Advanced Statistical Methods
Homework 7

Brandon Hosley

University of Illinois - Springfield

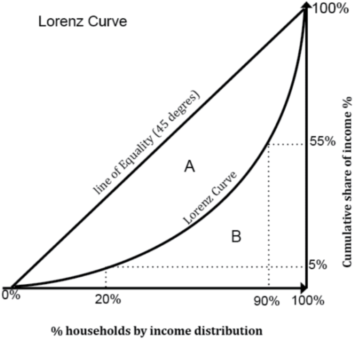
November 15, 2020

Overview

1. Q1A: Gini Index
2. Q1B: Information Gain
3. Q2: Hastie and Tibshirani Summary

Gini Index

- Measurement of a distribution’s inequality
  i.e. How far from a 1 to 1 ratio two normalized traits occur
- Often used to measure income inequality



www.economicshelp.org

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|}{2n^2\bar{x}}$$

Information Gain

- Represents a reduction in entropy by including addition variables
- How much the conditional distribution will decrease by including
  the additional information
- Actual value will be *up to* the true value of mutual information

$$IG_{X,A}(X,a) = D_{KL}(P_X(x|a)||P_X(x|I))$$

Hastie and Tibshirani Lecture: Tree-Based Methods

Pros:

- Relatively Simple
- Good for Classification
- Work well in ensembles

Cons:

- Underperform when compared to modern techniques

Hastie and Tibshirani Lecture: Bootstrap and Bagging

- Using the Bootstrap discussed earlier
- Ensemble the trees trained on each subset

Pros:

- Reduces Bias and Error
- Adding more trees never hurts

Cons:

- Much more computationally intensive than single tree