CSCI3230
Entropy, Information and Decision tree

CHEN Ran

Fall 2019

A scenario

- Suppose you are a police officer, and trying to find out whether a guy is guilty or not by asking him questions
- Initially, you are uncertain about the truth(he is only a suspicion), so P(guilty) =0.5 and P(not guilty) = 0.5
- Your aim, is to figure out the truth (certain), which means that figure out whether P(guilty) = 1 (P(not guilty) = 0) or P(not guilty) =1 (P(guilty) =0)
- Your boss want you to do it efficiently.(short time, less question)

Certainty and Probability

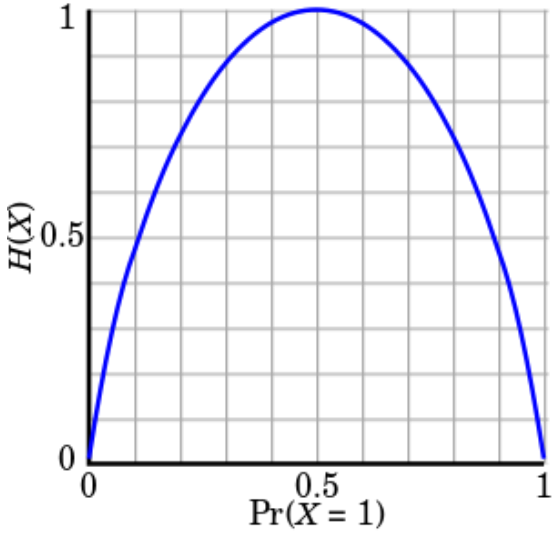
- In previous example: uncertain => P(guilty) = P(not guilty) =0.5

  certain => P(guilty) =1 or P(not guilty) = 1
- Conclusion: more uniformly distributed, more uncertain

  more concentrated, more certain
- It’s only empirical, any mathematical measure of uncertainty?

Entropy and information

- Entropy =$\sum -p_i \log(p_i)$
- Entropy comes from information theory. The higher the entropy, the less the information content



- Any other measure of uncertainty?
- Yes, Gini-Index

Order and efficiency

- Question: How to make your interrogation efficient as the policy officer?
- Answer: to make your questions informative, the more informative a question is, the earlier you should ask it

What is Informative?

- You have some information(entropy) before you ask a question.
- Your information (entropy) is updated after a question.

- Informative means the change of your information after a question, so the change of entropy

- $\Delta\, information = |information_{new} - information_{old}| = |Entropy_{old} - Entropy_{new}|$

Computation

- Entropy_old: the entropy of the labels in the dataset

- Entropy_new:

1. by asking a question, you divide the whole scenario(dataset) into several smaller situations (sub dataset)
2. Each situation (sub dataset) may happen with different probability
3. The new entropy is the probability weighted entropy of each sub dataset

Example

- Whether a student like the movie Titanic

Gender	Major	Like
Male	Math	Yes
Female	History	No
Male	CS	Yes
Female	Math	No
Female	Math	No
Male	CS	Yes
Male	History	No
Female	Math	Yes

- Initial Entropy:

$$E(Like) = -\frac{4}{8}\log\left(\frac{4}{8}\right) - \frac{4}{8}\log\left(\frac{4}{8}\right) = 1$$

Example

- Ask Major?

Gender	Major	Like
Male	Math	Yes
Female	History	No
Male	CS	Yes
Female	Math	No
Female	Math	No
Male	CS	Yes
Male	History	No
Female	Math	Yes

P(Major=Math) = 0.5

$$E(Like|Major = Math) = -\frac{2}{4}\log\left(\frac{2}{4}\right) + -\frac{2}{4}\log\left(\frac{2}{4}\right)=1$$

P(Major=History) = 0.25

$$E(Like|Major = histor) = -\frac{2}{2}\log\left(\frac{2}{2}\right) + -\frac{0}{2}\log\left(\frac{0}{2}\right)=0$$

P(Major=CS) = 0.25

$$E(Like|Major = CS) = -\frac{2}{2}\log\left(\frac{2}{2}\right) + -\frac{0}{2}\log\left(\frac{0}{2}\right)=0$$

$$Entropy = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$$

Example

- Ask Gender?

