Decision Trees

Resources

- Introduction to Data Mining - Tan, Steinbach, Kumar
- Introduction to Machine Learning with Python

Recap - Classification

- Given a collection of records (training set).
- Each record contains a set of attributes.
- One of the attributes is the class we want to predict, called the target variable.
- Find a model for the class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.

Classification - Testing

- A test set is used to determine the accuracy of the model.
- Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

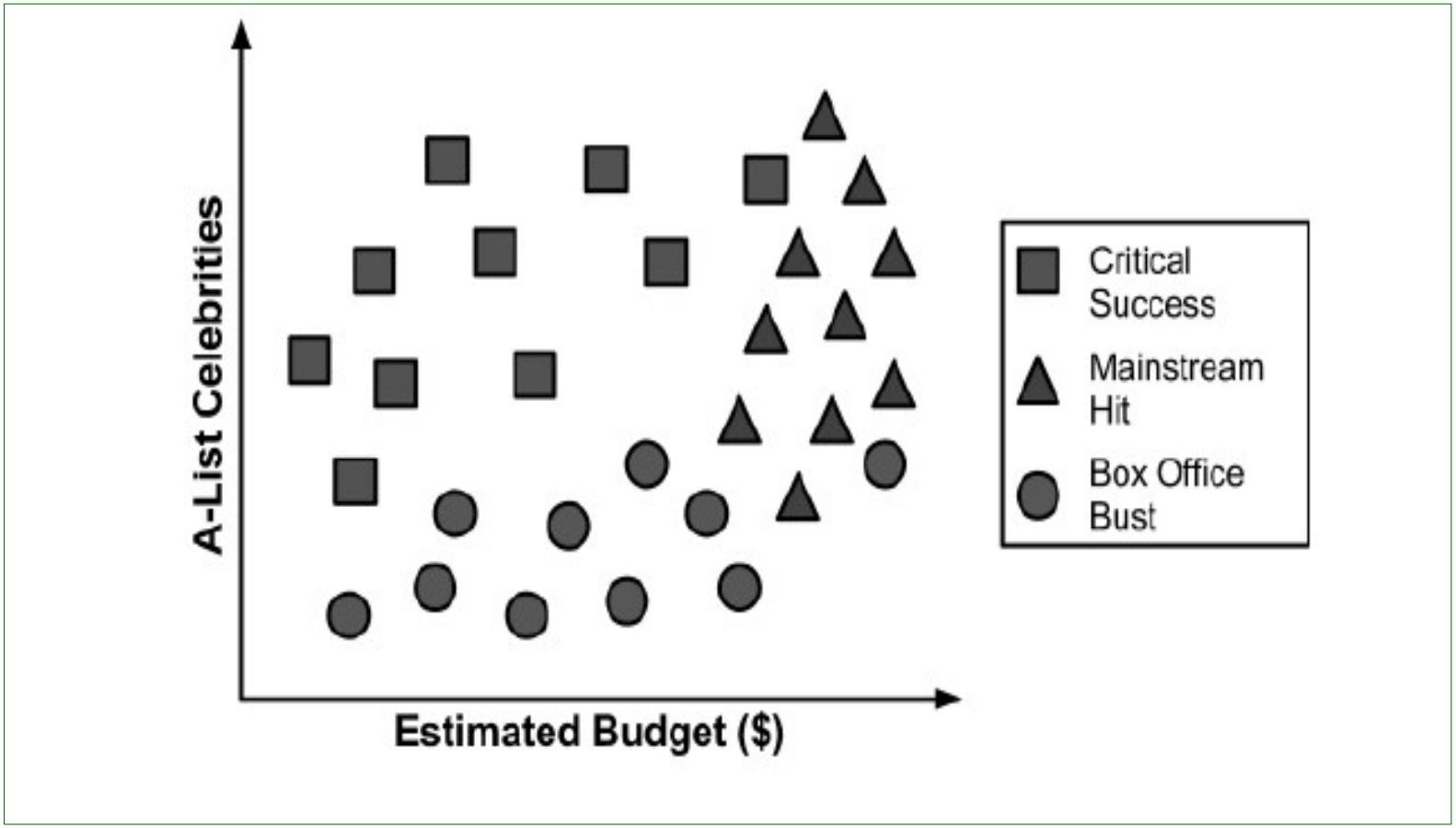
Decision Trees

- Choose a feature that best separates the instances of the target variable (class)
- This feature is then used to partition all the instances creating subtrees.
- The process is repeated until a stopping criterion is reached.

