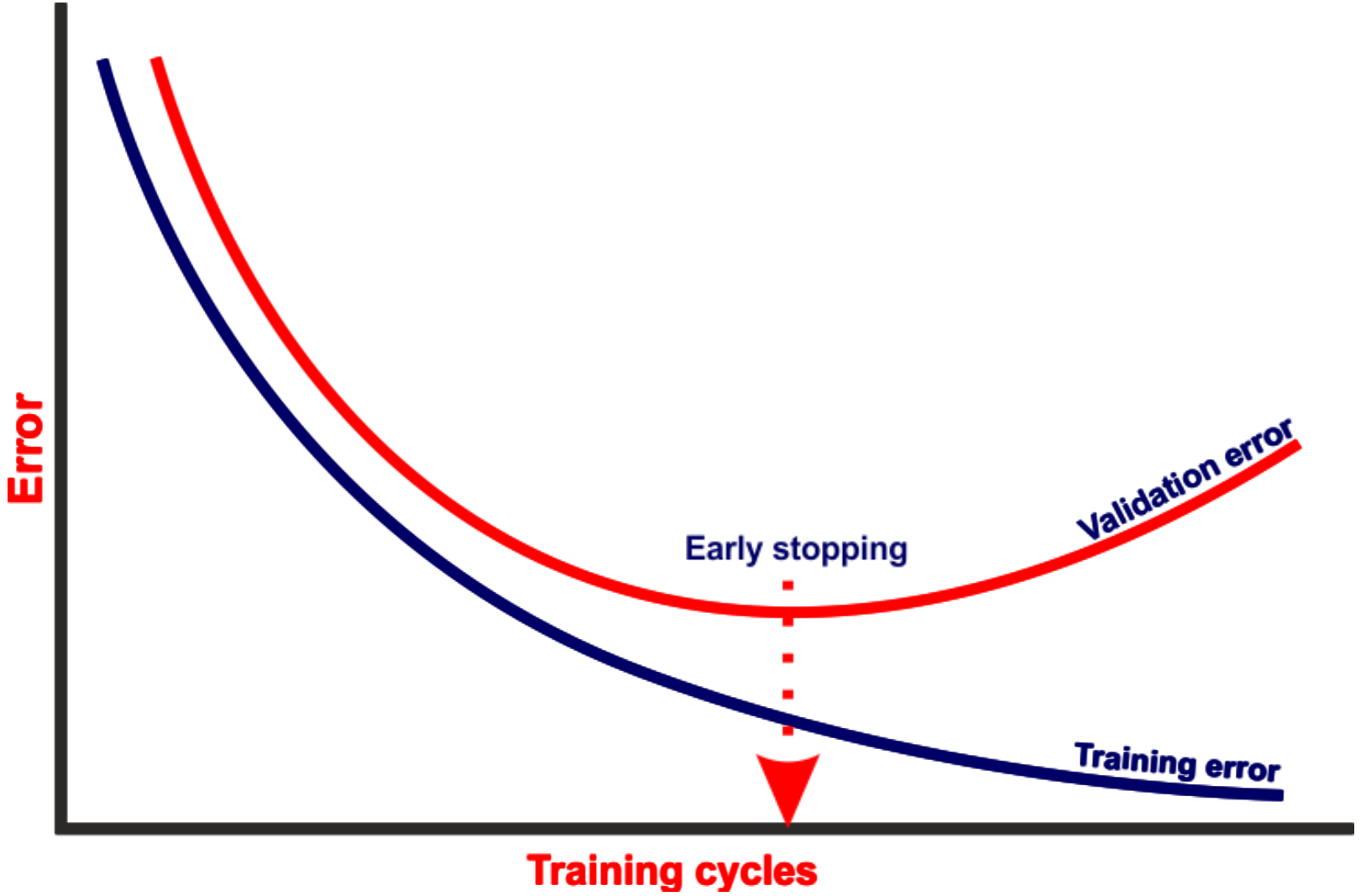
1. Announcements
   - Assignment 1 (**15%**)
   - Helpdesk sessions
2. Sets
   - Training and Test sets
   - Validation set
3. Datasets
   - Instances
   - Features and feature vectors
   - Class label
4. 3-K Algorithms
   - k-Nearest Neighbour
   - k-Means Clustering
   - k-fold Cross Validation
5. Decision Trees
   - DT learning vs learned DT
   - Impurity measure Conditions
   - Pruning
6. Other Questions

