Decision trees

Monday, February 28, 2022 4:33 PM

Confusion matrix demo. Stored in folder 7-knn.
Use PCA to reduce the dimension of the data, and speed up the program.

Center the data before using PCA.
-mean center it
-mean center and divide it by the std.

Hunt's algorithm

- The attribute that would best split your data
- Generates splits in our data to show separation in the data
- Yes and No$_s$ in the decision
- Multi-way split
- Binary split

Continuous attributes

- Discretization to form an ordinal categorical attributes
- Binary decision

How to determine the best split

- Binary -> multiway split, which one is better?
- Greedy approach:
    - ○ Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity.

Measure of node impurity

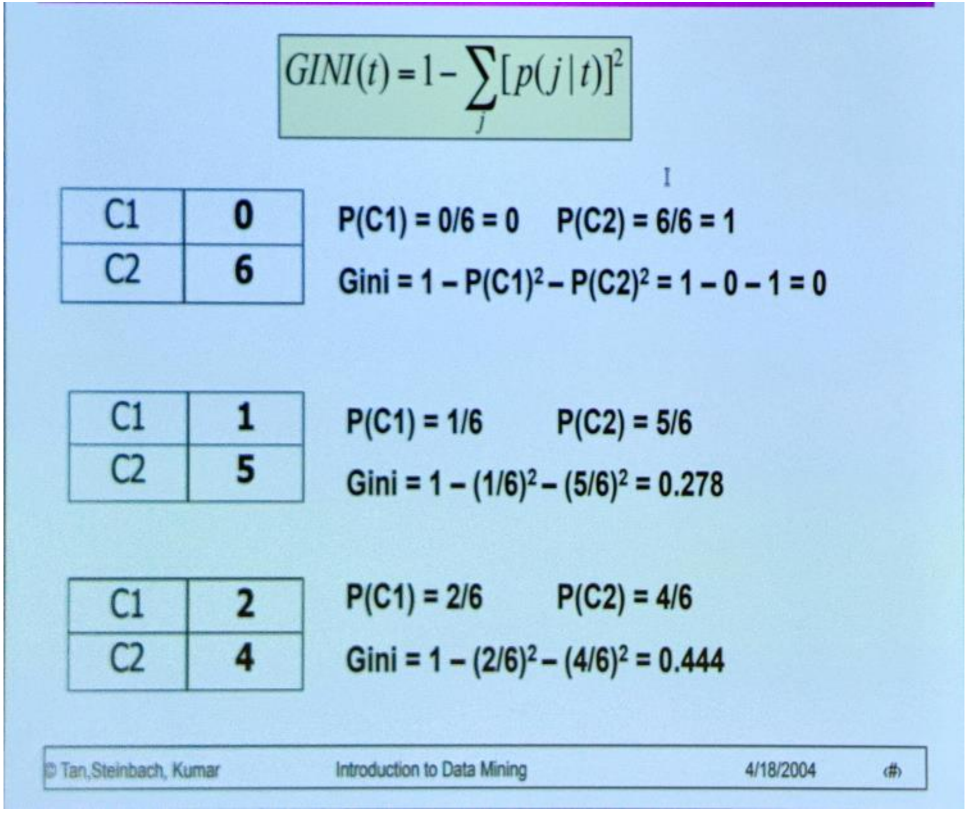
- Gini index

How to determine the best split

- Degree impurity
- And comparing between impurity
- Impurity before split - impurity after impurity

GINI

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



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

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – $P(C1)^2$ – $P(C2)^2$ = 1 – 0 – 1 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278

