Decision trees

- Build a good tree using information gain
- Dealing with continuous values
- Correctness – Precision, Recall

Overfitting

Generative models

Toss a coin N times. Heads comes H times

Estimate $p_H = \frac{H}{N}$

Why?

Assumption: Coin tosses are generated by some $\hat{p}_H$

Given $\hat{p}_H$ — for each $M \leq N$, compute probability of seeing M heads in N tosses

My estimate $p_H = \frac{H}{N}$ maximizes probability of H heads out of N

Maximum Likelihood Estimator

Overfitting

Performs well on training

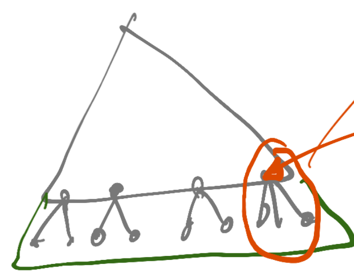
Performs worse on "new" data than some other model

Decision tree

Overfitting = Asking too many questions

Prefer shallower trees

Grow full tree



Compare "error rate" if this leaf node is not expanded

Prune the tree

Now, this description does violence to statistical notions of sampling and confidence limits, so the reasoning should be taken with a large grain of salt. Like many heuristics with questionable underpinnings, however, the estimates that it produces seem frequently to yield acceptable results.

```
physician fee freeze = n: democrat (168/2.6)
physician fee freeze = y: republican (123/13.9)
physician fee freeze = u:
|   mx missile = n: democrat (3/1.1)
|   mx missile = y: democrat (4/2.2)
|   mx missile = u: republican (2/1)
```

```
physician fee freeze = n:
|   adoption of the budget resolution = y: democrat (151)
|   adoption of the budget resolution = u: democrat (1)
|   adoption of the budget resolution = n:
|   |   education spending = n: democrat (6)
|   |   education spending = y: democrat (9)
|   |   education spending = u: republican (1)
physician fee freeze = y:
|   synfuels corporation cutback = n: republican (97/3)
|   synfuels corporation cutback = u: republican (4)
|   synfuels corporation cutback = y:
|   |   duty free exports = y: democrat (2)
|   |   duty free exports = u: republican (1)
|   |   duty free exports = n:
|   |   |   education spending = n: democrat (5/2)
|   |   |   education spending = y: republican (13/2)
|   |   |   education spending = u: democrat (1)
physician fee freeze = u:
|   water project cost sharing = n: democrat (0)
|   water project cost sharing = y: democrat (4)
|   water project cost sharing = u:
|   |   mx missile = n: republican (0)
|   |   mx missile = y: democrat (3/1)
|   |   mx missile = u: republican (2)
```

Dealing with "weak" classifiers

- Build a number of models
- Take majority answer   Voting

Ensemble classifiers

Bagging

Start with training set of size N

Pick N samples with uniform probability
with replacement

Prove that the expected number of distinct items is $\simeq 0.6N$

Build a model

Repeat $2n+1$ times, Vote

Random Forest TM

K attributes overall

At each node pick M of K at random

Explore only these M to choose next attribute to explore

Boosting

Initially all training data has equal "weight" : $w_i = \frac{1}{N}$

Build model M_1

M_1 makes errors on some inputs

Increase weight of erroneous inputs

Build M_2 — recompute weights for wrong inputs

⋮

Unsupervised

Clustering

Market Basket Analysis

Given a list of shopping baskets

Assume a fixed size

Set of Items I

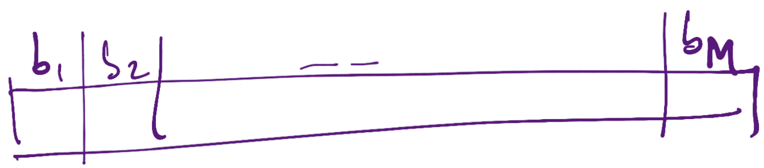
List of baskets B

$0 < t < 1$
= threshold

Identify $X \subseteq I$ s.t. X occurs in $\geq tB$

$|I| = N$ no. of possible baskets is large

Too large to count naively



How many such frequent subsets can there be?

How many individual items are frequent

10^6 items $= |I|$

10^8 baskets

Suppose t is 0.01 (i.e. 1%)

Each frequent item appears 10^6 times

Across 10^8 baskets — 10^9 items occur

At most $\dfrac{10^9}{10^6} = 10^3 = 1000$ items can be frequent

Observation

If $\{i_1, i_2\}$ is frequent,

then $\{i_1\}$, $\{i_2\}$ must be frequent

$\therefore$ if $\{i_1\}$ is not frequent, no subset involving i_1 can be frequent

If $\{i_1, i_2, i_3\}$ is frequent

$\{i_1\}$, $\{i_2\}$, $\{i_3\}$ is frequent

also $\{i_1, i_2\}$, $\{i_1, i_3\}$, $\{i_2, i_3\}$

"A Priori" observation

Algorithm

Compute F_1, frequent items
└ frequent sets of size 1

C_2 - candidates for F_2
└ $\{(i,j) \mid i \neq j,\ i,j \in F_1\}$

C_3 - candidates for F_3
$\{(i,j,k) \mid \{i,j\}, \{j,k\}, \{i,k\} \in F_2\}$

C_k - all k-size sets with all $k-1$ subsets in F_{k-1}

Tedious

Better strategy

Fix an enumeration of I: $i_1 < i_2 < \dots < i_N$

Write each set in ascending order

$$\left.\begin{array}{l}\{i_1, i_2, i_3, \dots, i_{k-1}\} \\ \{j_1, j_2, j_3, \dots, j_{k-1}\}\end{array}\right\} \in F_{k-1}$$