Over-fitting

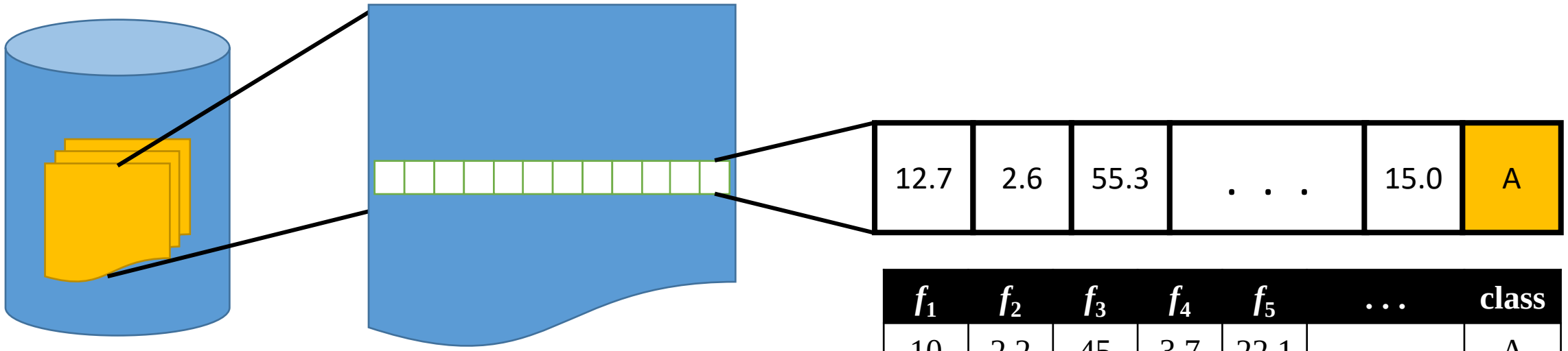


Validation Set

- What?
- Why?
- How?



Datasets and Instances



f_1	f_2	f_3	f_4	f_5	. . .	class
10	2.2	45	3.7	22.1	. . .	A
3.7	7.9	12	2.1	17.5	. . .	A
22.8	27.9	11.4	36	77	. . .	B
90.4	6.34	2.77	15.8	53.7	. . .	A
74.6	4.78	84.9	15.9	103	. . .	B
2.89	14.7	3.11	10	52	. . .	B

