greatlearning

Decision Trees

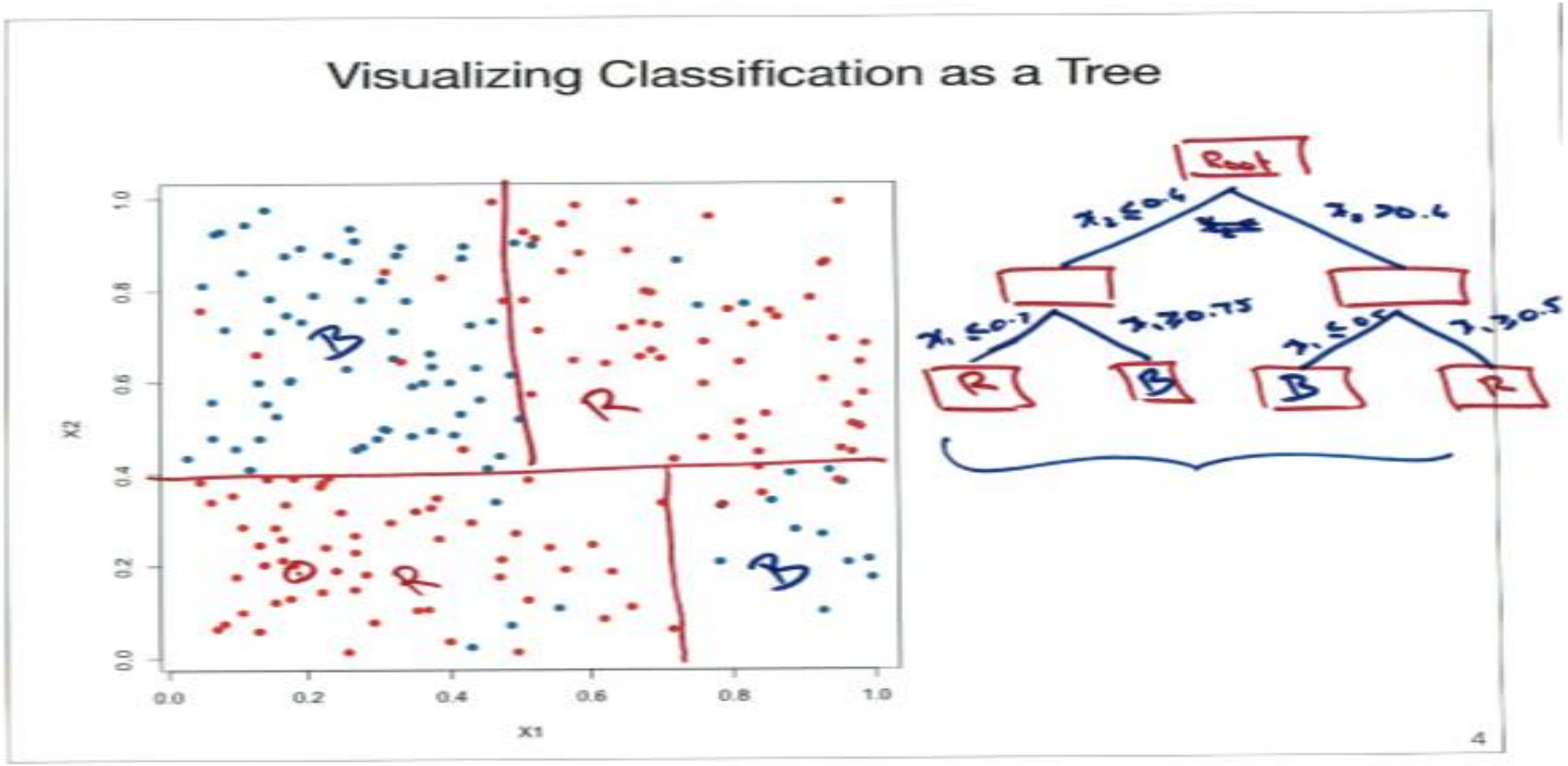
Learning Objectives

- 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

greatlearning

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.

