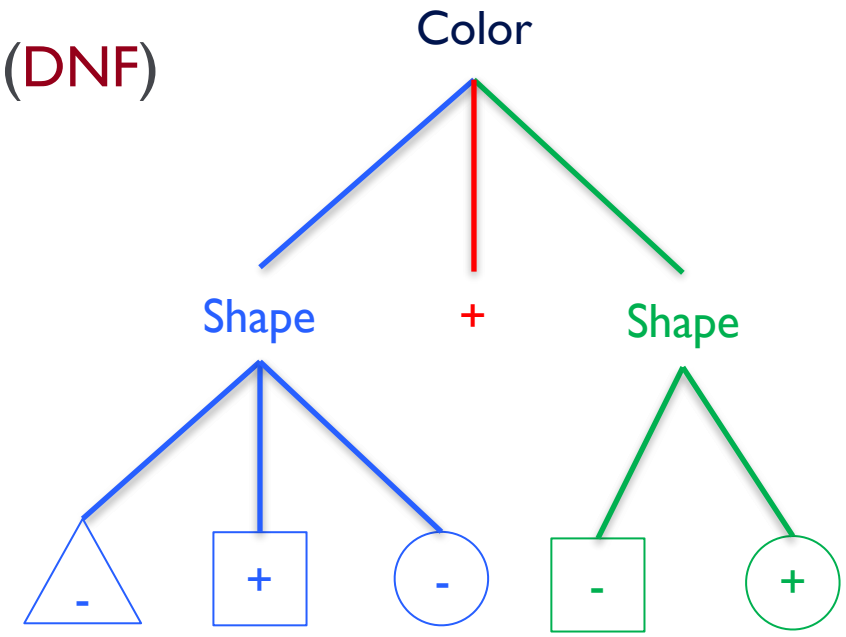
CIS 519 Recitation 3

Decision Tree

- A hierarchical data structure that represents data by implementing a divide and conquer strategy
  - color={red, blue, green} ; shape={circle, triangle, rectangle} ; label= {+, -}
- Nodes are tests for feature values, leaves specify the category (labels)
- Can be rewritten as rules in Disjunctive Normal Form (DNF)
- What is the hypothesis space here?
  - 2 features: color and shape
  - 3 values each: color(red, blue, green), shape(triangle, square, circle)
  - |X| = 9: (red, triangle), (red, circle), (blue, square) …
  - |Y| = 2: + and -
  - $|H| = 2^9$



Learning a Decision Tree: ID3

- Let S be the set of Examples
  - Label is the target attribute (the prediction)
  - Attributes is the set of measured attributes
- We want attributes that split the examples to sets that are relatively pure in one label
- when p_i is the fraction of examples labeled i:

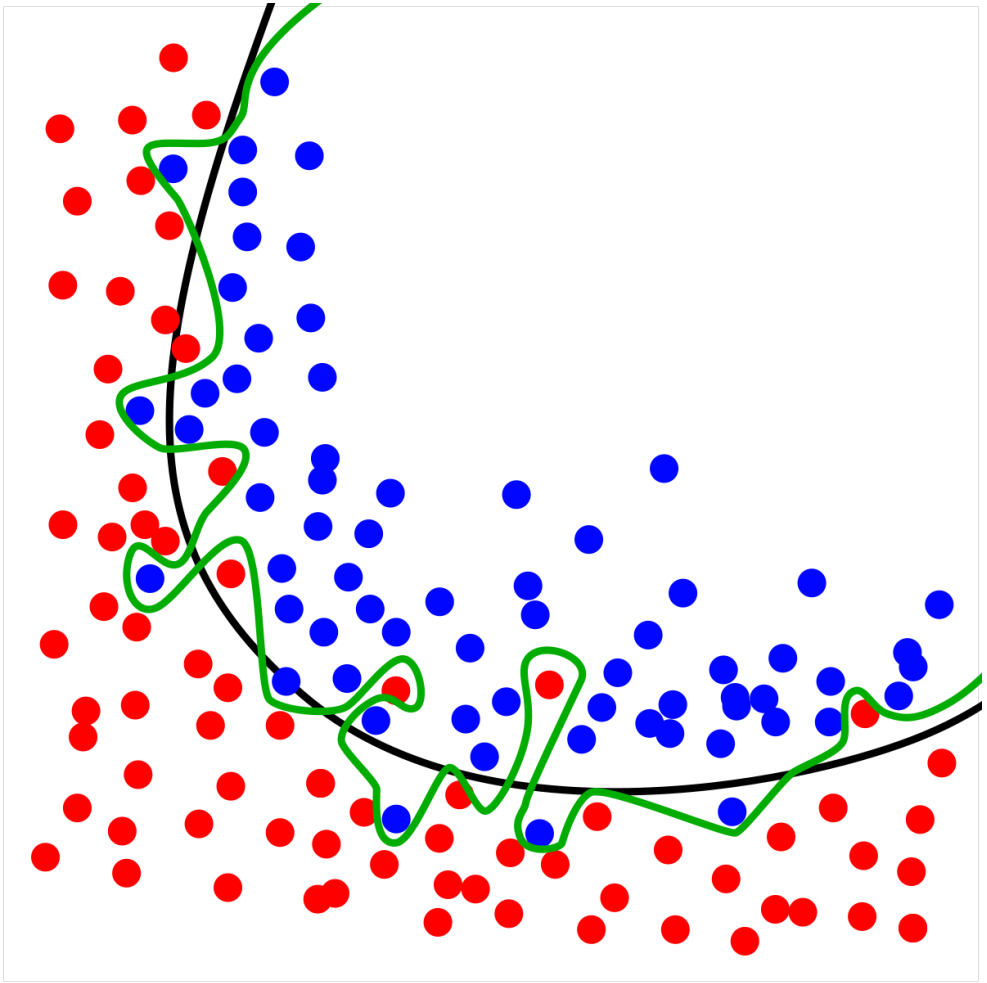
$$Entropy(S[p_1, p_2 \ , \ldots, p_k]) = -\sum_1^k p_i \log(p_i)$$

$$Gain(S, a) = Entropy(S) \ - \sum_{v \in values(S)} \frac{|S_v|}{|S|} \ Entropy(S_v)$$

where S_v is the subset of S for which attribute a has value v

Overfitting

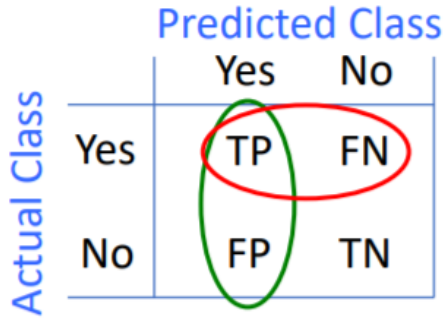
- A decision tree overfits the training data when its accuracy on the training data goes up but its accuracy on unseen data goes down
- Overfitting results in models that are more complex than necessary: after learning knowledge they "tend to learn noise"
- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records



Performance Measures

Confusion Matrix

- Given a dataset of P positive instances and N negative instances:



$$\text{accuracy} = \frac{TP + TN}{P + N}$$

- Imagine using classifier to identify positive cases (i.e., for information retrieval)

