Introduction to Machine Learning
Decision Trees

Nikolay Manchev

6 December 2016

London Machine Learning Study Group

Housekeeping

Next events

http://www.meetup.com/London-Machine-Learning-Study-Group

Follow me

https://twitter.com/nikolaymanchev

Slides and code

Available at https://github.com/nmanchev/MachineLearningStudyGroup

Machine Learning Models

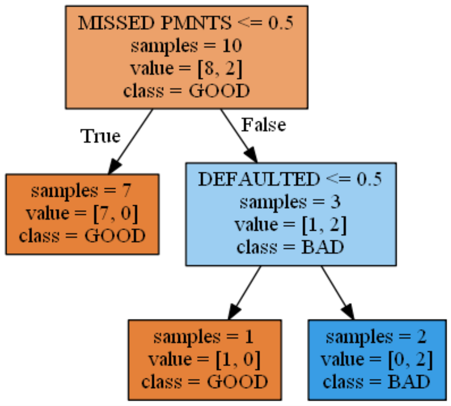
Flach talks about three types of Machine Learning models [Fla12]

- Geometric models
- Logical models
- Statistical models

Decision Trees

Decision Tree Example

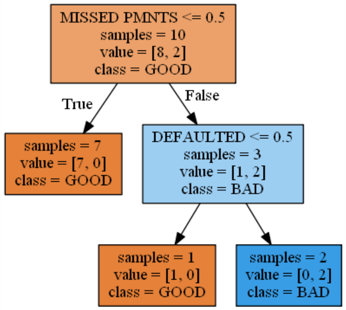
< 2 years at current job	missed payments	default	rating
N	N	N	GOOD
Y	N	Y	GOOD
N	N	N	GOOD
N	N	N	GOOD
N	Y	Y	BAD
Y	N	N	GOOD
N	Y	N	GOOD
Y	Y	Y	BAD
Y	N	N	GOOD
Y	N	N	GOOD



Credit rating example. Based on [Lew07]

Terminology

- **Root Node** – That is where the tree starts
- **Sub-tree** – A sub-section of the tree
- **Splitting** – The process of dividing a node into sub-nodes
- **Decision Node** – Node that splits into sub-nodes
- **Terminal Node** – Node that does not split



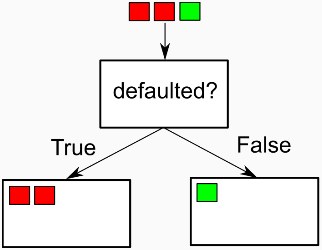
Algorithm

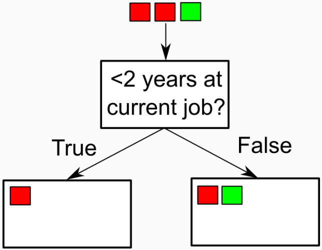
Algorithm 1 Pseudocode for training a Decision tree

function $\mathrm{TREE}(examples, attributes, default\_value)$
  if $examples$ is empty **then return** a node with $default\_value$
  if all $examples$ have the same class **then return** a node with that class
  if $attributes$ is empty **then return** MODE($examples$)
  $attr \leftarrow$ SELECT_BEST($examples$, $attributes$)
  for each value v_i of $attr$ **do**
    $examples_i \leftarrow \{examples$ where $attr = v_i\}$
    $subtree \leftarrow$ TREE($examples_i$, $attributes - attr$, MODE($examples_i$))
    add a decision node v_i and $subtree$ to $tree$
  return $tree$

Attribute Selection

$<$ 2 years at current job	missed payments	default	rating
N	N	N	GOOD
Y	N	Y	GOOD
N	N	N	GOOD
N	N	N	GOOD
N	Y	Y	BAD
Y	N	N	GOOD
N	Y	N	GOOD
Y	Y	Y	BAD
Y	N	N	GOOD
Y	N	N	GOOD

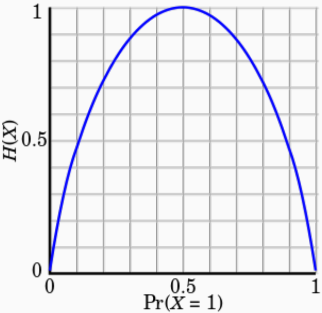




Entropy[1/3]