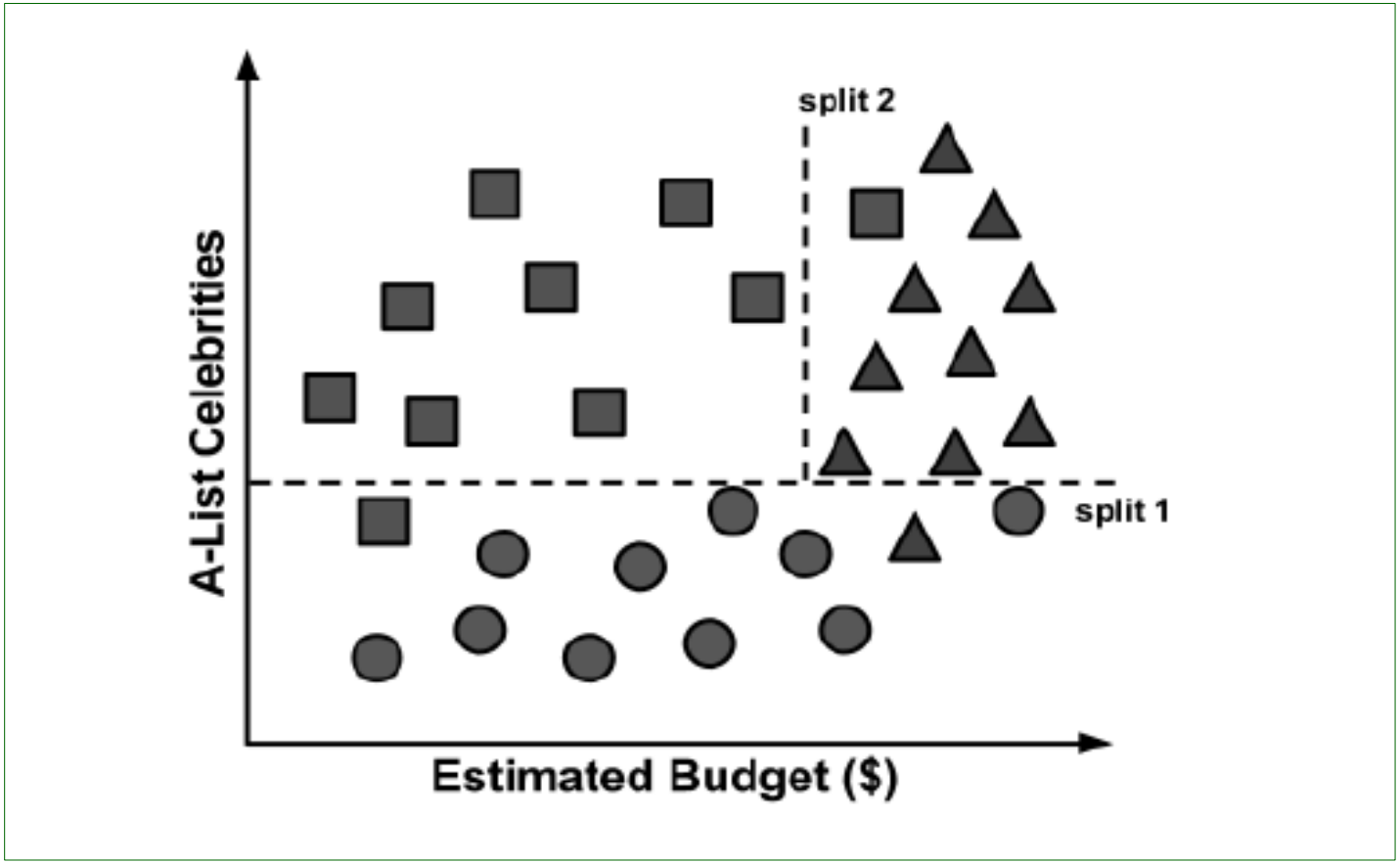
Decision Trees

- The algorithm might stop at a node if:
  - All (or nearly all) of the examples at the node have the same class
  - There are no remaining features to distinguish among examples
  - The tree has grown to a predefined size limit

Simple Example



Decision Boundaries



Decision Tree

- Split the data on number of A-list celebrities.
- Then split again on budget.
- Notice we can keep spliting to get a perfect fit on the training data.
- This can lead to overfitting.

The Decision Tree