Gender	Major	Like
Male	Math	Yes
Female	History	No
Male	CS	Yes
Female	Math	No
Female	Math	No
Male	CS	Yes
Male	History	No
Female	Math	Yes

P(Gender=male) = 0.5

$$E(Like|Gender = male) = -\frac{1}{4}\log\left(\frac{1}{4}\right) + -\frac{3}{4}\log\left(\frac{3}{4}\right) = 0.475$$

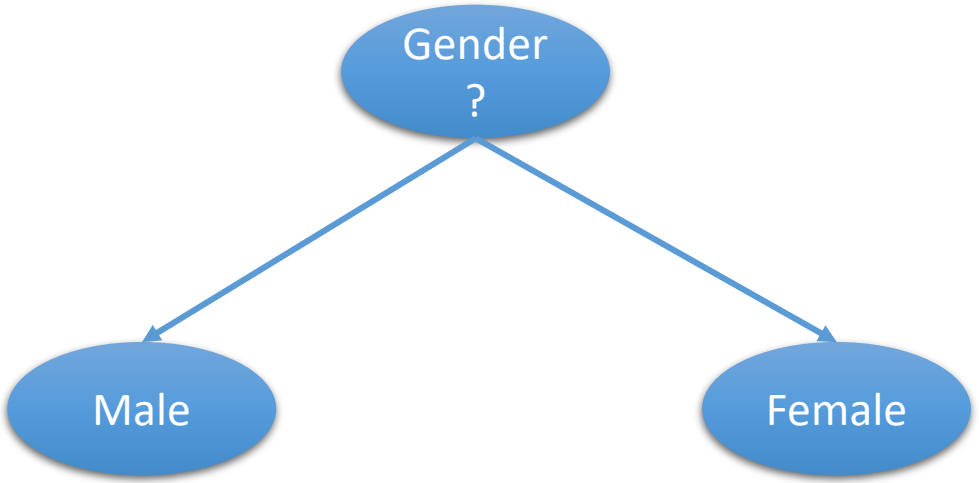
P(Gender =female) = 0.5

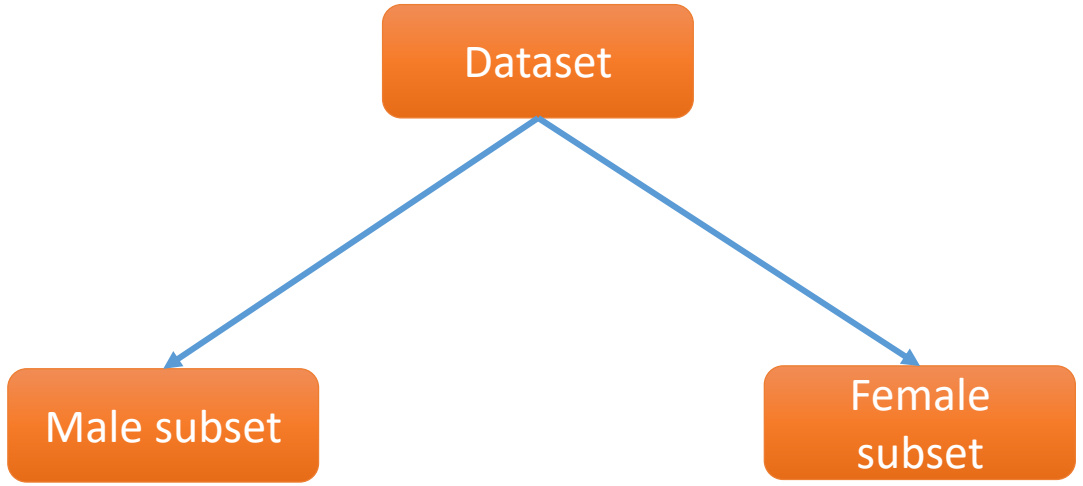
$$E(Like|Gender = female) = -\frac{3}{4}\log\left(\frac{3}{4}\right) + -\frac{1}{4}\log\left(\frac{1}{4}\right) = 0.475$$

$$Entropy = 0.5 * 0.475 + 0.5 * 0.475 = 0.475$$

Which one first?

- Ask Gender first because $1 - 0.5 < 1 - 0.475$!





What are these subsets ?

