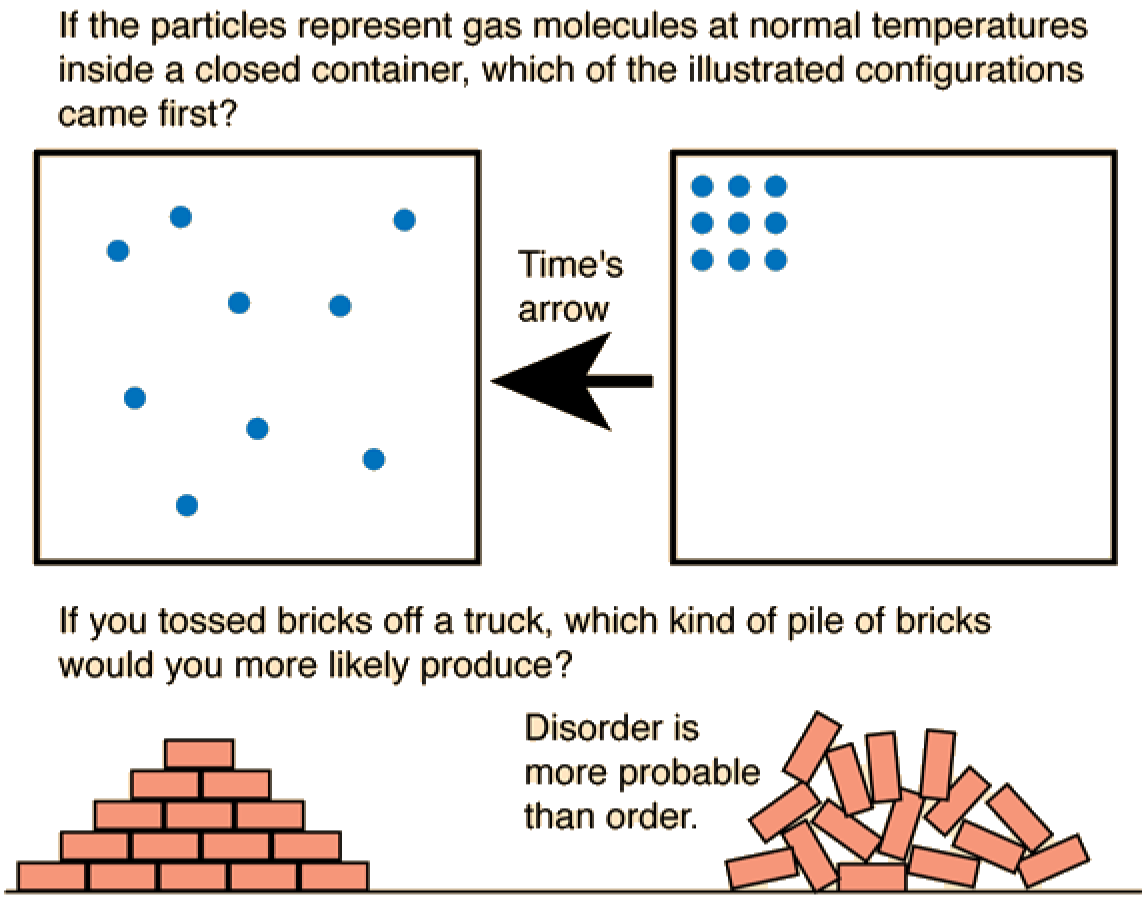
Decision Trees – Part 4

Other Impurity Measures

Entropy

- In thermodynamics, the measure of disorder in a system.
- The second law of thermodynamics states that the entropy of an isolated system never decreases over time.
- Systems evolve towards thermodynamic equilibrium, the state with maximum entropy.
- Put two gasses of roughly the same weight into a container and they will tend to mix.

Entropy


If the particles represent gas molecules at normal temperatures inside a closed container, which of the illustrated configurations came first?
Time's arrow
If you tossed bricks off a truck, which kind of pile of bricks would you more likely produce?
Disorder is more probable than order.

