

2/28/22

KNN.py

- uses pickle library, pickle.dump()
- compare it & get accuracy
- plot confusion matrix
- PCA
- granularity matters

Splitting Based on Continuous Attributes

Discretization: form an ordinary categorical attribute
= static / dynamic

Binary Decision: $(A < v)$ or $(A \geq v)$ for continuous value

How to determine Best Split

Split 1: Not a good split \therefore 50%, 50% Not very useful

Split 2: Have more information about the question asked

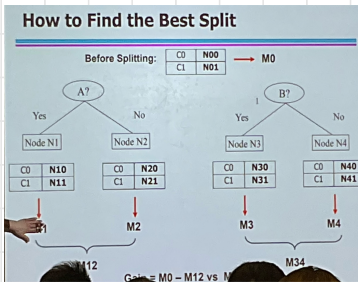
Split 3: Good Split, Not a useful split.

Approach:

need to measure **impurity**

Measurement for impurity:

Gini Index



Measure of Impurity: GINI

- Gini Index for a given node t:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

(NOTE: $p(j|t)$ is the relative frequency of class j at node t.)

- Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

C1	0
C2	6
Gini	=0.000

C1	1
C2	5
Gini	=0.278

C1	2
C2	4
Gini	=0.444

C1	3
C2	3
Gini	=0.500

The lower the GINI, the better

Stopping Criteria for Tree Induction

- have similar attribute values
- records belong to same class
- early termination

ex: pass a certain frequency / Gini Index

Cost v.s. Accuracy

Cost-sensitive Measurement

Methods of Estimation:

holdout: reserve $\frac{2}{3}$ for training
 $\frac{1}{3}$ for testing



Cross Validation: partition to k subsets

k -fold

$k = n$