Decision Trees – Part 2

Decision Tree - Advantages

- Good all purpose classifier that does well on most problems.
- Can handle number and categorical features as well as missing data.
- The algorithm works out what the most relevant features are.
- If the tree is small can easily be interpreted.
  - Results shared
  - Results demonstrated for legal reasons.
- Fast.

Decision Tree - Disadvantages

- Large trees difficult to interpret.
- Large trees can look counter intuitive.
- Small changes in the training data can lead to large changes in the tree.
- All splits are parallel to an axis so can have trouble modelling some shapes.
- Easy to overfit the data.

Applying the decison tree

- This is fairly obvious.
- A decision tree is essentially a set of rules.
- For example
  - If A-List < 2 then
    - Bust
  - Esle
    - If budget < X then
      - Critical Success
    - Else
      - Box Office Success

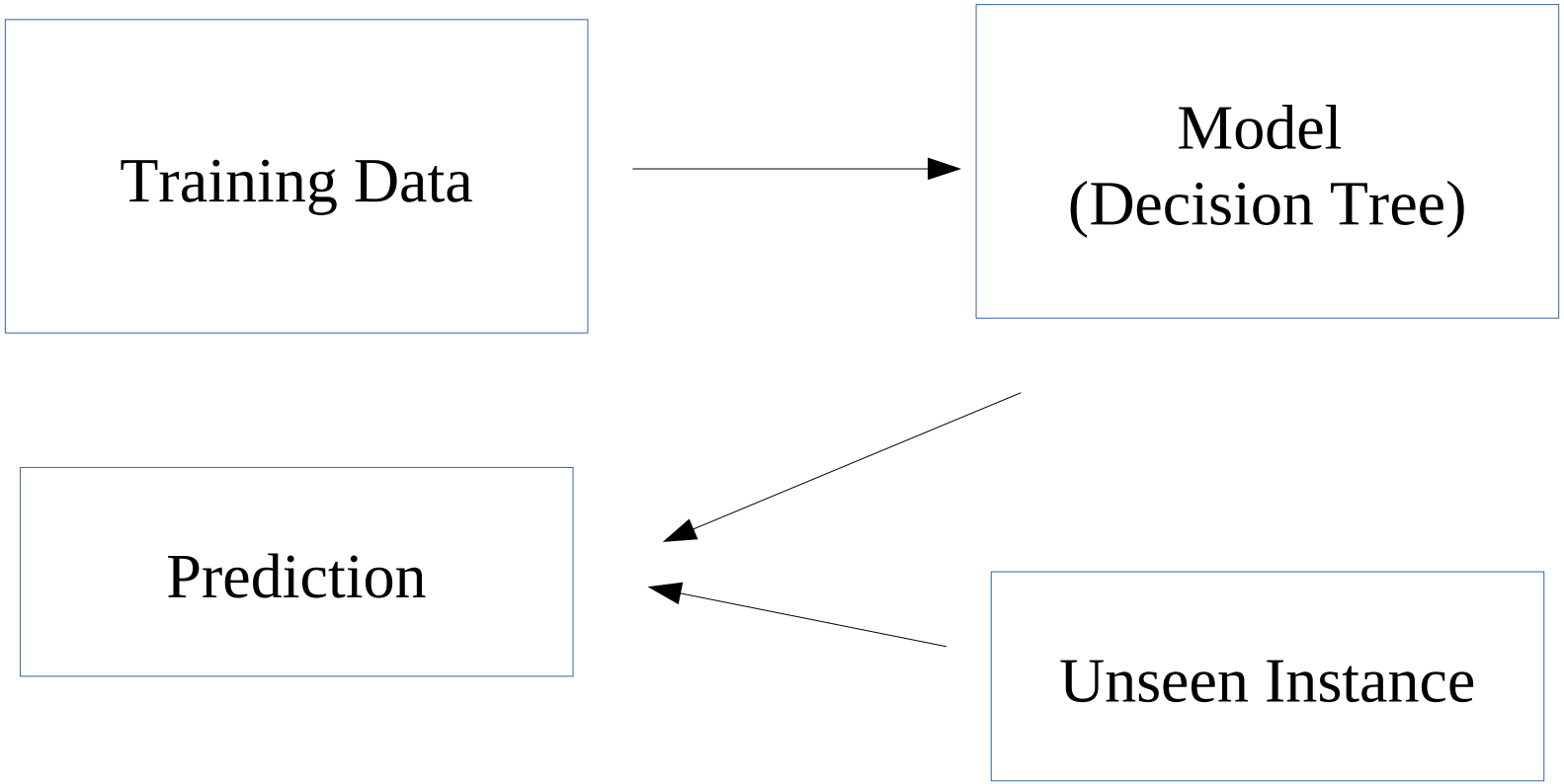
Applying the decison tree

- For given values of the two variables, simply follow the rules to obtain the calssification.
- This correspons to traversing the tree from the root to the one of the leaf nodes which gives a classification.
- The leaf node might contain all instances of the one class.
- Or classification done by choosing the majority class in the leaf node.

Early Stopping

- One common way of stopping the generation of the decision tree is to use a ‘bucket size’.
- Once the number of instances in a node of the tree goes below this bucket size, then it is not split any more.
- For example, 15 instances of class A and 2 instances of class B gives a classification A obviously.