Sub dataset

Male subset

Major	Like
Math	Yes
CS	Yes
CS	Yes
History	No

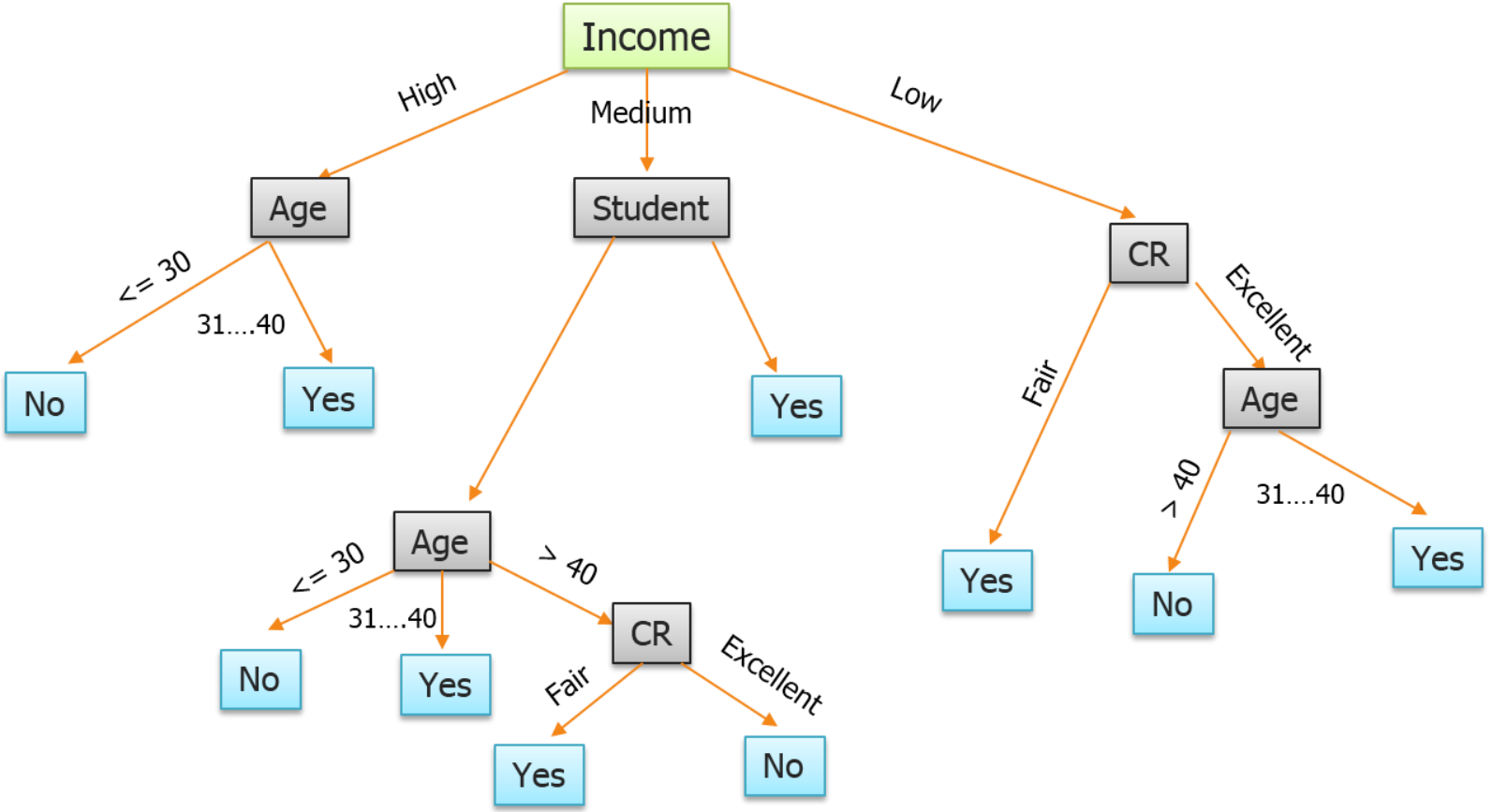
Female subset

Major	Like
History	No
Math	No
Math	No
Math	Yes

- You can ask a new question (make a new decision)on each subset, and going on and on…

Decision tree

- By making decisions in this way, you get an decision tree



- No Tutorial next week.
- Tutorial the week after next week(Nov 13,14):

  Data Mining

  Weka