Entropy – Decision Trees

$$Entropy(n) = -\sum_{c} p(c|n) \log p(c|n)$$

- Again P(c|n) is the probability of an instance in node n being of class c.
- And this is also the relative frequency of class c in node n.
- (Log to the base 2 is used)

Entropy

- Maximum value is log n_c when records are equally distributed among all classes ($\log_2 2 = 1$)
- Minimum is 0 when all records belong to one class, implying most information

Entropy As a Measure of Information Gain

- Entropy was applied to data communication by Claude Shannon.
- Considering information and noise in a signal.
- Low probability events corresponds to information.
- [Roughly speaking, in a decision tree, homogenous nodes are unlikely to happen at random.
- And when they do they provide us with information.]

Information Gain

- This is defined as the entropy of the parent node minus the weighted sum of the entropy values of the child nodes.
- If child nodes are ‘purer’ than parent nodes then we have gained information.

Entropy

C1	**0**
C2	**6**

P(C1) = 0/6 = 0   P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6   P(C2) = 5/6

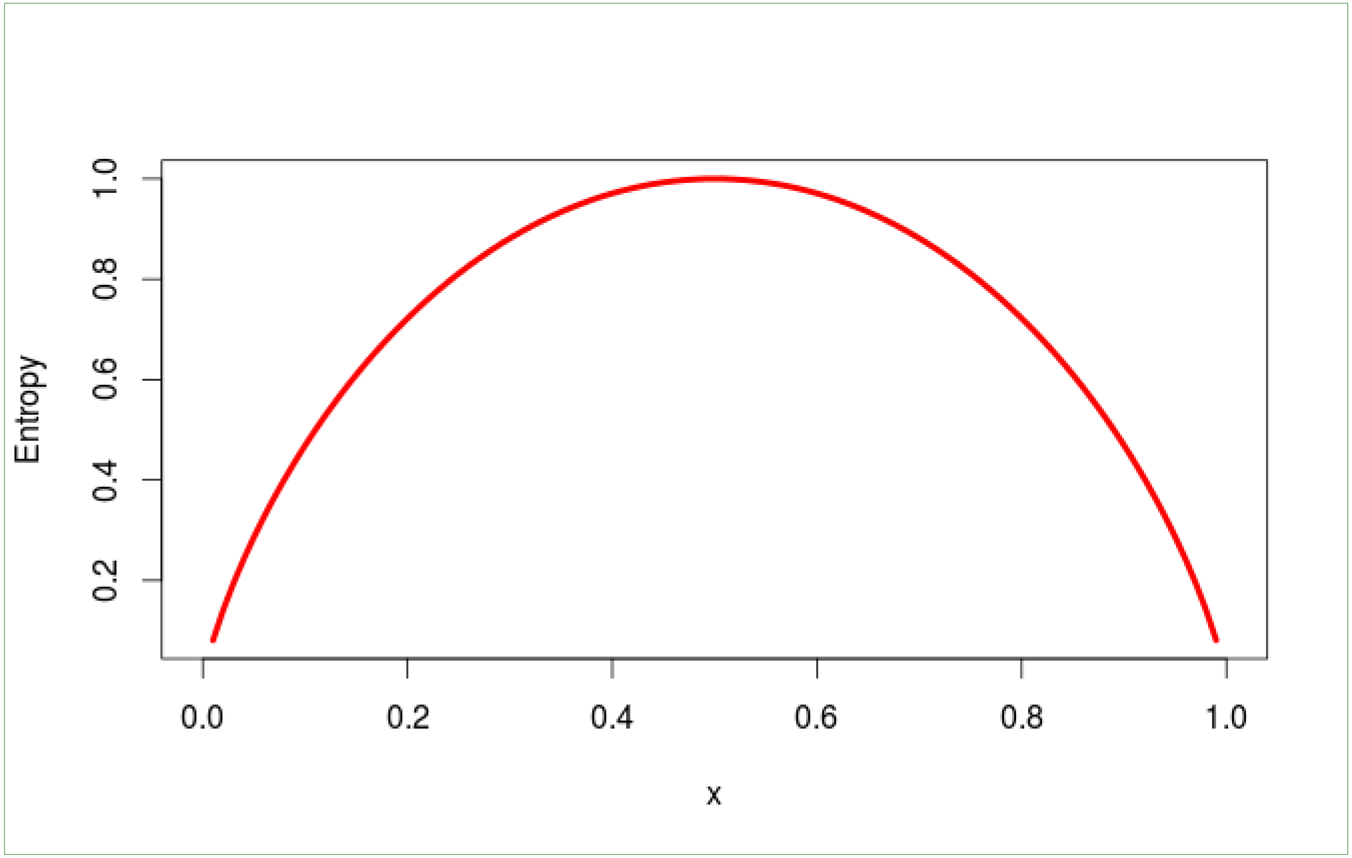
Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6   P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Entropy for 2 classes



