greatlearning
Power Ahead

Decision Trees

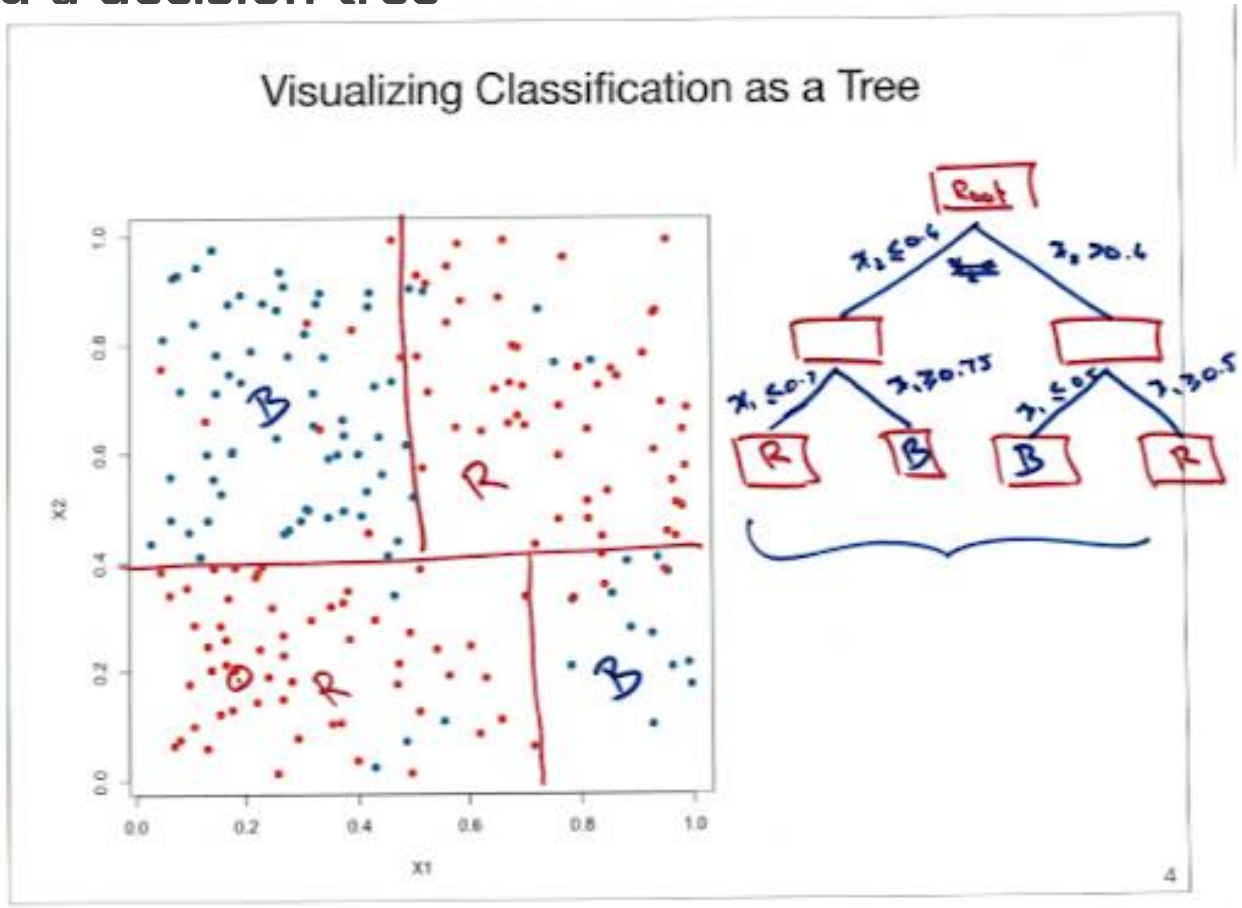
Learning Objectives

- Decision tree introduction
- Entropy
- Gini Index
- Pruning
- Case study

Introduction to Decision Tree

- Decision Tree is used for regression and classification, more often classification.
- Can be used for binary classification such as whether an applicant for loan is likely to turn into defaulter or not.
- Decision tree algorithm finds the relation between the target column and the independent variables and express it as a tree structure.
- It does so by binary splitting data using functions based on comparison operators on the independent columns.

