Decision Tree

Data Set with Categorical Features

Each path in Decision Tree is decision rule

- internal nodes → attributes.
- Leaf nodes → classification

path : root → leaf : Conjuction of attribute test

Algorithm

function dtree(*examples*, *attributes*, *default*) **returns** a decision tree
if *examples* is empty **then**
  return *default*
else if all *examples* have the same class label **then**
  return the class label
else if *attributes* is empty **then**
  return majority_class(*examples*)
else
  best ← choose_attribute(*attributes*, *examples*) Table
  tree ← a new decision tree with root test *best*
  for each value v_i of *best* **do**
    $examples_i$ ← {elements of *examples* with *best* = v_i}
    m ← majority_class($examples_i$)
    subtree ← dtree($examples_i$, *attributes* − *best*, m)
    add a branch to *tree* with label v_i and subtree *subtree*
  end for
  return *tree*
end if

best var
v_1 / v_2 v_3

How to choose attributes ?

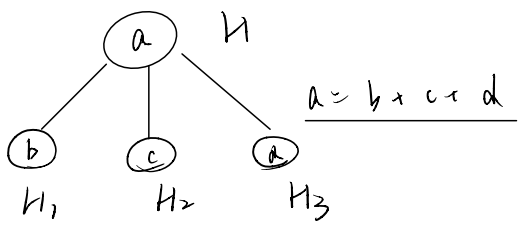
Choose the attributes that leads to the maximum reduction in impurity

Entropy:

p_i : fraction of example belongs to class i

Entropy: $-\sum_{i=1}^{k} p_i \log p_i$ ($0 \log_2 0 = 0$ by convention)

Information Gain: $= H - \left(\frac{b}{a} H_1 + \frac{c}{a} H_2 + \frac{d}{a} H_3\right)$



Entropy: $H(X) = \sum_x P(x) \log \frac{1}{P(x)}$

Conditional entropy:

given $X = x$. $H(Y|X=x) = \sum_x P(Y|X=x) \log \frac{1}{P(Y|X=x)}$

given X : $H(Y|X) = \sum_x P(X=x) H(Y|X=x)$

Mutual Information:

$I(X;Y) = H(X) - H(X|Y)$

Average reduction in uncertainty about X from learning the value of Y

or Average of information Y conveys about X.

Mutual Information ≡ information gain

Mutual Information & KL divergence

$I(X;Y) = \sum_x P(x) \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \log \frac{1}{P(x|y)}$

- Note that:

$$I(X;Y) = \sum_X P(X) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)}$$
$$= \sum_{X,Y} P(X,Y) log \frac{1}{P(X)} - \sum_{X,Y} P(X,Y) log \frac{1}{P(X|Y)}$$
$$= \sum_{X,Y} P(X,Y) log \frac{P(X|Y)}{P(X)}$$
$$= \sum_{X,Y} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)} \quad \text{equivalent definition}$$
$$= KL(P(X,Y)||P(X)P(Y)) \quad \text{symmetric}$$

a. b c

$I(X;Y) = I(Y;X)$

$d(a,c) \le d(a,b) + d(b,c)$

- Due to equivalent definition:

$$I(X;Y) = H(X) - H(X|Y) = I(Y;X) = H(Y) - H(Y|X)$$

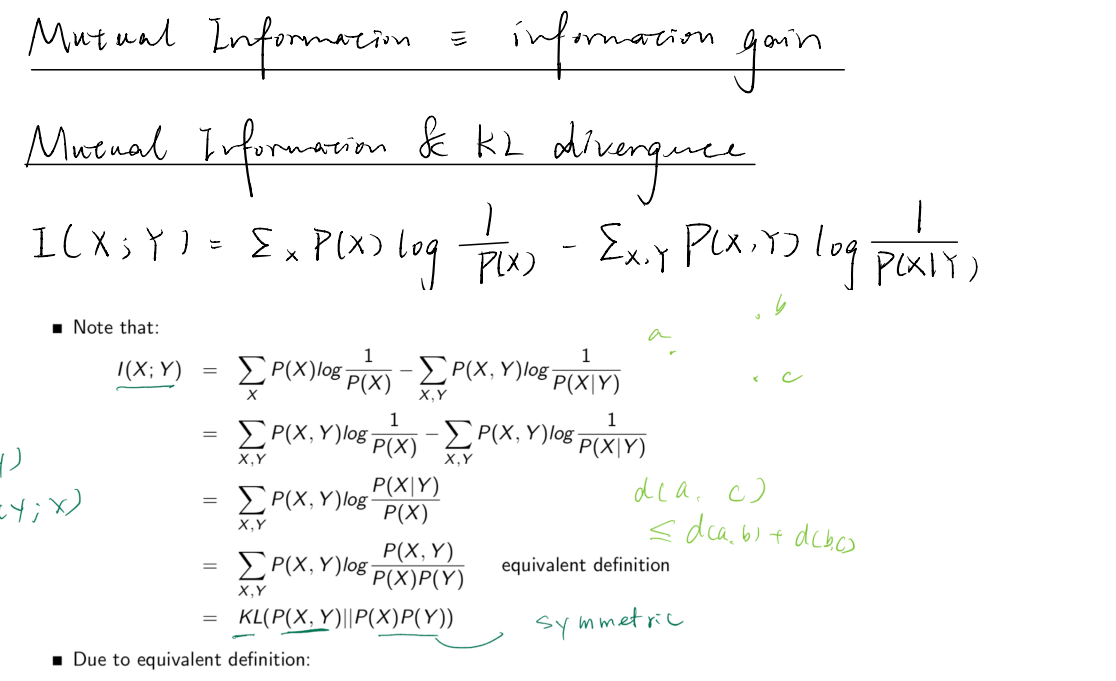
$I(X;Y) \ge 0$

when $I(X;Y) = 0$? when X & Y independent

- As a decision tree gets deeper, fewer examples are associated with the deeper nodes, often leading to overfitting with large variance.
- It may not be good to fully expand a tree or else overfitting may occur.
- Pruning may be applied to reduce the size of the tree

Decision Tree Ensemble

① Bagging.



Reduce variance !

Dependent !

② Random forest

- The decision trees in a random forest are like ordinary decision trees with only one exception in the learning algorithm: *each decision node is created by selecting from a random subset of the attributes (instead of all the attributes)*, called **attribute bagging**.
- A random forest is random in two senses:
  - Sample bagging: random sampling of training data points when building trees
  - Attribute bagging: random subsets of features considered when splitting nodes.

Continuous-valued Attributes

"Discretize"

Boolean Variable A_c : if $A \le c$

How to pick c ? median