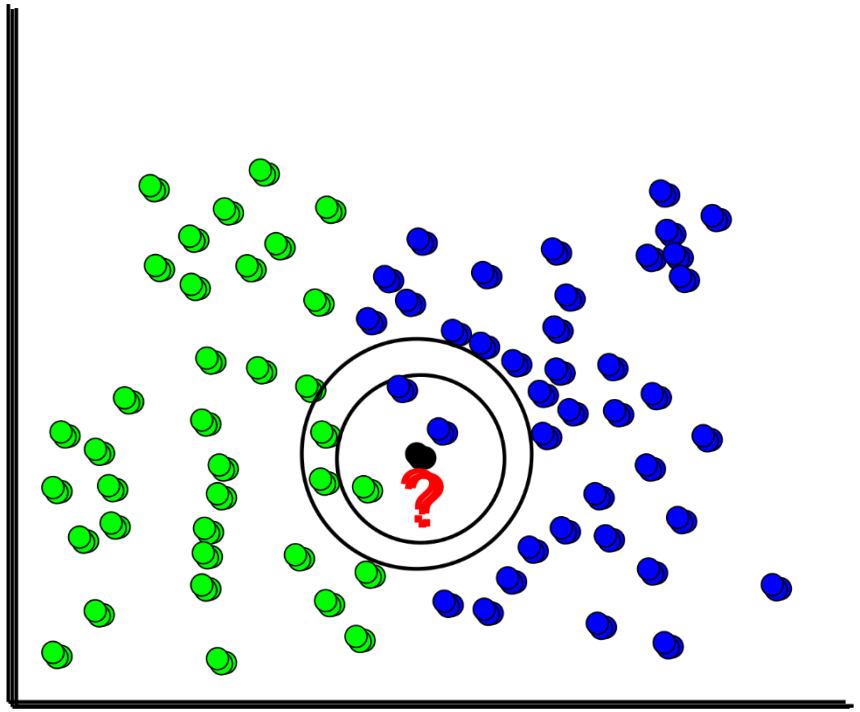
K-Nearest Neighbour

Training Set

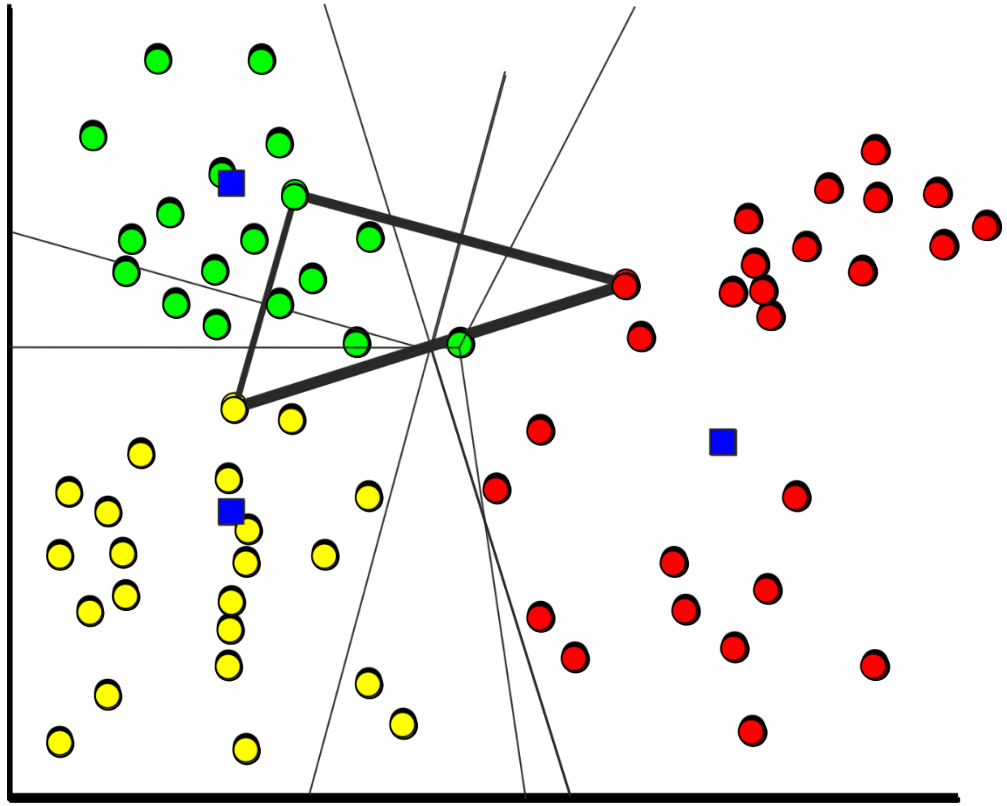
$$d(\vec{u}, \vec{v}) = \sqrt{\sum_{i=1}^{N} (\vec{u}_i - \vec{v}_i)^2}$$

