2/28/22

kNN.py

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

Splitting Based on Continous Attributes

Discretization: form an ordinary categorical attribute

: Static / Dynamic

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

How to determine Best Split

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

Split 2: Have more information about the question asked

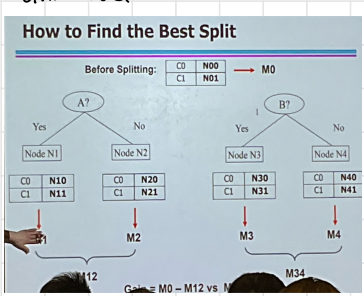
Split 3: Good Split, Not a useful split.

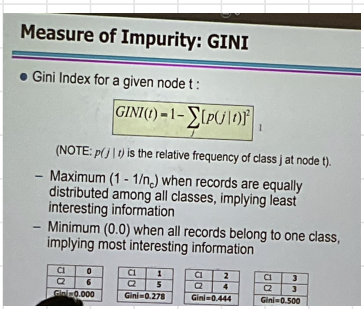
Approach:

need to measure impurity

Measurement for impurity:

Gini Index





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 2/3 for training
1/3 for testing

Cross Validation. partition to k subsets

k-fold

k=n