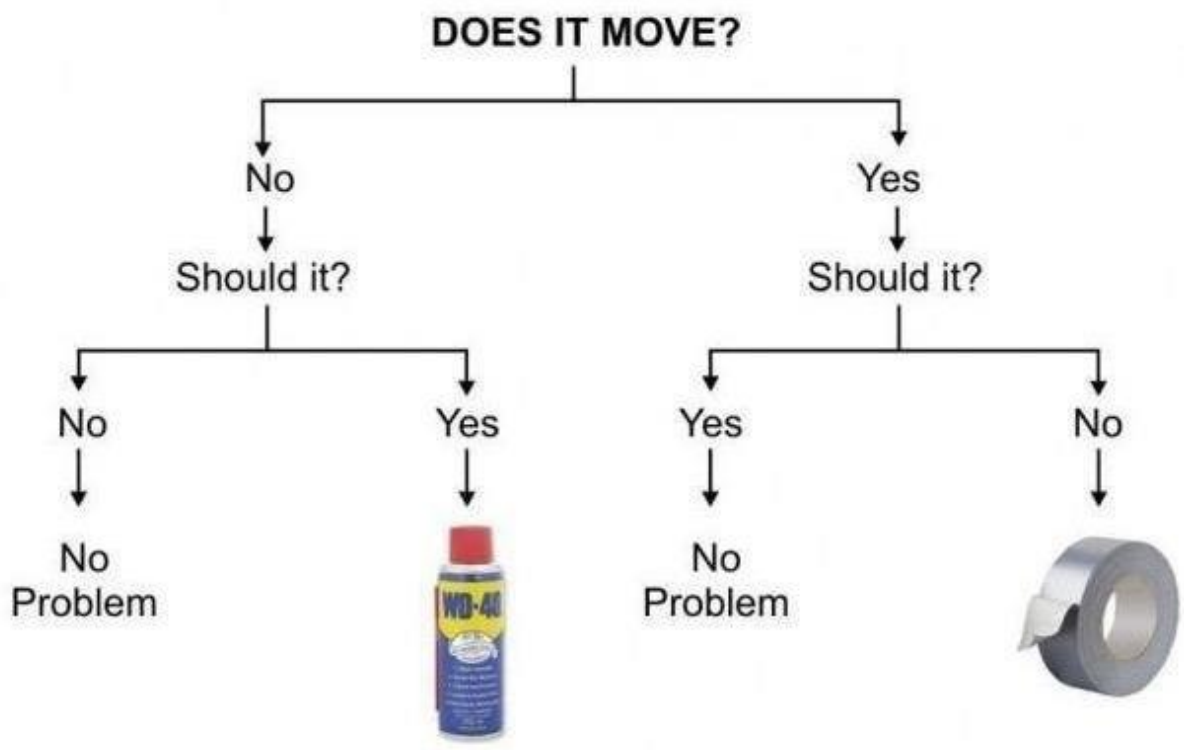
Lecture 5

Classification and Regression Trees (CART)

Decision trees are used in:

- **classification** when the predicted outcome is the class (discrete) to which the data belongs,
- **regression** when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

The decision procedure is based on a sequence of questions, to each of which answer allows to split the domain of possibilities till further splitting is not possible or unnecessary. A decision tree generates a set of rules that follow a "IF Variable A is X THEN…" pattern.



Source: https://www.flickr.com/photos/dullhunk/7214525854

There are many different algorithmic tree structures, method of constructing, and visualisation techniques. Examples of decision tree algorithms:

- ID3 (Iterative Dichotomiser 3), https://en.wikipedia.org/wiki/ID3_algorithm,

- C4.5 (successor of ID3), https://en.wikipedia.org/wiki/C4.5_algorithm,
- CART (Classification And Regression Tree), https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees_.28CART.29,
- Chi-square automatic interaction detection (CHAID), https://en.wikipedia.org/wiki/Chi-square_automatic_interaction_detection,
- MARS, https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_spline.

CART algorithm[1] is a classification algorithm for building a decision tree based on Gini's impurity index as splitting criterion:

1. Find each feature's best split. For each feature with K different values there exist K-1 possible splits. Find the split, which maximizes the splitting criterion. The resulting set of splits contains best splits (one for each feature).
2. Find the node's best split. Among the best splits from Step 1 find the one, which maximizes the splitting criterion.
3. Split the node using best node split from Step 2 and repeat from Step 1 until stopping criterion is satisfied.

Split Criteria

- Misclassification Error: Does the split make the model more or less accurate?
- Gini Index: Favors large partitions. Uses Squared class probabilities.
- Information Gain: Favors smaller partitions. Uses the base-two log of the class probabilities.
- Gain Ratio: Normalizes Information Gain to penalize many small splits.
- ANOVA: Used in Regression Trees. Minimizes the variance in each node.

Stopping Criterion

The most common stopping procedure is to use a minimum count on the number of training instances assigned to each leaf node. If the count is less than some minimum then the split is not accepted and the node is taken as a final leaf node.

[1] Breiman L (1984) Classification and regression trees. The Wadsworth and Brooks-Cole statistics-probability series. Chapman & Hall.

Pruning is a data compression technique. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

1. Split randomly training data into 10 folds.
2. Select pruning level of tree (level 0 equals to full decision tree).
3. Use 9 folds for creation of 9 new pruned trees and estimate error on last 10th fold.
4. Repeat from Step ii until all pruning levels are used.
5. Find the smallest error and use the pruning level assigned to it.
6. Until pruning level is reached, remove all terminal nodes in the lowest tree level and assign decision class to parent node. Decision value is equal to class with higher number of cases covered by node.

Gini Index, Gini impurity

A number between 0 and 1, which indicates the likelihood of new, random data being wrongly classified if it were given a random class label according to the class distribution in the dataset.

- 0 denotes that all elements belong to a certain class or if there exists only one class,
- 1 denotes that the elements are randomly distributed across various classes.
- 0.5 denotes equally distributed elements into some classes.

Consider a dataset *D* that contains samples from *k* classes. The probability of samples belonging to class *i* at a given node can be denoted as p_i. Then the Gini Impurity of *D* is defined as:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

An attribute with the smallest Gini Impurity is selected for splitting the node.

Example (https://www.learndatasci.com/glossary/gini-impurity/)

	Count		Probability		Gini Impurity
	n_1	n_2	p_1	p_2	$1 - p_1^2 - p_2^2$
Node A	0	10	0	1	$1 - 0^2 - 1^2 = 0$
Node B	3	7	0.3	0.7	$1 - 0.3^2 - 0.7^2 = 0.42$
Node C	5	5	0.5	0.5	$1 - 0.5^2 - 0.5^2 = 0.5$

Task 1. Titanic Data

The data has 1309 observations characterised by 6 variables:

- pclass: passenger class,
- survived: died or survived,
- sex: male or female,
- age: age in years,
- sibsp: number of siblings or spouses aboard,
- parch: number of parents or children aboard.

Construct a decision tree to predict the chance of surviving the Titanic sinking.