Measure of uncertainty

- Information theory entropy
  For a discrete random variable $X \in \{\chi_1, \ldots, \chi_n\}$ and probability mass function $\mathsf{P}(X)$
  $$\mathsf{H}(X) = -\sum_{i=1}^{n} \mathsf{P}(x_i)\log_2 \mathsf{P}(x_i)$$
- Example – 10 samples in the credit rating data (8 GOOD, 2 BAD):
  $\mathsf{H}(X) = \mathsf{H}(\frac{8}{10}, \frac{2}{10}) = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} \approx 0.72$

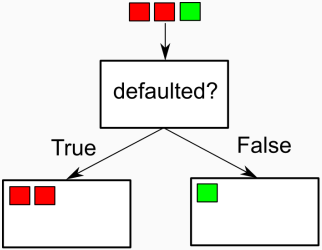


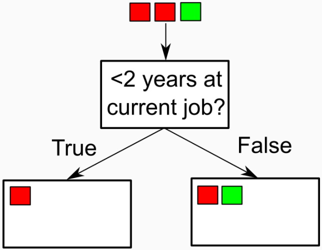
Binary entropy plot [BD]

Entropy[2/3]

Measure of uncertainty

- For the (defaulted? = True) terminal node:
  $\mathrm{H}(\frac{2}{2}) = -\frac{2}{2}\log_2\frac{2}{2} = 0$
- For the (<2 years? = False) terminal node:
  $\mathrm{H}(\frac{1}{2}, \frac{1}{2}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$

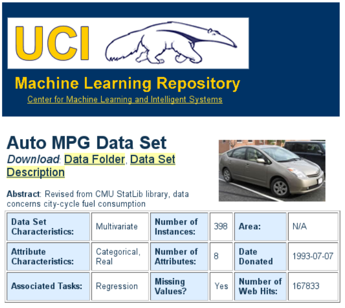




Data Set Example

UCI Machine Learning Repository –
`archive.ics.uci.edu/ml`

- Great resource for Machine Learning data sets
- Over 330 freely available sets
- Auto MPG Data Set
  - Fuel consumption in MPG
  - Attributes: mpg, cylinders, displacement, horsepower, weight, acceleration etc.



Entropy[3/3]

Entropy of a split

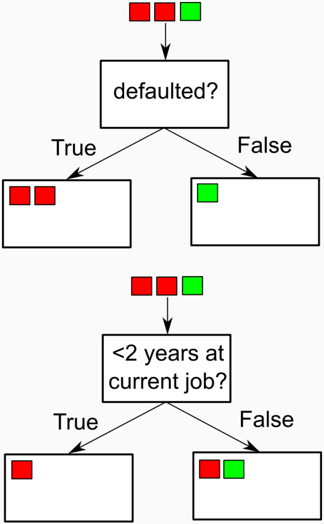
We can compute the weighted average over all sets in the split

$$\mathsf{I}(X, A) = \sum_{i=1}^{m} \frac{|X_i|}{|X|} \times H(X_i)$$

where m is the number of distinct values in A, $|X|$ is the size of X, and $|X_i|$ is the size of X_i

Example:

$$\mathsf{I}(X, \text{defaulted}) = \tfrac{2}{3} \times 0 + \tfrac{1}{3} \times 0 = 0$$
$$\mathsf{I}(X, \text{ <2 years?}) = \tfrac{1}{3} \times 0 + \tfrac{2}{3} \times 1 = \tfrac{2}{3}$$



Selecting the best attribute

Information Gain

Expected decrease of entropy after splitting on an attribute

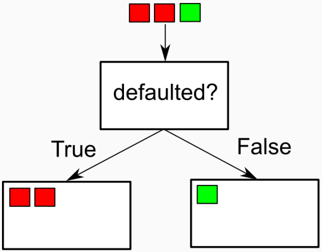
$\mathsf{IG}(X, A) = \mathsf{H}(X) - \mathsf{I}(X, A)$

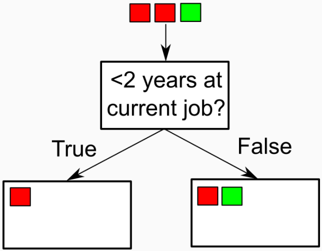
Example:

$\mathsf{H}(X) = \mathsf{H}(\frac{2}{3}, \frac{1}{3}) - \frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.92$