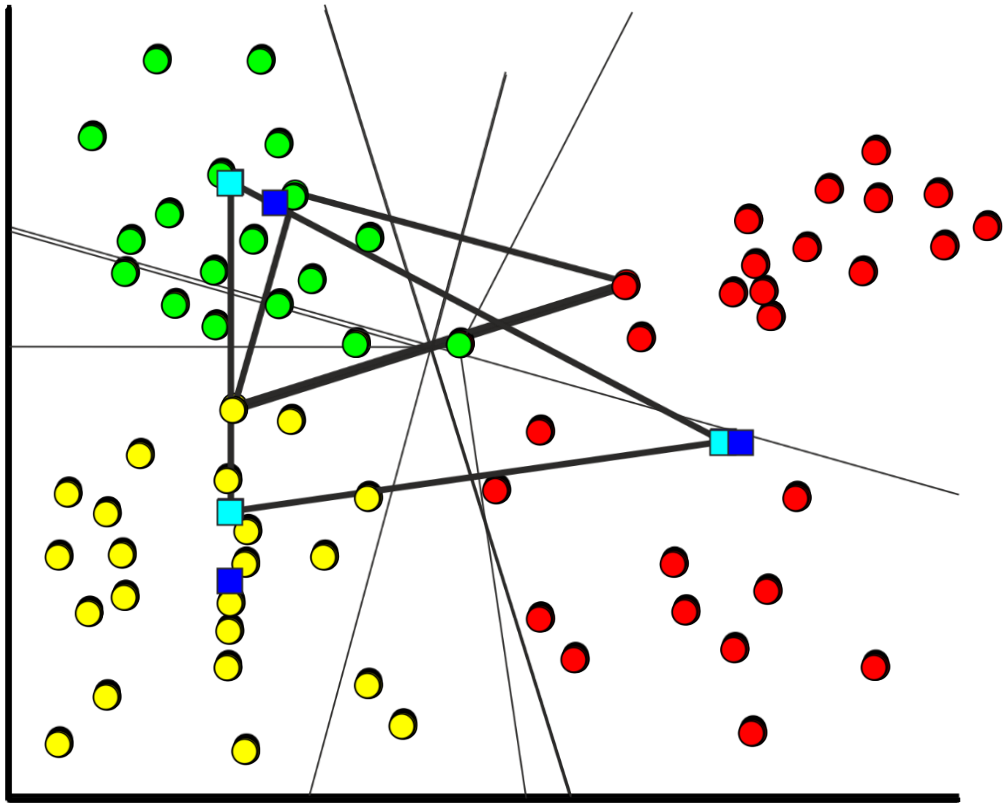
f_1	f_2	f_3	class		
10.3	45.7	2.7	A	$d(\cdot,\cdot) =$	14.84
7.1	80.5	1.1	A	$d(\cdot,\cdot) =$	47.40
22.3	20.4	9.6	B	$d(\cdot,\cdot) =$	24.57
30.5	21.2	17.9	B	$d(\cdot,\cdot) =$	33.65
5.2	67.1	7.7	A	$d(\cdot,\cdot) =$	33.88
15.6	18.6	11.4	B	$d(\cdot,\cdot) =$	21.19
11.9	53.4	6.3	A	$d(\cdot,\cdot) =$	22.24