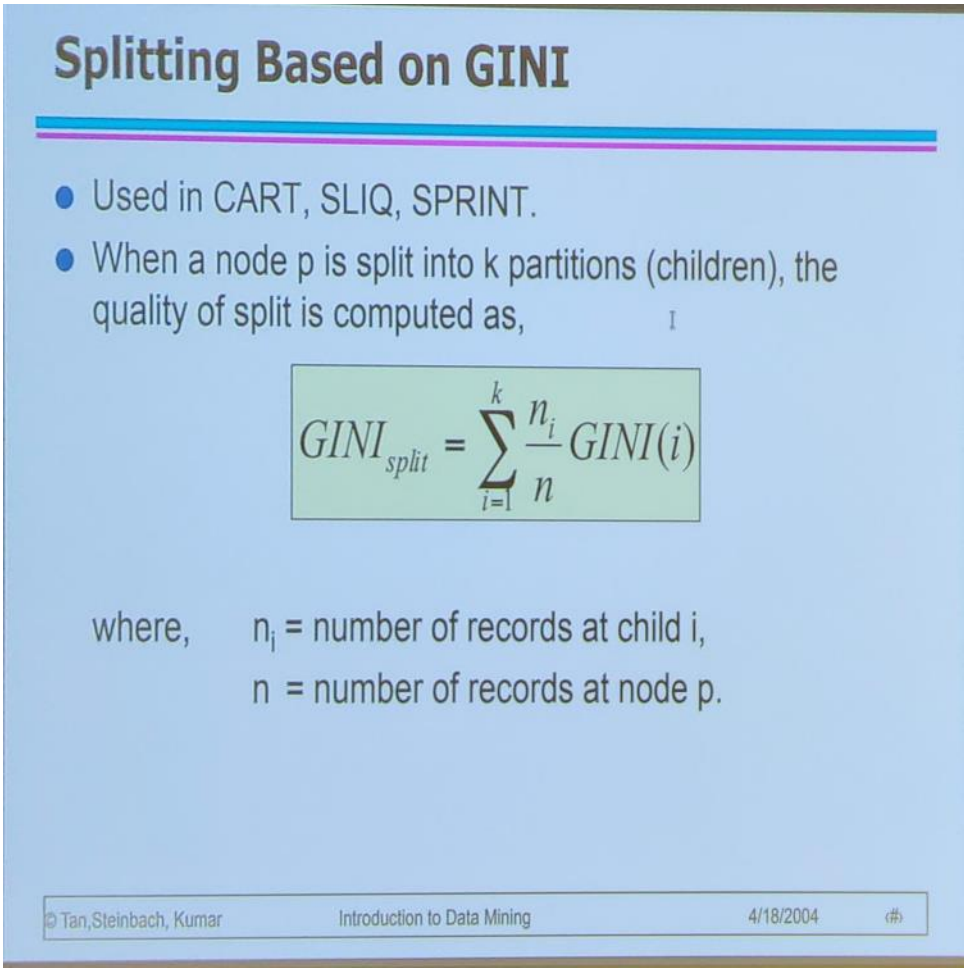
C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

-



Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

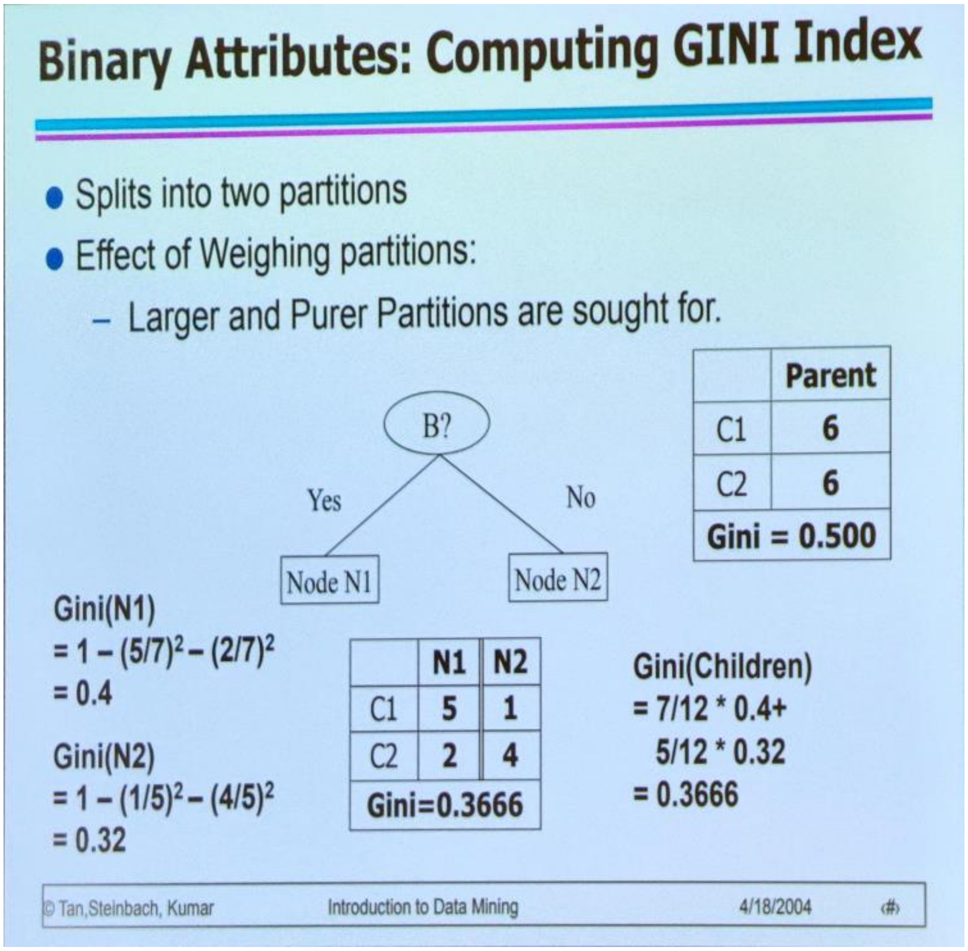
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i,

n = number of records at node p.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

Binary Attribute Computing GINI Index


B?
Yes
No
Node N1
Node N2

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.