$\mathsf{IG}(X, \text{defaulted}) = \mathsf{H}(\frac{2}{3}, \frac{1}{3}) - \mathsf{I}(X, \text{defaulted}) = 0.92 - 0 = 0$

$\mathsf{I}(X, \text{ <2 years?}) = \mathsf{H}(\frac{2}{3}, \frac{1}{3}) - \mathsf{I}(X, \text{<2 years?}) = 0.92 - \frac{2}{3} = 0.26$





Decision Trees

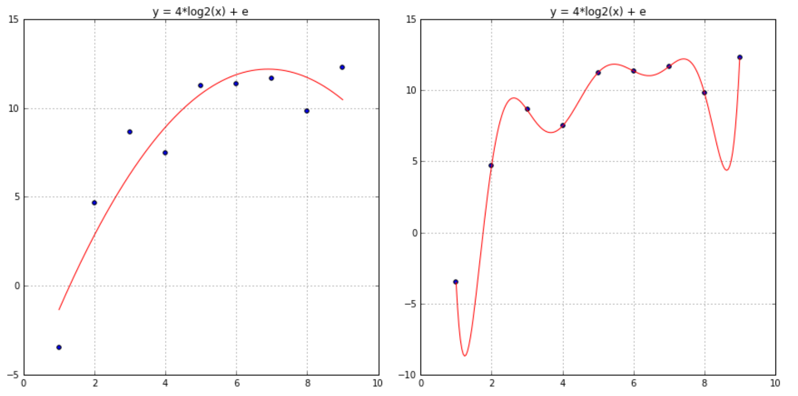
Pros

- Interpretability (easy to understand)
- Mixed data type – numerical and categorical variables in the same model
- Less data preparation

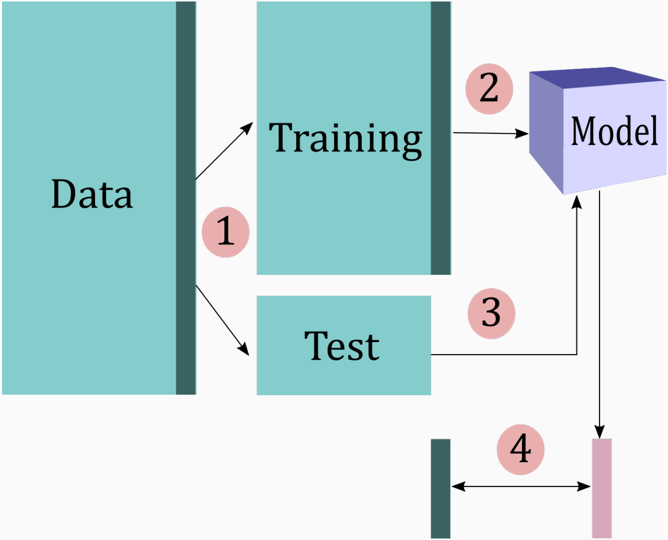
Cons

- They tend to get complex and overfit

Overfitting



Holdout method



Excessive Branching

- An attribute with high cardinality is typically considered a good candidate for a split
    - Worst case would be a unique identifier – splitting on it produces pure nodes (one example per node), but such tree would be useless for prediction
    - One way to avoid this is by using different metrics (e.g. Gain Ratio, which captures the ratio of information gain to the intrinsic information)

Some improvements

Reducing complexity

- Other selection criteria
  - Gain Ratio
  - Chi-Square
- Restriction on splits
  - Binary and multi-way splits
  - Minimum samples to split
- Early stopping
- Pruning
  - Reduced Error Pruning

Decision Trees

Pros

- Interpretability (easy to understand)
- Mixed data type – numerical and categorical variables in the same model
- Less data preparation

Cons

- They tend to get complex and overfit
- Instability
- Inadequate for predicting continuous values

Decision Trees vs. Linear Regression

Decision Trees

- Can solve both classification and regression problems
- If the relationship is non-linear it will outperform Linear Regression
- Can build models that are easy to explain

Linear Regression

- In a linear relationship Linear Regression will likely outperform Decision Trees
- Cannot easily handle categorical variables

References I

- Brona and A. Damato, *Binary entropy plot*.
- Peter Flach, *Machine learning: The art and science of algorithms that make sense of data*, Cambridge University Press, New York, NY, USA, 2012.
- Michael S. Lewicki, *Artificial intelligence: Learning and decision trees*, Carnegie Mellon, 2007.