Data Mining, CSCI 347, Fall 2019
Decision Trees, Sept. 18

Information gain = info. before splitting - info. after splitting

Info. before splitting is the information on top

Info. after splitting is the sum of the information at each branch, weighted by the number of attributes that took that branch

Problem – highly branching attributes

- Subsets of highly branching attributes are likely to be pure, making the "information gain" high.
- This can result in overfitting (selecting an attribute that is non-optimal for predication.)
- Avoid selection of highly branching attributes by using a "gain ratio"

- Gain ratio – Gain from Attribute / Intrinsic Value of Split
- Intrinsic information – entropy of distribution of instances into branches ignoring the class values