Data Mining and Machine Learning
ID3 and Regression

Gergely Horváth

September 29, 2022

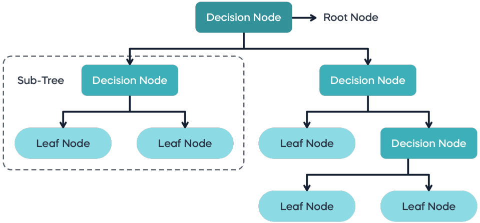
Outline

1. Decision trees
   - ID3
   - ID3
2. Regression

Decision trees

Classification and regression trees (categorical & numerical data handling)

- Splits dataset into small subsets
- Final result: tree with
  - root node
  - decision nodes: branches → possible values for the attribute
  - leaf nodes: represents a classification
- Which feature splits the data better (which is the best attribute)?



ID3 – Iterative Dichotomiser 3

- Core algorithm for building decision trees
  - top-down, greedy search to test each attribute at every node of the tree
- Which is the best attribute?
  - the one which will result in the smallest tree
  - choose the attribute that produces the "purest" nodes
  - information gain (IG)
    - information before splitting – information after splitting
    - is used to construct a tree
- Entropy: a measure of randomness
  - unbiased coin toss (head and tail is equally likely): $E = 1$
  - biased (2 head): $E = 0$
  - ID3 uses entropy to calculate the homogeneity of a sample
  - $E(p_1, p_2, ..., p_N) = -p_1 log(p_1) - p_2 log(p_2) - ... - p_N log(p_N)$

ID3 – Iterative Dichotomiser 3

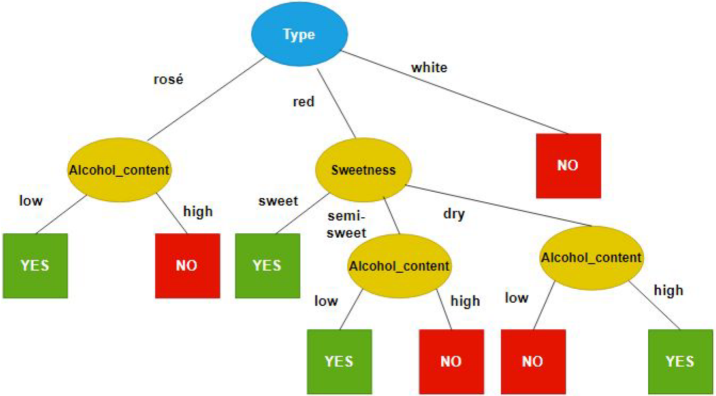
General recipe:

1. Compute the overall entropy of the class distribution
2. Choose a feature and compute the entropy of the class distribution for each unique value
3. Calculate a weighted sum of the unique values' entropy (weight is its relative overall presence in the examined dataset)
4. Subtract this value from the overall entropy to receive the information gain
5. Repeat the previous 3 steps for every attribute and choose the one with the highest information gain
6. Make the first split with the best performing attribute
7. Split the data with respect to the chosen feature
8. Where the subsets are uniform in class values, that class value should be assigned to that node which will make it a leaf node, otherwise, the previous steps should be repeated for the assigned subsets

ID3 – Iterative Dichotomiser 3

Alcohol_content	Sweetness	Type	(Year)	Popular
low	sweet	rosé	2012	yes
low	dry	red	2009	no
low	semi-sweet	red	2008	yes
high	sweet	rosé	2013	no
low	dry	white	2013	no
low	sweet	white	2006	no
high	semi-sweet	red	2011	no
high	sweet	red	2007	yes
high	dry	red	2005	yes

ID3 – Iterative Dichotomiser 3



Regression – I.

The "reality":

$$y = \sum_{i}^{n} (p_i \cdot x_i) + \epsilon$$

The model:

$$\hat{y} = \sum_{i}^{n} (p_i \cdot x_i)$$

Let us have matrix notations:

$$Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ . \\ . \\ y^{(m)} \end{bmatrix} \qquad X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ . \\ . \\ x^{(m)} \end{bmatrix}$$

Where $Y \in \mathbb{R}^{m \times 1}$, $X \in \mathbb{R}^{m \times n}$ and $p \in \mathbb{R}^{n \times 1}$.

Regression – II.

Notation repetition:

$$Y = X \cdot p + \epsilon$$
$$\hat{Y} = X \cdot p$$

The next step is to define the error made by the model (in this case the – arguably – simplest convex function will be used):

$$L(D,\ p) = \left\| Y - \hat{Y} \right\|_2 = \| Y - (X \cdot p) \|_2$$

The error is expressed with a 2-norm, so the next step is to minimize the loss, to find the global minimum of this function. → This is going to be easy and trivial since we have a convex function.

Logistic regression will be a model, where our linear regression model is being fed to a sigmoid function:

$$sigmoid(model) = \frac{1}{1 + e^{-model}}$$