Classification Error

$$\boxed{Error(n) = 1 - \max_{c} P(c|n)}$$

- Classification error at a node t :
- Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes
- Minimum 0 when all records belong to one class, implying most interesting information

Classification Error

$$Error(n) = 1 - \max_{c} P(c|n)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

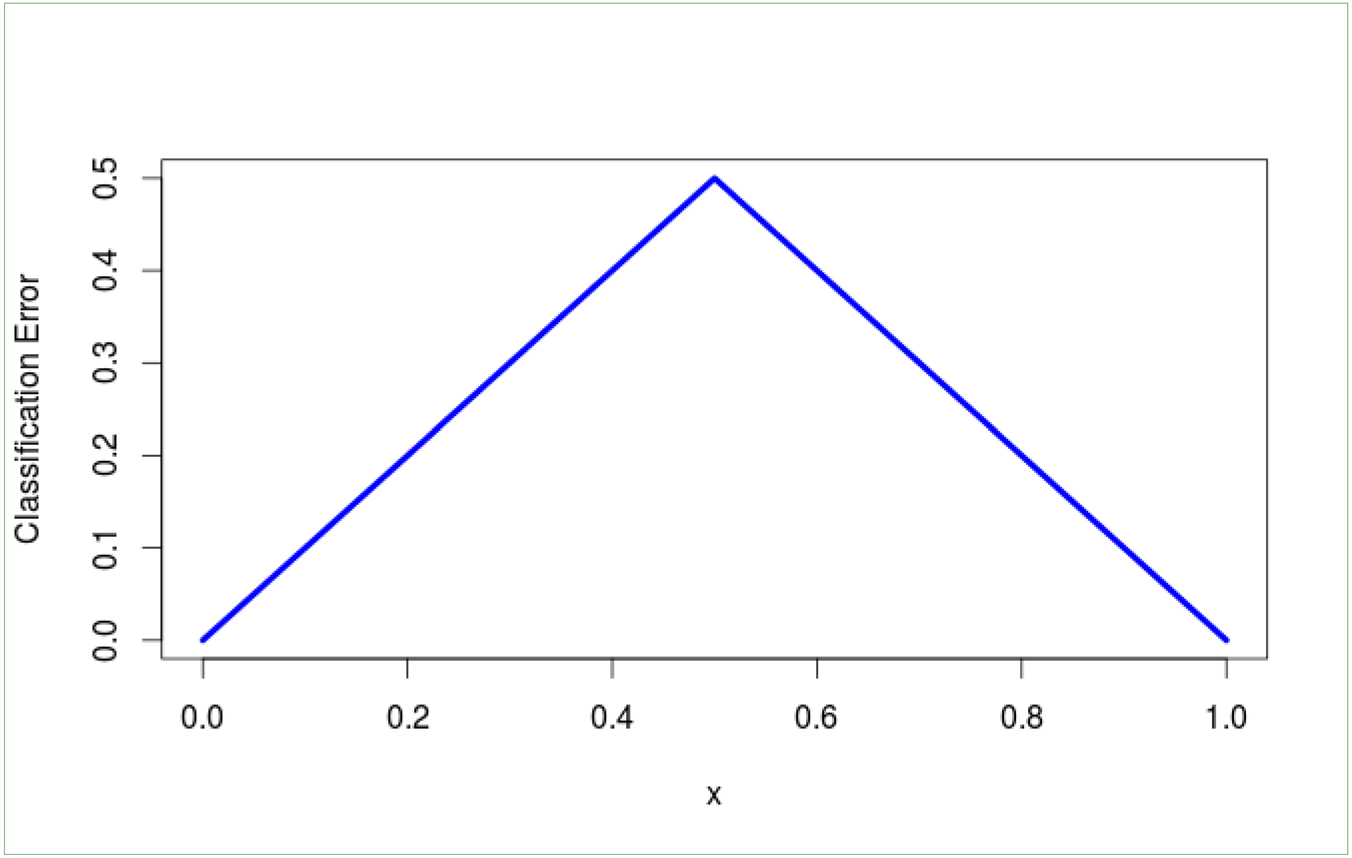
Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

