Decision Trees

Prof. Ankur Sinha

Indian Institute of Management Ahmedabad

Gujarat India

Decision Trees

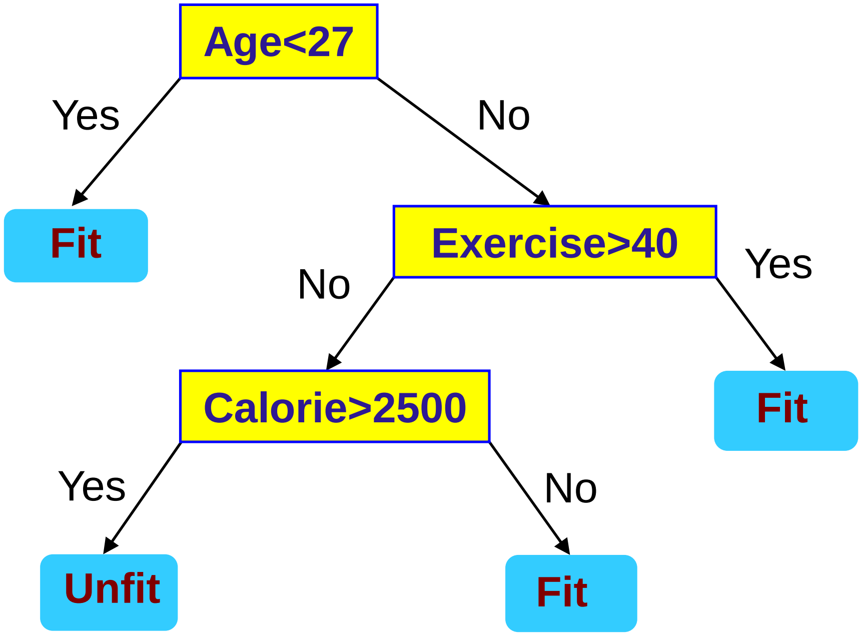
- Used for classification
  - Legitimate or fraudulent credit card transactions
  - Grant a loan or not
  - Tumor is benign or not
  - News item is on Finance, Politics, Sports, or Arts

Example

Person	Calorie Intake	Exercise Duration	Age	Fit (Yes/No)
Person 1	2089	20	47	0
Person 2	2569	54	23	1
Person 3	2790	58	28	1
Person 4	1882	20	41	1
Person 5	2160	55	20	1
Person 6	2408	22	29	1
Person 7	2740	44	25	1
Person 8	2700	8	29	0
Person 9	2635	52	33	1
Person 10	1918	22	40	1

Example

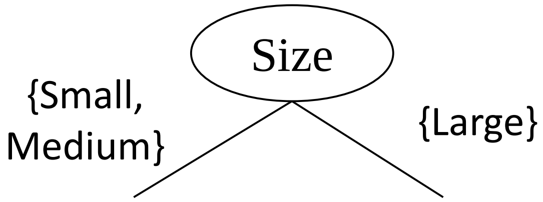
Person	Calorie Intake	Exercise Duration	Age	Fit (Yes/No)
Person 1	2089	20	47	0
Person 2	2569	54	23	1
Person 3	2790	58	28	1
Person 4	1882	20	41	1
Person 5	2160	55	20	1
Person 6	2408	22	29	1
Person 7	2740	44	25	1
Person 8	2700	8	29	0
Person 9	2635	52	33	1
Person 10	1918	22	40	1

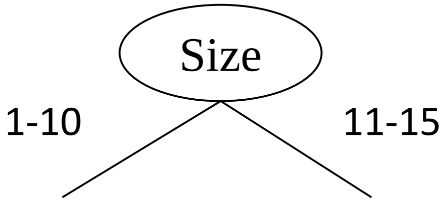


Which attribute to choose at each node?
How to split the attribute?
What is the depth of the tree?

