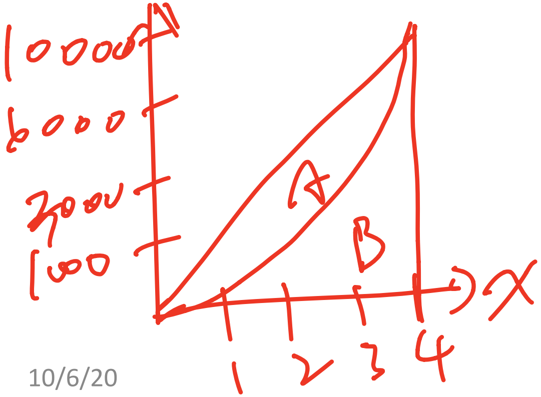
Review quizzes

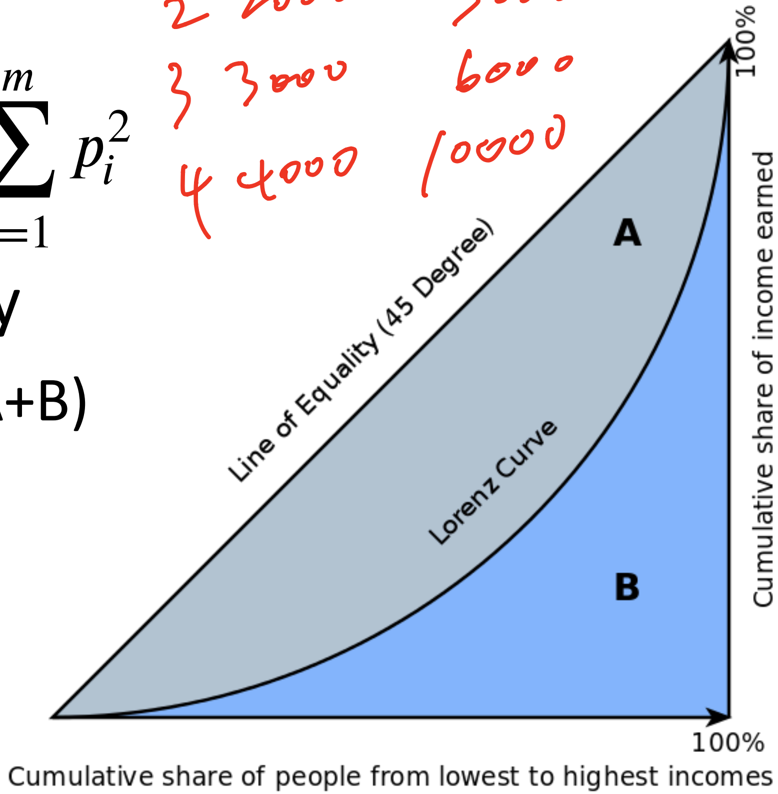
- What is L2-norm?
- What is L1-norm?
- What are cosine similarity and cosine distance?
- What is entropy?
- What is the main difference between information gain and gain ratio?

Two different gini-indices

- Decision tree
  - $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$
- Income inequality
  - Gini-index = A/(A+B)

1	1000	1000
2	2000	3000
3	3000	6000
4	4000	10000




100%
Cumulative share of income earned
A
Line of Equality (45 Degree)
Lorenz Curve
B
100%
Cumulative share of people from lowest to highest incomes

When p_i's give maximum entropy and minimum entropy? (1/3)

- Entropy:

$$\mathrm{Ent}\left(p_1, \ldots, p_n\right) = -\sum_{i=1}^{n} p_i \log p_i$$

- If $n = 2$

$$\mathrm{Ent}\left(p_1, p_2\right) = -p_1 \log p_1 - \left(1 - p_1\right) \log\left(1 - p_1\right)$$

- What is the value of p_1 to maximize and minimize $\mathrm{Ent}\left(p_1, p_2\right)$?
  - Possible points include
    - Extreme points: $p_1 = 0$, $p_1 = 1$
    - $\nabla_{p_1} \mathrm{Ent}\left(p_1, p_2\right) = 0$

When p_i's give maximum entropy and minimum entropy? (2/3)

$$\mathrm{Ent}(p_1, p_2) = -p_1 \log p_1 - (1 - p_1)\log(1 - p_1)$$

- If $p_1 = 0,\ \mathrm{Ent}(p_1, p_2) = 0$
- If $p_1 = 1,\ \mathrm{Ent}(p_1, p_2) = 0$
- If $\nabla_{p_1}\mathrm{Ent}(p_1, p_2) = 0$

$$\Rightarrow -\log p_1 - \frac{p_1}{p_1} - (-1)\log(1 - p_1) - (1 - p_1)\frac{-1}{1 - p_1} = 0$$

$$\Rightarrow -\log p_1 - 1 + \log(1 - p_1) + 1 = 0$$

$$\Rightarrow \log(1 - p_1) = \log p_1$$

$$\Rightarrow 1 - p_1 = p_1$$

$$\Rightarrow p_1 = \frac{1}{2},\ \mathrm{Ent}(p_1, p_2) = 1$$

When p_i's give maximum entropy and minimum entropy? (3/3)

- Values of $p_1, \ldots, p_n$ $(n > 2)$ to maximize and minimize entropy?
- Possible points:
  - $p_i = 1, p_{-i} = 0$ $(p_{-i} = \left[p_1, . , , , . p_{i-1}, p_{i+1}, \ldots, p_n\right])$
  - $\nabla_{p_i} \mathrm{Ent}\left(p_1, \ldots, p_n\right) = 0,$ $\sum p_i = 1, 0 \leq p_i \leq 1\ \forall i$
    - This can be solved by Lagrange multiplier, which will be discussed in future lectures

Exercise 2

- Requirement
  - Implement a decision tree classifier using Python. (50%)
    - You **cannot** use existing decision tree libraries (e.g., `sklearn.tree.DecisionTreeClassifier`)
  - Use your classifier to predict the class of the iris plants based on the Balance Scale Data Set (http://archive.ics.uci.edu/ml/datasets/Balance+Scale). (40%)
    - Separate the data into training (70%) and test (30%) datasets. Please make sure the dataset is split in a stratified fashion, i.e., the class distributions in the training and the test datasets are the same as the class distribution in the entire dataset.
    - Report both the training and the test error for *k* = 1, 2, 3, …, 20
  - A brief discussion of the results. (10%)
- Please submit your code and report to new ee-class
- Due date: 10/19 23:59:59