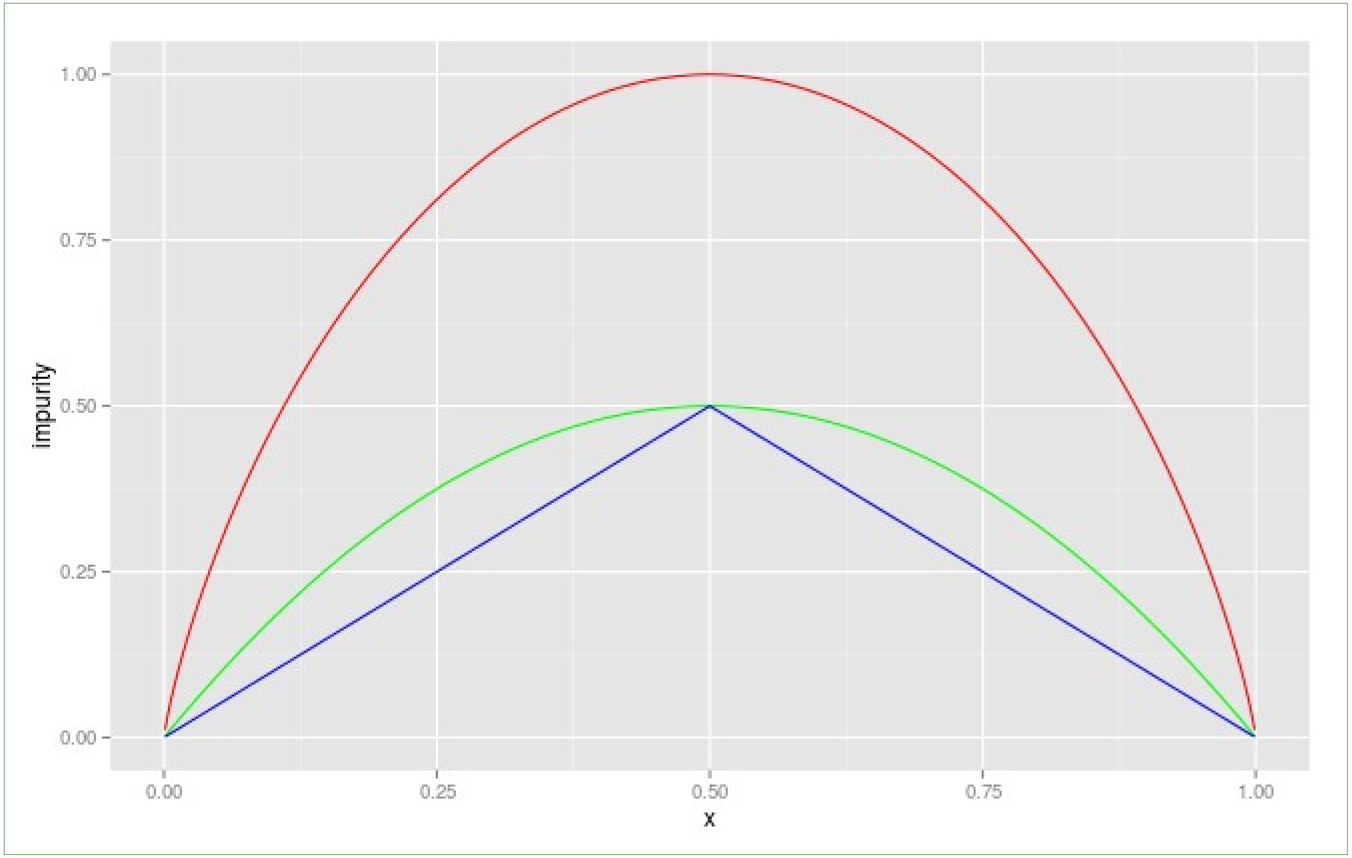
Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Classification Error for 2 classes



Comparison of Impurity Measures

For a 2-class problem:



Recap

- Three measures of node impurity.
- Different measures in different algorithms
- GINI
  - CART
- Information Gain (Entropy)
  - ID3, C4.5, C5.0
- Misclassification Error