

Entropy and Information Gain*(find and work with a partner)*

Given the movie data below (Liked (Li) is the response variable), select the feature that maximizes the information gain. The steps below will guide you through the process.

Movie	Type	Length	Director	Famous actors	Liked?
m1	Comedy	Short	Adamson	No	Yes
m2	Animated	Short	Lasseter	No	No
m3	Drama	Medium	Adamson	No	Yes
m4	Animated	Long	Lasseter	Yes	No
m5	Comedy	Long	Lasseter	Yes	No
m6	Drama	Medium	Singer	Yes	Yes
m7	Animated	Short	Singer	No	Yes
m8	Comedy	Long	Adamson	Yes	Yes
m9	Drama	Medium	Lasseter	No	Yes

1. First compute the probability that the outcome is “yes”:

$$P(\text{Li} = \text{yes}) =$$

2. Next compute $H(\text{Li})$ using a calculator.

$$H(\text{Li}) =$$

3. Finally, given the conditional entropy values below, compute the information gain of each feature.

$$H(\text{Li} \mid \text{T}) = 0.61$$

$$H(\text{Li} \mid \text{Le}) = 0.61$$

$$H(\text{Li} \mid \text{D}) = 0.36$$

$$H(\text{Li} \mid \text{F}) = 0.85$$

$$\text{Gain}(\text{Li}, \text{T}) =$$

$$\text{Gain}(\text{Li}, \text{Le}) =$$

$$\text{Gain}(\text{Li}, \text{D}) =$$

$$\text{Gain}(\text{Li}, \text{F}) =$$

4. Based on your results, which feature is most informative for the outcome?

Based on materials by Jessica Wu and Ziv Bar-Joseph