02_Decision_Tree

Test attributes sequentially, arbitrary boolean function can be represented.

Alogrithm

1. Selection of attribute: Choose attributes that maximaize information gain.
2. Generate tree. Internal node: attribute. Leaf node: Class
3. Prouning

Information Gain

The Shannon information content of an outcome is: $-log_2 p(x_i)$

- **Entropy**(↑) denotes measure of uncertainty(↑) or unpredictability.
  Less entropy more purification.

$$H(X) = \sum_i -p(x_i) log_2 p(x_i)$$

  $p(x_i)$: probability for random toss coin $X = x_i$,
  $p(x_i) = 0.5, H(x) = 1\ bit.\ maximal\ entropy, highest\ uncertainty$
  $p(x_i) = 0\ or\ 1, H(x) = 0.\ minimal\ entropy, no\ uncertainty$

- **Conditional Entropy** denotes uncertainty of random variable Y given random variable X.

$$H(Y \mid X) = \sum_i p(X = x_i) \sum_j -p(Y = y_j \mid X = x_i) log_2 p(Y = y_j \mid X = x_i)$$

$$note: \sum p = 1$$

- **Information Gain**
  Information gain of attribute A in train data D,utlizing entropy and conditional entropy

$$g(D, A) = H(D) - H(D \mid A)$$

Information gain is the change in information entropy H from a prior state to a state that takes some information. It denotes reduction of class D's entropy after acquiring attribute A.

- Gini impurity:
  Another definition of predictability (impurity).

$$\sum_i p_i(1 - p_i) = 1 - \sum_i p_i^2$$

Prouning

- **Overfitting**
  Good results on training data, but generalizes poorly. Happens when:
  $Non\ representative\ sample\ (few\ sample), Noisy\ examples, Too\ complex\ model$

Choose a simpler model and accept some errors for the training examples

- **Occam's Razor**

  The *simplest explanation* compatible with data tends to be the right one.
- **Reduced-Error Prouning**

Split data into training and validation set
Do until further pruning is harmful:

1. Evaluate impact on validation set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves validation set accuracy

Produces smallest version of most accurate subtree