Dis Fri

Friday, October 12, 2018 12:16 PM

(Quick Review + examples of Variance & Expectation)

Decision Tree

- a common classification method which can deal with nonlinear separation
- hierarchical structure:
  - node: attribute name
  - edge: attribute value
  - leaf: label
- Outputs are usually discrete categories
- Real valued outputs also possible (regression trees)

ID3 is the algorithm we use to build decision trees

1. If all examples have same label, return a single node tree with that label
2. Else create a root node for tree
3. A = attribute in Attributes that *best* classifies S
4. For each possible value v of fA
   a. add a new tree branch corresponding to A=v
   b. Let Sv be the subset of examples in S with A=v
   c. if Sv is empty: add leaf node with the common value of Label in S
   d. else: below this branch add the subtree ID3(Sv, Attributes - {a}, Label)
5. Return Root

How to choose the best attribute in step 3?

Information Gain and Entropy

Entropy measures uncertainty in the dataset

Information Gain is the difference between the entropies of two states

Entropy:

$$H[S] = -\sum_{v=1}^{K} P(S = a_v) \log_2 P(S = a_v)$$

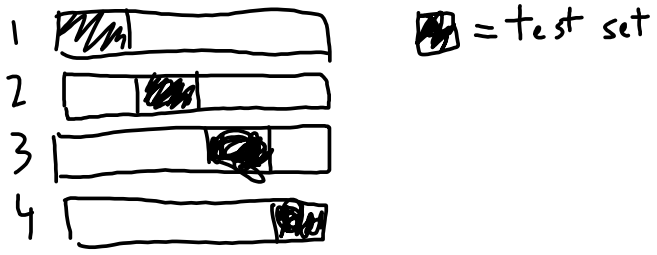
(same example as in lecture - play tennis? given outlook, temperature, humidity, wind)

Scikit-Learn

- a free software machine learning library for Python
- various classification, regression, clustering algorithms

Cross-Validation

- we split our data into training and test sets in multiple experiments:



(scikit-learn demo)