Applied Machine Learning

Classification - Random Forests - Splits

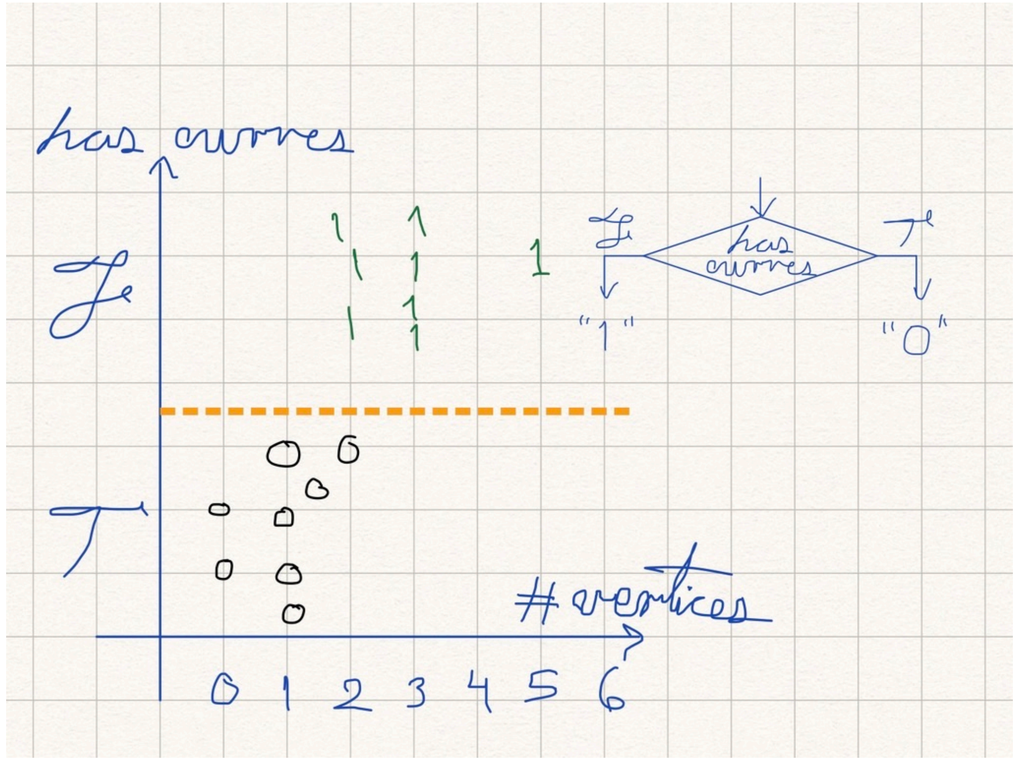
Random Forests - Splits

- Decision Trees and Random Forests
- Entropy
- Information Gain
- Dealing with missing values

Decision Trees

Decision Tree

- Each node is a test on some input feature
- The result of the test indicates what branch to take
- Each leave represents the resulting class



Decision Trees - Construction

- DecisionTreeExpand (branch, dataset)

  stop when

    depth(branch_node) >= max_depth

    size(dataset) <= min_leave_size

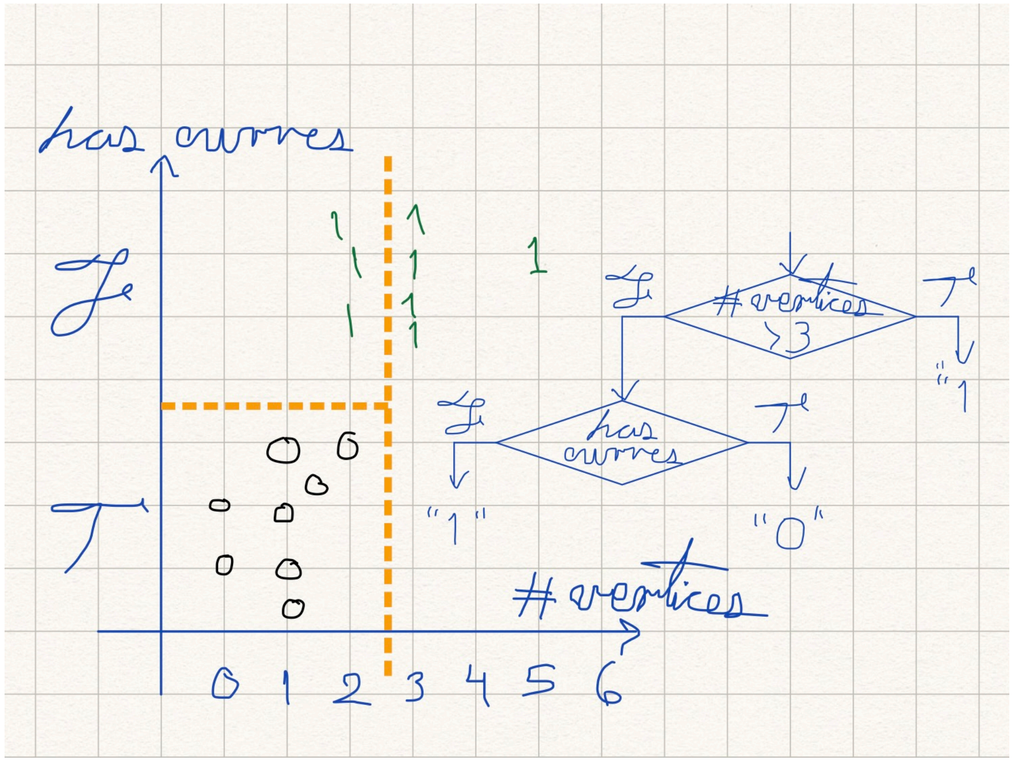
    all elements in dataset in same class