2.1	33.5	4.7	A



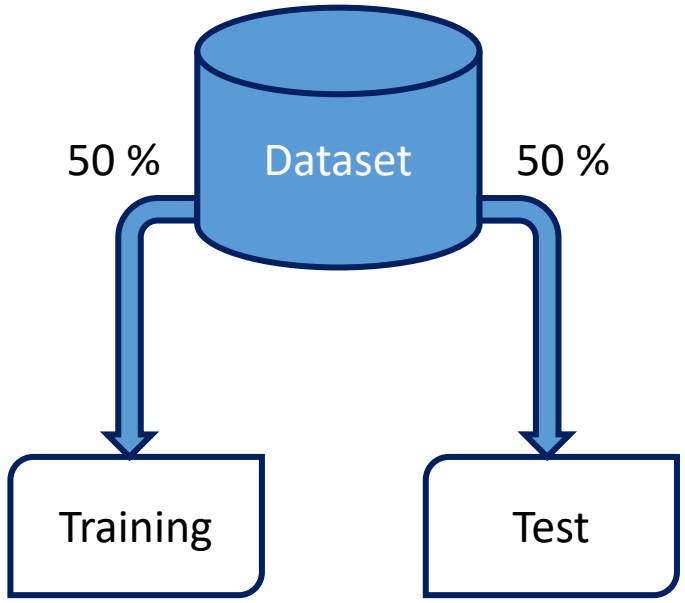
k-Means Clustering

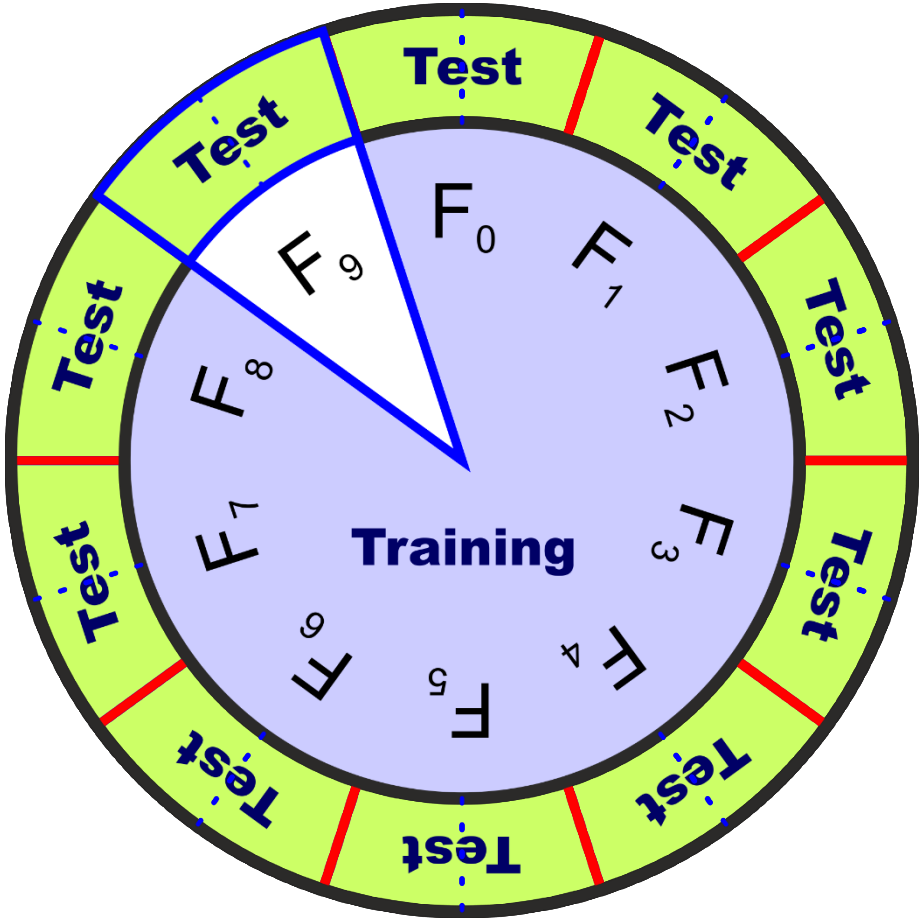




Demo

k-fold Cross Validation





75 %
92 %
43 %
76 %
30 %
82 %
83 %
47 %
50 %
73 %

65.10 %

- How do we specify **the number** of instances in each fold?
- How do we **select** those instances?

Decision Trees (DT)

- DT learning ≠ learned DT
- Impurity measure
  1. 0, if all instances belong to one class
  2. Max, equal number of instances for both classes
  3. Continuous –smooth
- Large vs. small trees?

Other Questions

- Can k-NN can cope with categorical data/features?
- Does k-Means always provide the same clusters?
- How do we select/initialise the seeds in k-Means?