Visualising a decision tree



Common measures of Impurity

Entropy

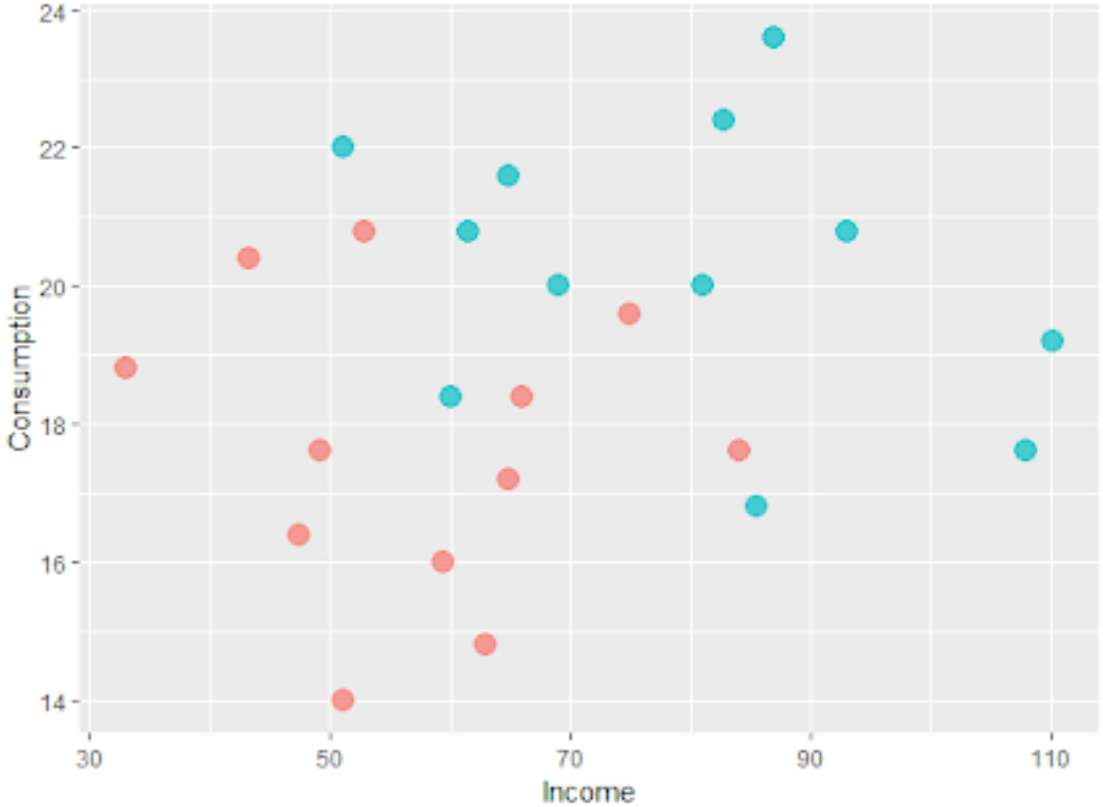
- A measure of uncertainty.
- Given that there are two possible outcomes for a given action.
- We can express the relation between probability and impurity of target column in a mathematical form.

Common measures of Impurity Contd

Gini Index

- Is calculated by subtracting the sum of the squared probabilities of each class from one.
- Perfectly classified, Gini Index would be zero
- Uses squared proportion of classes.