Decision Trees



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Categorizing news stories as finance, weather, entertainment, sports, etc

Types of Input Data

- Numeric
- Categorical Data
  - Nominal (no order)
    - Colour, Species etc.
  - Ordinal (defined order)
    - Low, Medium, High

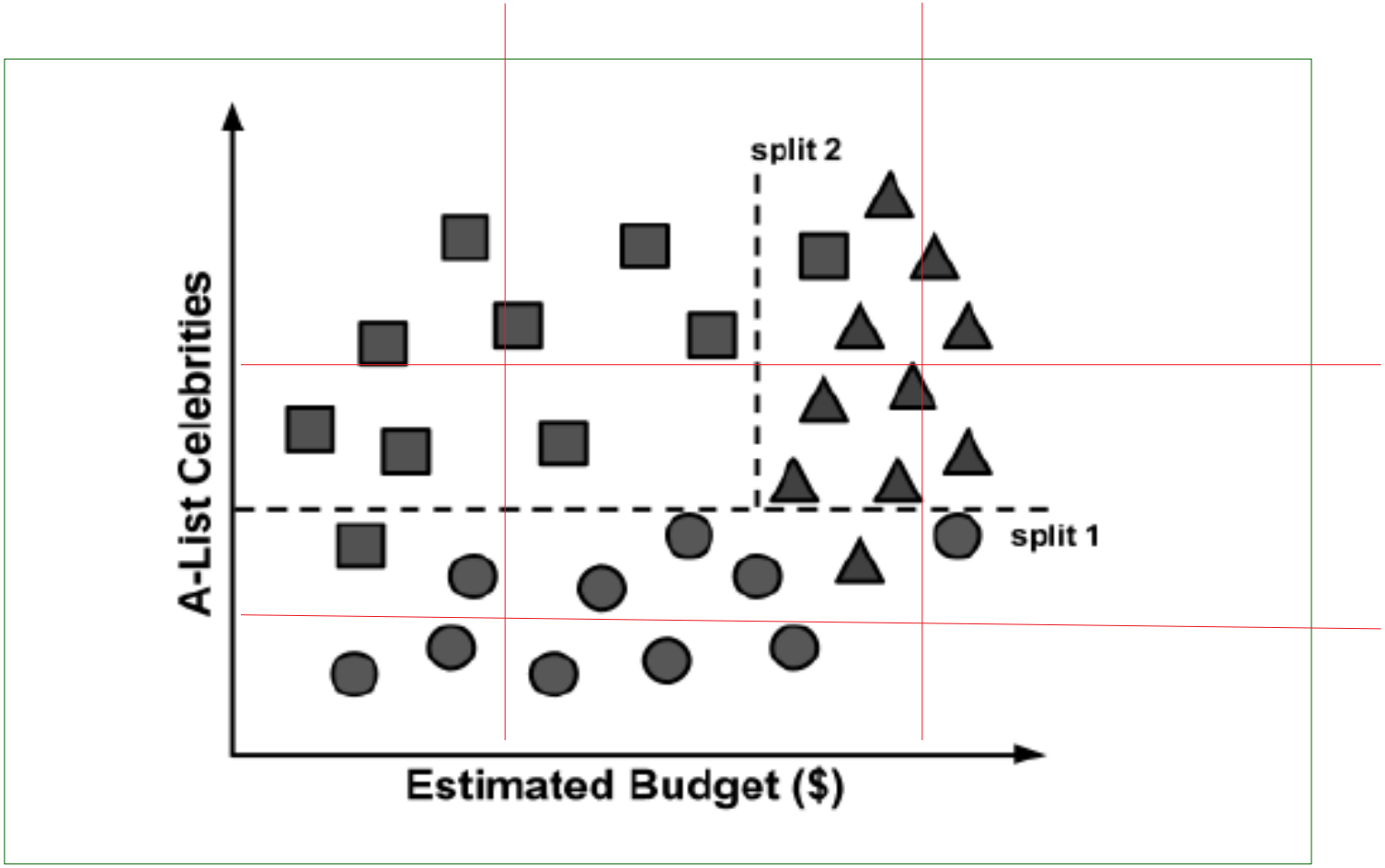
Decision Trees

- Can handle all types of data (numeric, nominal ordinal)
- Sometimes doesn’t work so well with a large number of numeric factors.
- Or with a large number of nominal factors with many different values.
- But these are not hard and fast rules.

Brief Overview of Algorithm

- Decision Tree algorithm considers all possible splits on all possible features.
- Chooses split that best separates the classes.
- This means that child nodes produced by the split should be as homogeneous or pure as possible.

Brief Overview of Algorithm



Brief Overview of Algorithm

- The red lines are alternative splits for the data.
- They do not produce as pure child nodes as the chosen splits.
- The decision tree algorithm considers all such splits before deciding on the best split.
- What is required to do this is a way of measuring the “purity” of a set of instances with respect the target class.

Measures of ‘Impurity’

- A set of instances is pure if they are mostly of the same class
- What is used to drive the decision tree algorithm is a measure of impurity.
- And the decision tree algorithm aims to create nodes in the decision tree with a low level of node impurity.