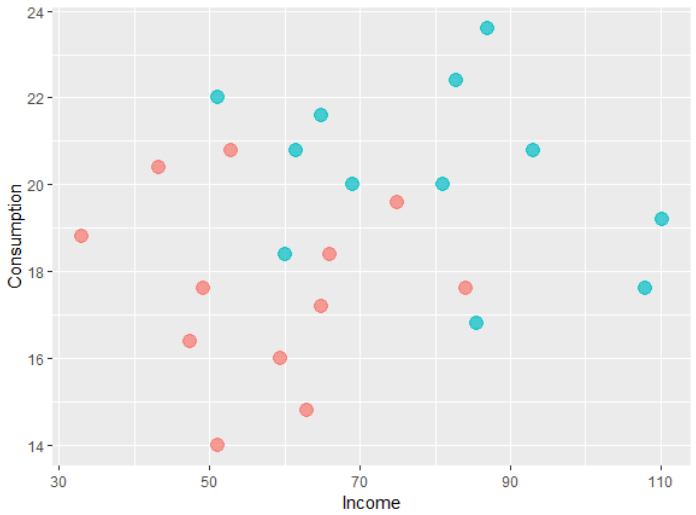
Common measures of Impurity Contd

greatlearning

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

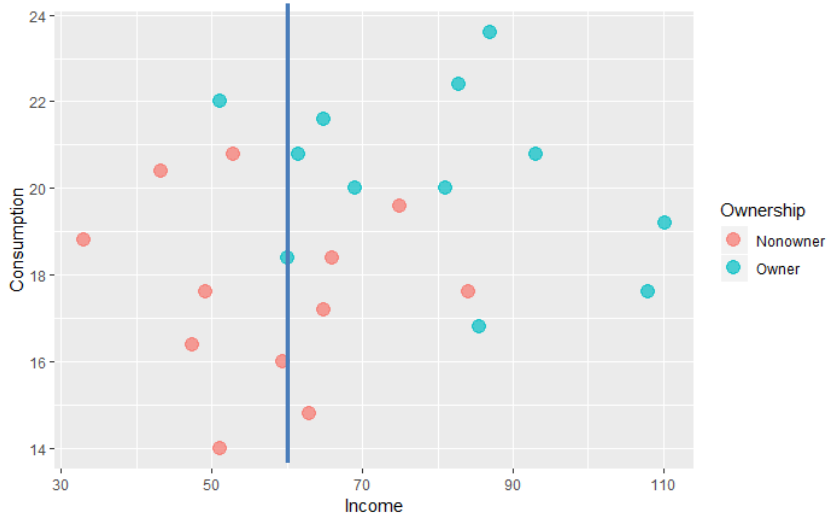


$$I(A) = 1 - \sum_{k=1}^{m} p_k^2,$$

2. Calculation of GINI Index for left and right rectangles

⇒

GINI index for the left & right

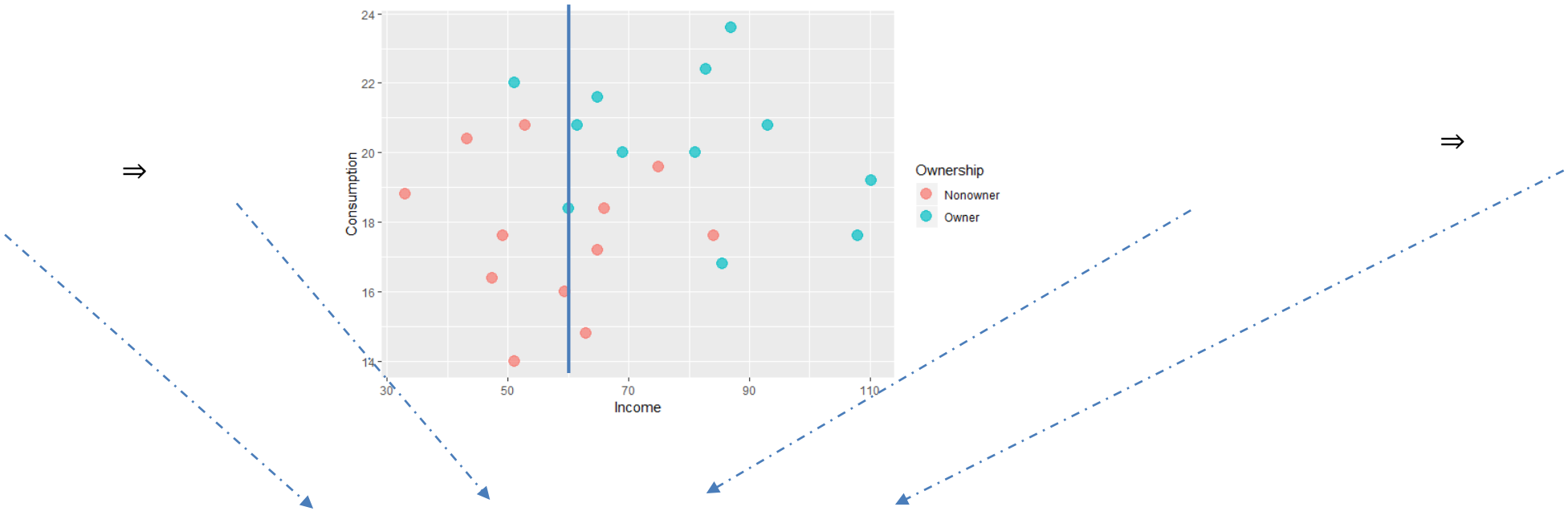


$$I(A) = 1 - \sum_{k=1}^{m} p_k^2,$$

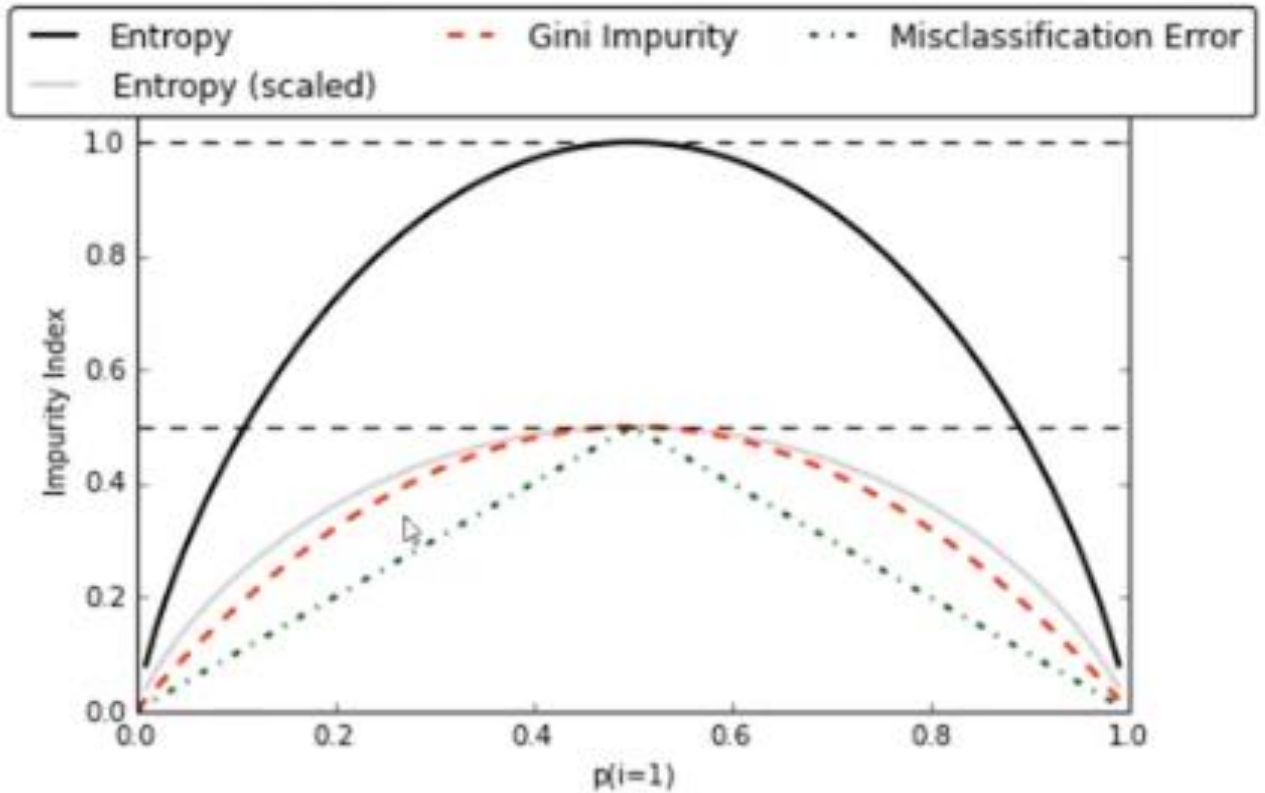
⇒

3. Weighted average of impurity measures

GINI index for the left & right



Decision Trees – Gini , Entropy , Misclassification Error



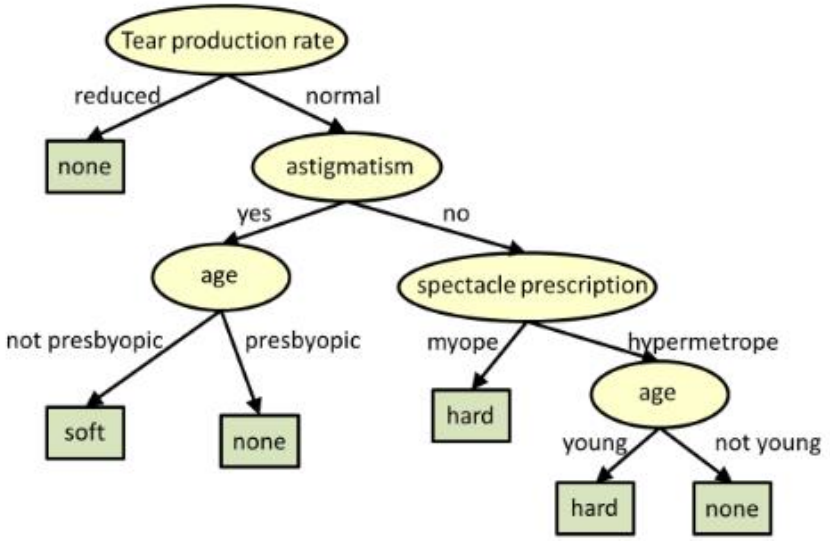
Note: Misclassification Error is not used in Decision Trees