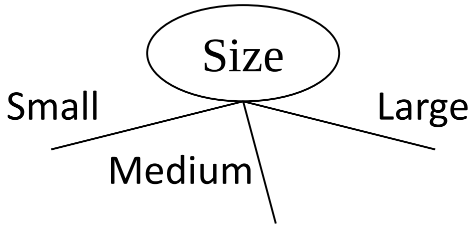
Decision Trees

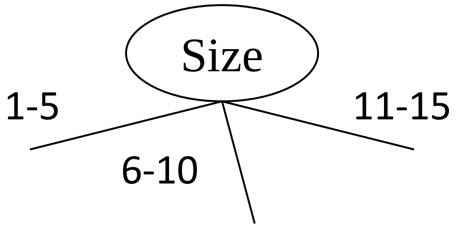
- Binary Split





- Multiway Split





Gini Index

- Gini index measures impurity
- Used in Classification and Regression Tree (CART) algorithm

$$Gini(t) = 1 - \sum_{j} p_j^2$$

At node t

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Parent node is split into k partitions

Number of objects in partition i is n_i

Example

Person	Calorie Intake	Exercise Duration	Age	Fit (Yes/No)
Person 1	2089	20	47	0
Person 2	2569	54	23	1
Person 3	2790	58	28	1
Person 4	1882	20	41	1
Person 5	2160	55	20	1
Person 6	2408	22	29	1
Person 7	2740	44	25	1
Person 8	2700	8	29	0
Person 9	2635	52	33	1
Person 10	1918	22	40	1
Person 11	2218	41	59	1
Person 12	2461	36	48	0
Person 13	2057	49	26	1
Person 14	2394	19	39	0
Person 15	2319	53	38	1
Person 16	2190	23	43	0
Person 17	2589	11	18	0
Person 18	2640	29	57	0
Person 19	2508	59	55	1
Person 20	2419	38	28	1
Person 21	2998	10	57	0
Person 22	2155	50	36	1
Person 23	1959	16	26	1
Person 24	1904	24	45	1
Person 25	1980	42	37	1
Person 26	1937	55	30	1
Person 27	2433	4	32	0
Person 28	2773	1	27	0
Person 29	1914	58	25	1
Person 30	1913	30	37	1

Find the Gini Index for the data
$=1-(10/30)^2-(20/30)^2$

Find out the best criterion to split on such that the purity increases

Entropy

- Entropy measures impurity
- Information gain, Used in ID3 (Iterative Dichotomiser) algorithm, refers to difference between entropy before the split and average entropy after the split

$$Entropy(t) = -\sum_{j} p_j \log_2 p_j$$

At node t

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent node p is split into k partitions
Number of objects in partition i is n_i

Entropy

- Gain ratio, which is adjusted information gain is used by C4.5, an improvement of ID3