1. Calculate GINI for overall rectangle



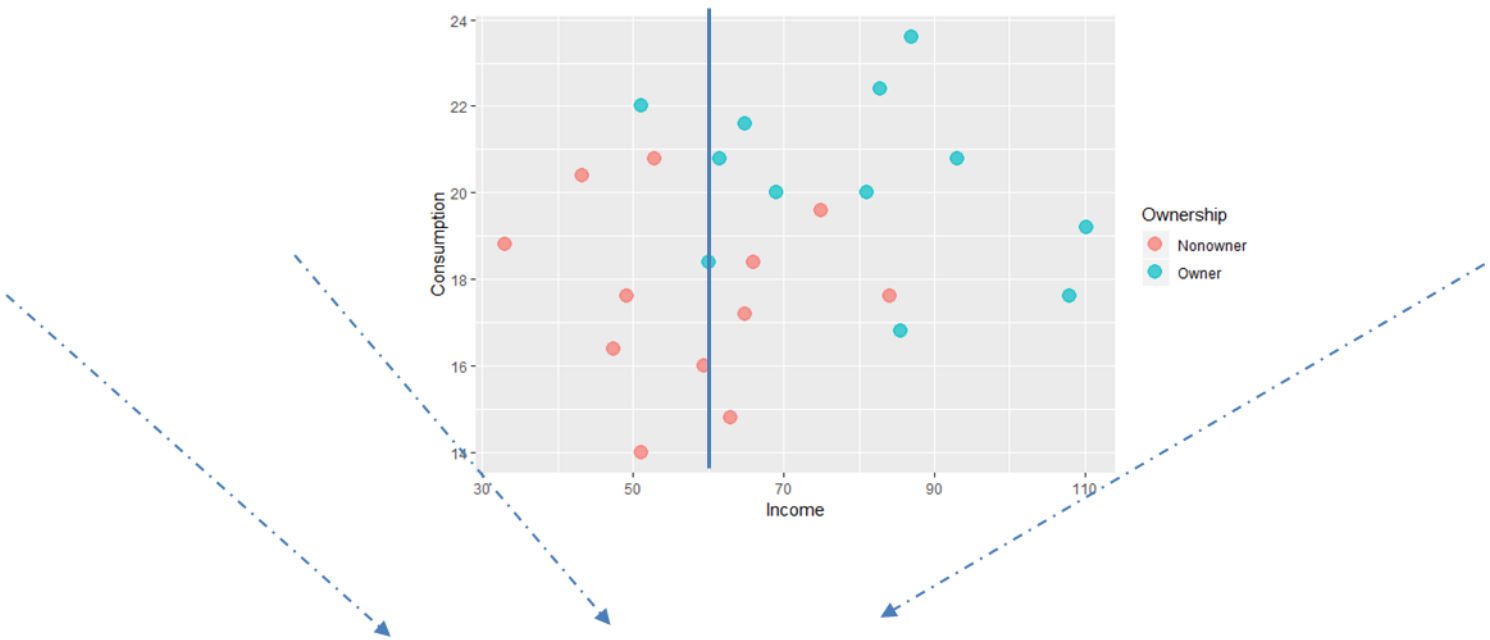
$$I(A) = 1 - \sum_{k=1}^{m} p_k^2,$$

2. Calculation of GINI Index for left and right rectangles

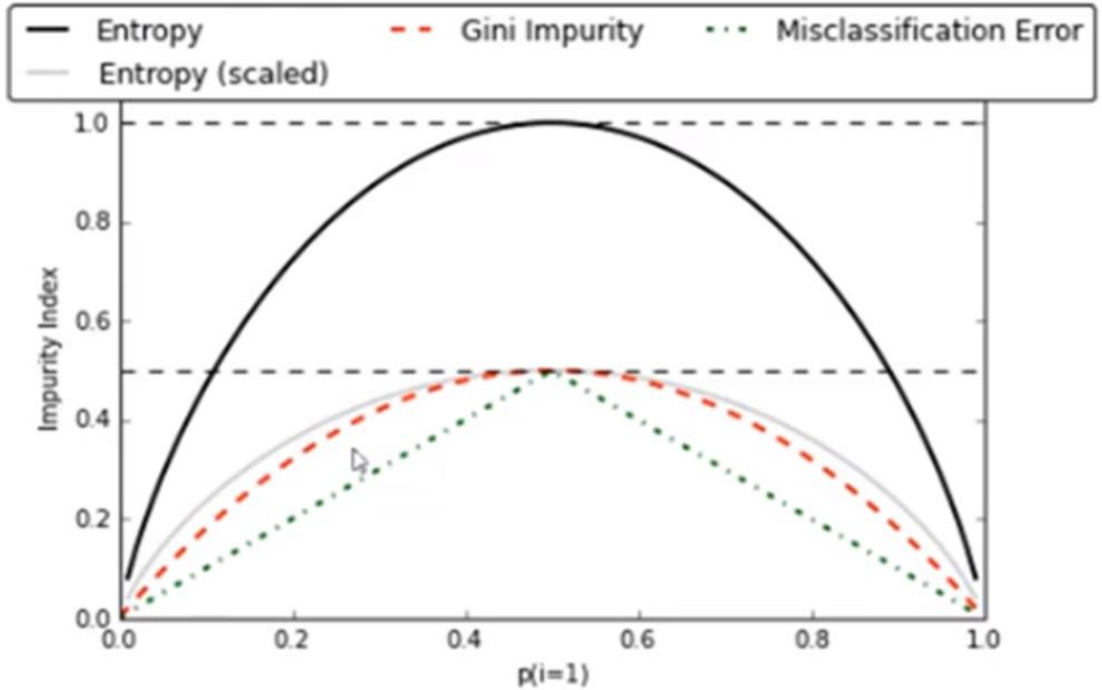
$$I(A) = 1 - \sum_{k=1}^{m} p_k^2,$$

3. Weighted average of impurity measures

GINI index for the left & right



Decision Trees – Gini , Entropy , Misclassification Error

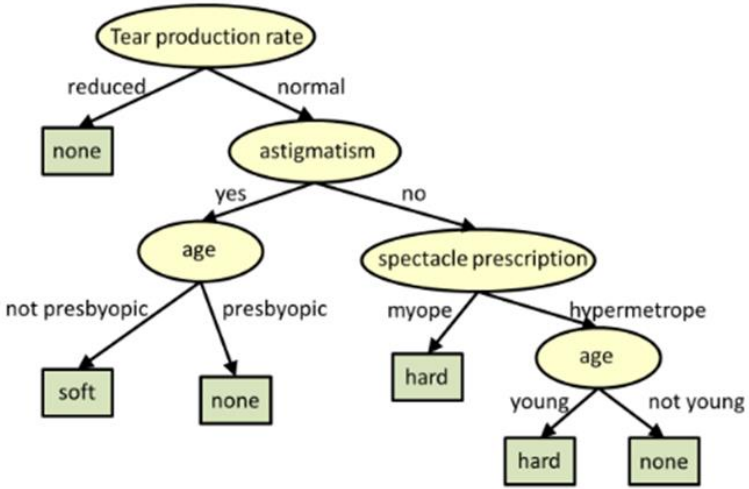


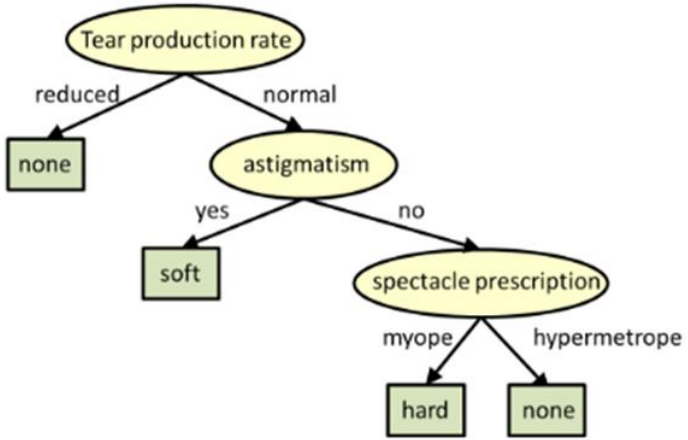
Note: Misclassification Error is not used in Decision Trees

Pruning

- Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances.
- Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Example





Hands on exercise on Decision Tree

Context:

- Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. This is a transnational data set which contains all the transactions occurring online retail.
- Ecommerce data is information relating to the visitors and performance of an online shop. It's mostly used by marketers e.g. in understanding consumer behavior and enhancing conversion funnels.

Objective:

- The objective is to find out the features which have the most information context to differentiate the positive class and negative class and also build a model to predict whether a customer will buy a product or not.

Description of attributes:

- Typically e-commerce datasets are proprietary and consequently hard to find among publicly available data. This is a transnational data set which contains all the transactions occurring online retail. Data Description:
- Out of the 12,330 customer samples in the dataset, 84.5% (10,422) were negative class samples (i.e. customers who did not end up buying the product), and the rest (1908) were positive class samples (i.e. customers who ended up buying).
- The dataset consists of 10 numerical and 8 categorical attributes.
- The 'Revenue' attribute can be used as the class label.
- "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration": These represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories.

Description of attributes:

- The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.
- The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.
- Bounce Rate: The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.
- Exit Rate: The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.
- Page Value: The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

Description of attributes:

- Special Day: The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction.
- The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.
- For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.
- The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.
- Ecommerce data is information relating to the visitors and performance of an online shop. It's mostly used by marketers e.g. in understanding consumer behavior and enhancing conversion funnels.

greatlearning
Power Ahead

Happy Learning !