  (subset_l, subset_r, test) = best_split(dataset)

  (child_r,child_r) = new_branch(branch, test, split_l, split_r)

  DecisionTreeExpand(child_l,subset_r)

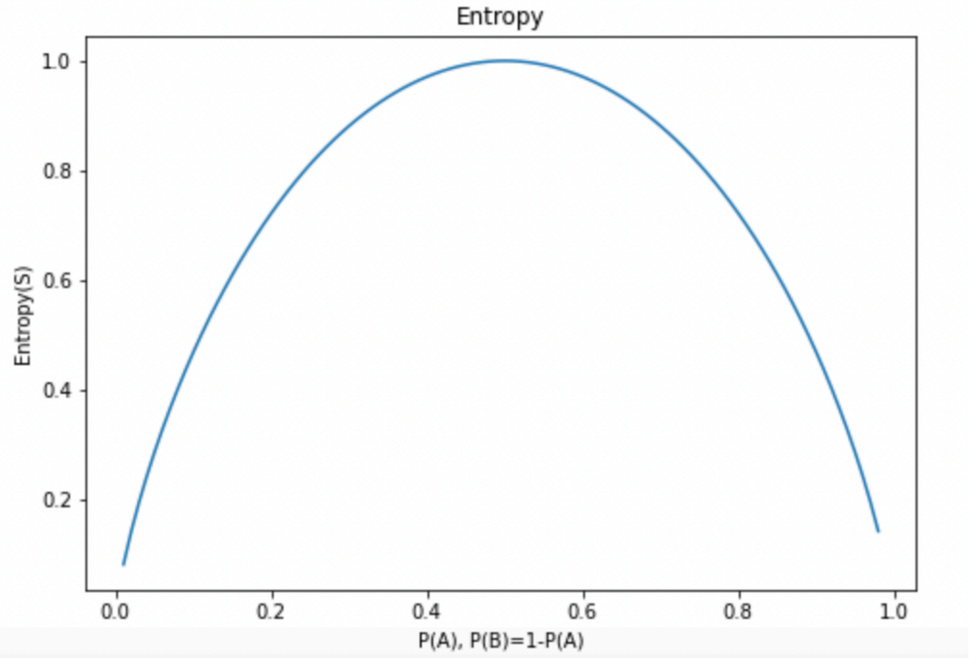
  DecisionTreeExpand(child_r,subset_r)



Entropy - 2 classes

$\|A\|$	$\|B\|$	Entropy(S)
88	12	0.53
0	100	0.00
1	99	0.08
12	88	0.53
25	75	0.81
49	51	0.99
50	50	1.00

- Diversity of dataset $S = A \cup B$
  - one subset per class
- $N = |A| + |B|$
- $P(A) = \frac{|A|}{N}, P(B) = \frac{|B|}{N}$
- Entropy$(S) = -P(A)\log_2 P(A) - P(B)\log_2 P(B)$



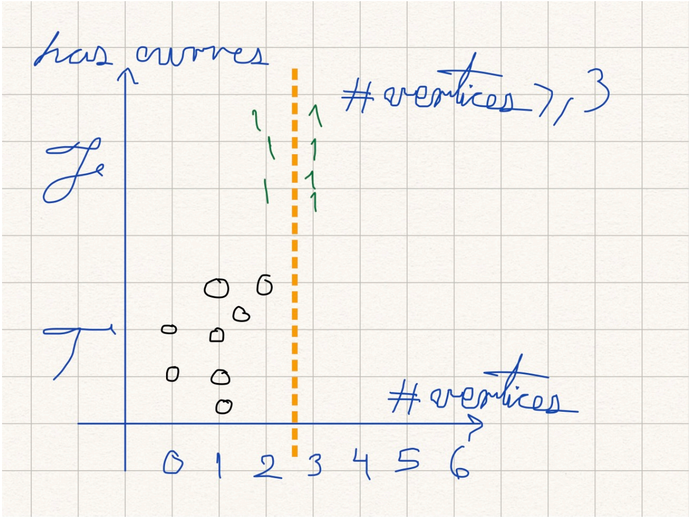
Entropy - C classes

More general case:

- The elements of S may belong to C different classes
- Each class i with probability P_i
- $\text{Entropy}(S) = -\sum_{i=1}^{C} P_i \log_2 P_i$

Information Gain

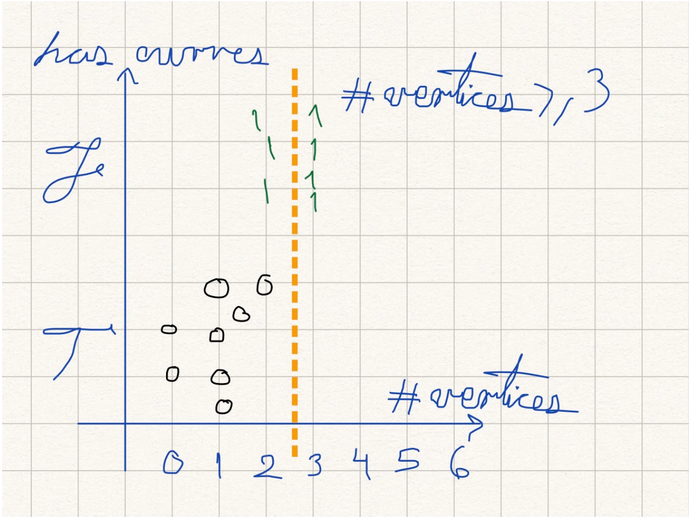
- Applying test on values of feature $x^{(i)}$ in set S results in subsets S_l and the S_r
  - $P_{S_l} = \frac{|S_l|}{|S|}, P_{S_r} = \frac{|S_r|}{|S|}$
  - $\text{Gain}(S; S_l, S_r) = \text{Entropy}(S) - (\frac{|S_l|}{|S|}\text{Entropy}(S_l) + \frac{|S_l|}{|S|}\text{Entropy}(S_l))$
- If the split is in n subsets:
  - $\text{Gain}(S; S_1, \ldots, S_n) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}\text{Entropy}(S_i)$

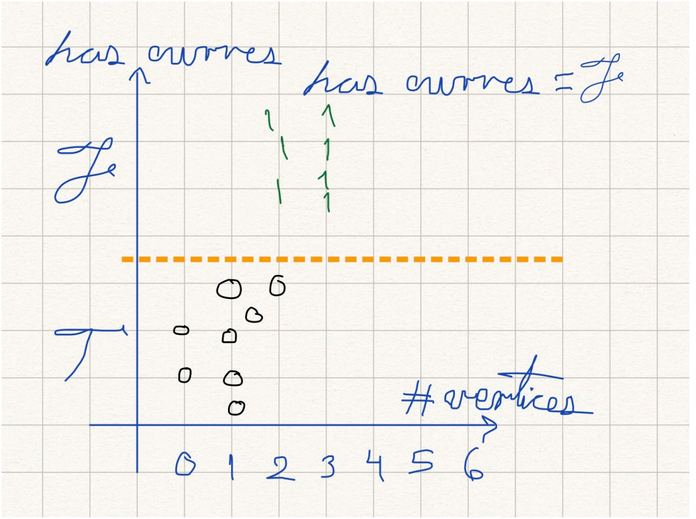


Information Gain

$$\text{Gain}(S; S_1, \ldots, S_n) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy}(S_i)$$

- Examples
  - $\text{Entropy}(S) = 0.996$
    - $\text{Gain}(S; S_{left}, S_{right}) = 0.99 - (\frac{11}{15} * 0.84 + \frac{4}{15} * 0.00) \approx 0.37$
    - $\text{Gain}(S; S_{up}, S_{down}) = 0.99 - (\frac{7}{15} * 0.00 + \frac{8}{15} * 0.00) \approx 0.99$





Dealing with missing values

- In splits, if an item misses the feature value that decide where it goes
  - Estimate it based on other examples
    - mode or mean
  - Consider only the examples in the corresponding branch

Random Forests - Splits

- Decision Trees and Random Forests
- Entropy
- Information Gain
- Dealing with missing values

Applied Machine Learning

Classification - Random Forests - Splits