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

$$Entropy(t) = -\sum_{j} p_j \log_2 p_j$$

At node t

Parent node p is split into k partitions
Number of objects in partition i is n_i

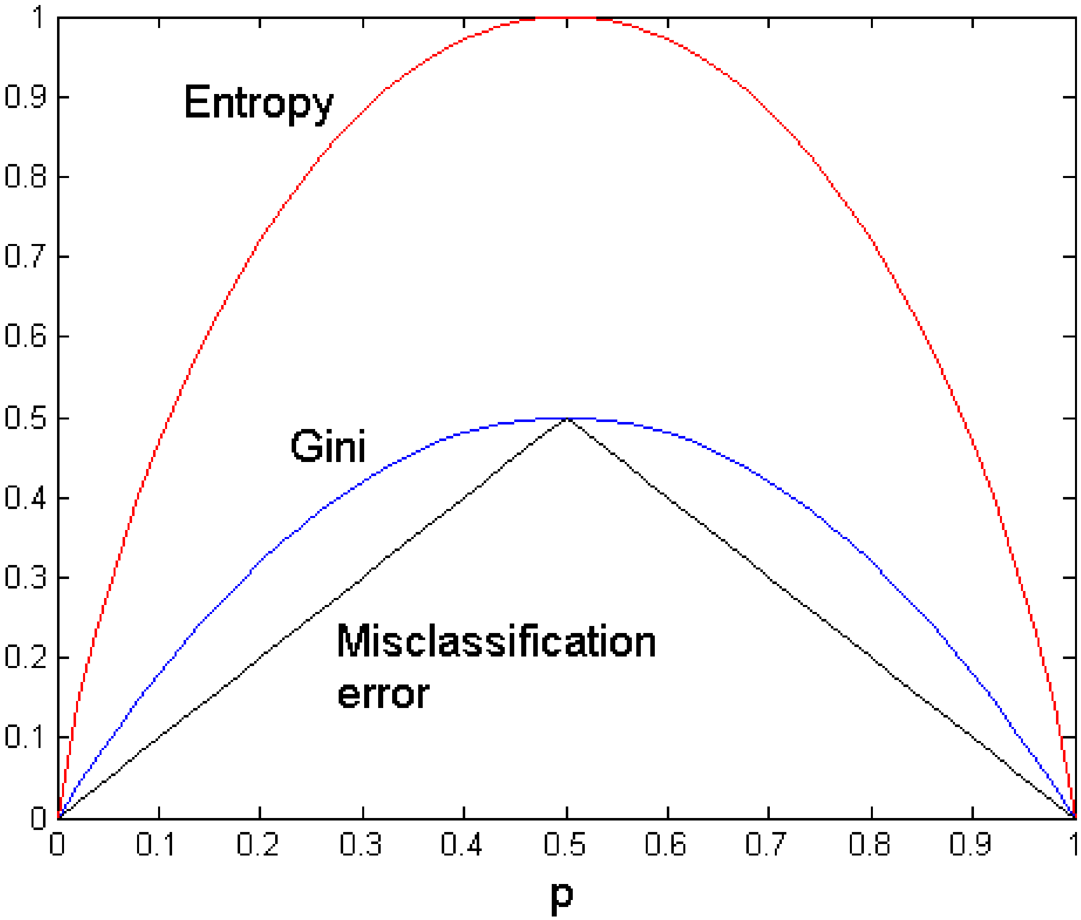
Classification Error

- Classification error measure impurity

$$\boxed{Error(t) = 1 - \max_{j} p_j}$$

At node t

Comparing Different Criteria



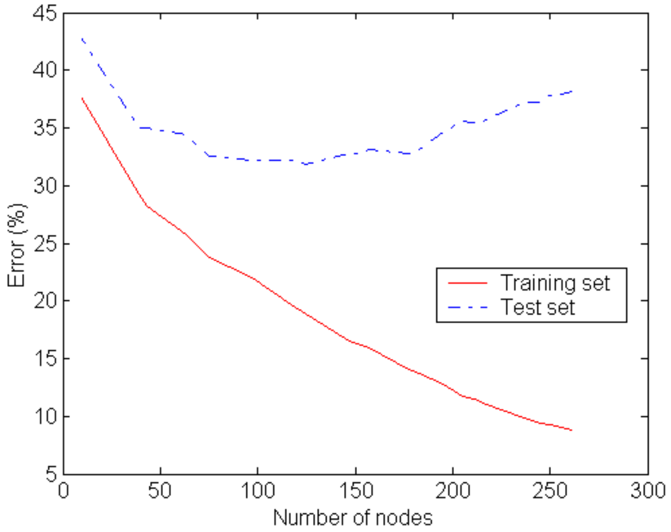
A two class problem with $p_1=p$ and $p_2=1-p$

Occam’s Razor

- A deep decision tree can fit almost any data
- Occam’s razor says that between two models of similar generalization errors, one should prefer the model which is simple

Addressing Overfitting

- Pre-pruning
  - Stop the algorithm when the tree becomes large
- Post-pruning
  - Trim the nodes in the bottom-up manner



Thank you