$$\text{precision} = \frac{TP}{TP + FP}$$

Probability that a randomly selected result is relevant

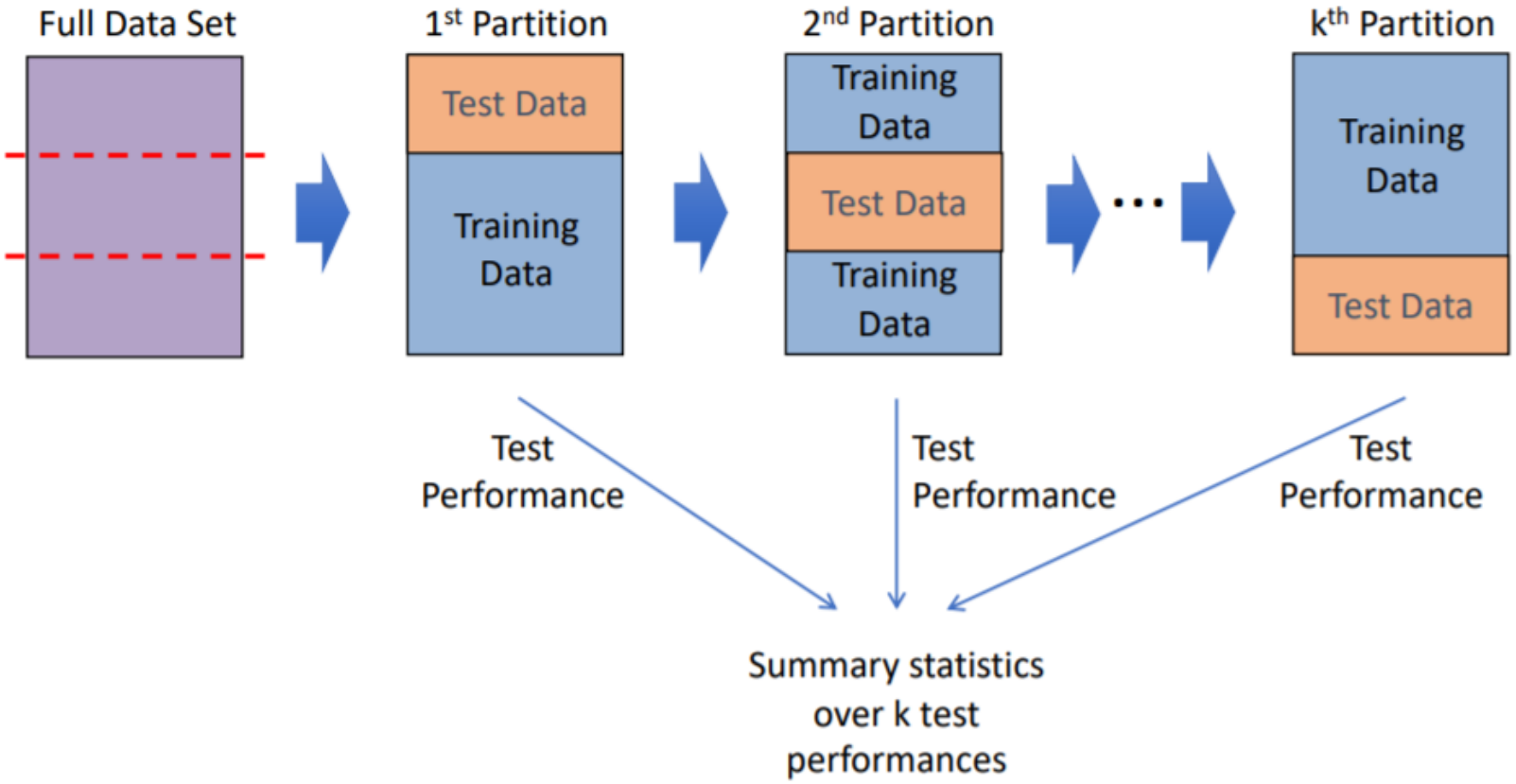
$$\text{recall} = \frac{TP}{TP + FN}$$

Probability that a randomly selected relevant document is retrieved

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Cross Validation

- What's it for?
  - model Choosing, parameter tuning…
- How does it work?



Scikit Learn

- Read the documentation!
  - http://scikit-learn.org/stable/
- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
- http://scikit-learn.org/0.16/modules/generated/sklearn.cross_validation.train_test_split.html

HW1 Q&A