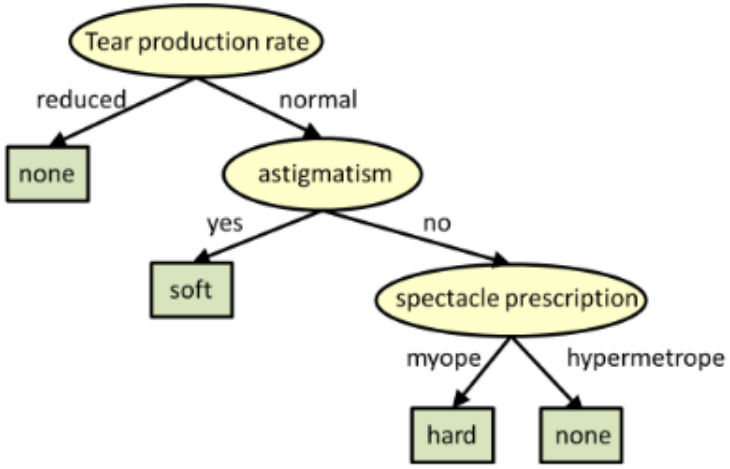
Pruning

- 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

- This datasets is related to red variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).
- These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

Problem Statement:

- Wine Quality Prediction- Here, we will apply a method of assessing wine quality using a decision tree, and test it against the wine-quality dataset from the UC Irvine Machine Learning Repository. The wine dataset is a classic and very easy multi-class classification dataset.

Description of attributes:

- 1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- 2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- 3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
- 4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- 5 - chlorides: the amount of salt in the wine
- 6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

Attribute information Contd.

- 7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- 8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content
- 9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- 10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- 11 - alcohol: the percent alcohol content of the wine
- Output variable (based on sensory data): 12 - quality (score between 0 and 10)

greatlearning

HappyLearning

Questions?