	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
$= 1 - (5/7)^2 - (2/7)^2$
$= 0.4$

Gini(N2)
$= 1 - (1/5)^2 - (4/5)^2$
$= 0.32$

	N1	N2
C1	5	1
C2	2	4
Gini=0.3666		

Gini(Children)
$= 7/12 * 0.4 +$
$5/12 * 0.32$
$= 0.3666$

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

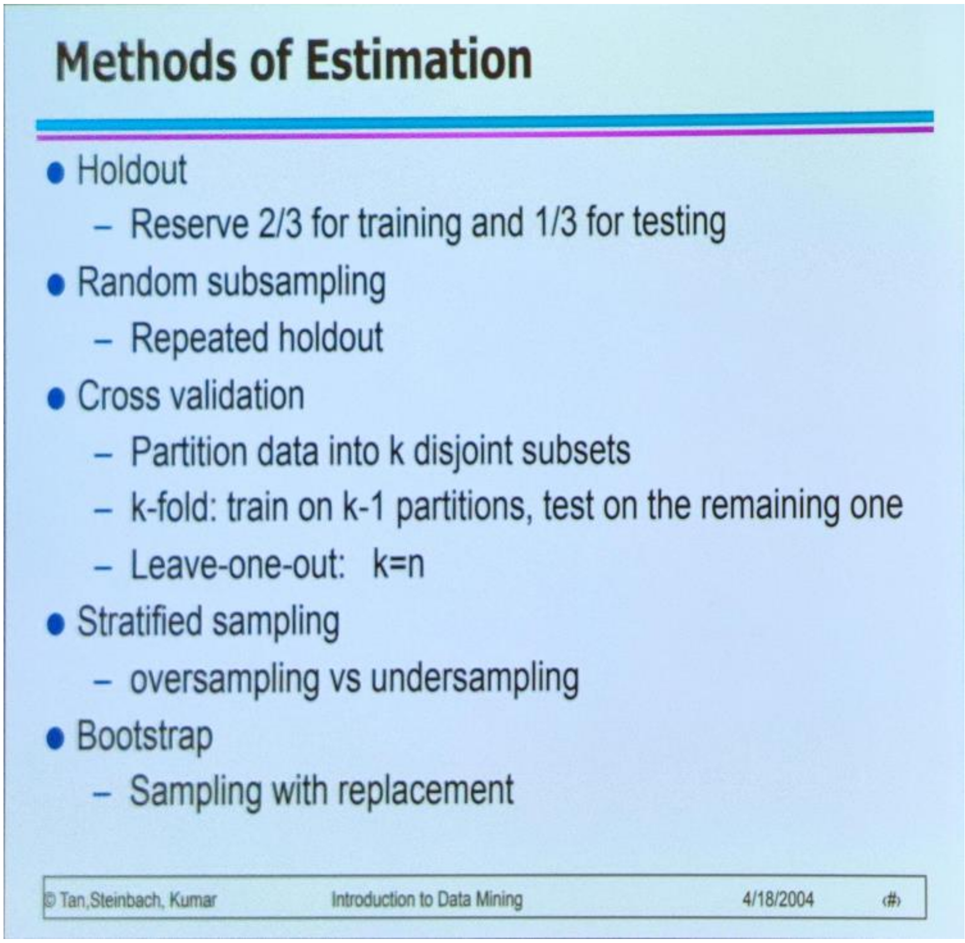
Continuous attributes: computing gini index

- Speed up by not recomputing things, but it is expensive algo

Skipped entropy and misclassification error.

Stopping criteria for tree induction
-stop expanding a node when all the records belong to the same class
-Stop expanding a node when all the records have similar attribute values
-early termination (to be discussed later)

Methods of Estimation:



Methods of Estimation

- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Random subsampling
  - Repeated holdout
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n
- Stratified sampling
  - oversampling vs undersampling
- Bootstrap
  - Sampling with replacement

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 ‹#›

Bagging
-sampling with replacement
-build classifier on each ootstrap sample

- Each sample has probability

Boosting
-an iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records.