Foundations of Data Science
Lecture 3, Module 1
Fall 2022

Rumi Chunara, PhD

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Data Preprocessing

Major Tasks in Data Preprocessing

- Data sampling
- Data cleaning
- Data integration
- Data reduction

Major Tasks in Data Preprocessing

- Data sampling
- Data cleaning
- Data integration
- Data reduction

Data Reduction

- **Goal**: obtain a reduced representation of the data set that is *much smaller* in volume but yet leads to very similar analytical results

- **Why data reduction?**
  - Complex data analysis may take a very long time to run
  - Important if you want to retain only the most relevant information about the data

What Do We Mean by *"Information"*?

"Any quantity that can reduce uncertainty about another quantity"

Information Example

If a carnival operator wanted to reduce the uncertainty about your weight, what information might he use?

Conditional thinking

So much of what we do in data science involves thinking in terms of…

$$E[Y|X]$$

And deciding whether or not $E[Y|X] \neq E[Y]$

In other words: are Y and X dependent?

From a decision making standpoint, conditional thinking is being able to make a better judgment about a potential outcome Y if we were to know the value of X.

Information and Data Reduction

If we have a target variable in our dataset, along with a set of features, we often wish to:

1) For each feature, determine whether it contains important information about the target variable (in other words, does a given feature reduce the uncertainty about the target variable?)

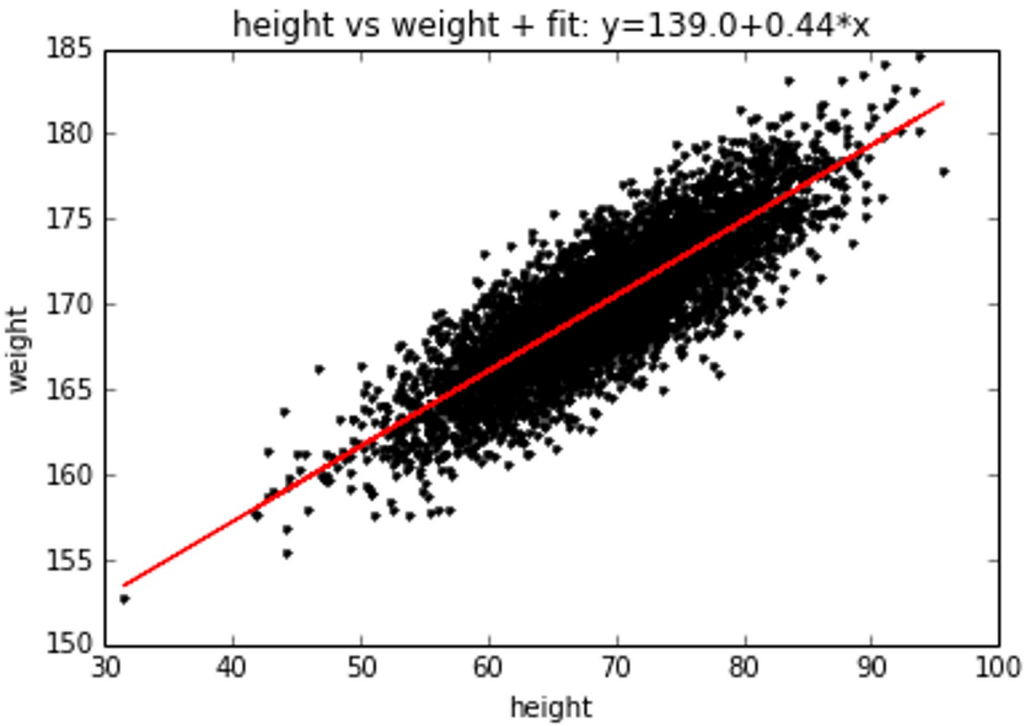
2) Select the features that are best suited for predicting the target variable

3) Rank each feature on its ability to predict the target variable

If certain features do not reduce the uncertainty, we may remove them and end up with a **reduced representation of our dataset!**

Reducing Uncertainty

- Let's assume the carnival operator is also a data scientist
  - She collects data on height vs. weight and sees that they are strongly linearly correlated!



We can then learn:

E[Weight|Height]
as opposed to
E[Weight]

Information and Entropy

- **Entropy**: a measure of the unpredictability or uncertainty of a given information content
  - Certainty: close to 0% or 100% sure that something will happen
  - Fundamental concept used for ranking and selection of features

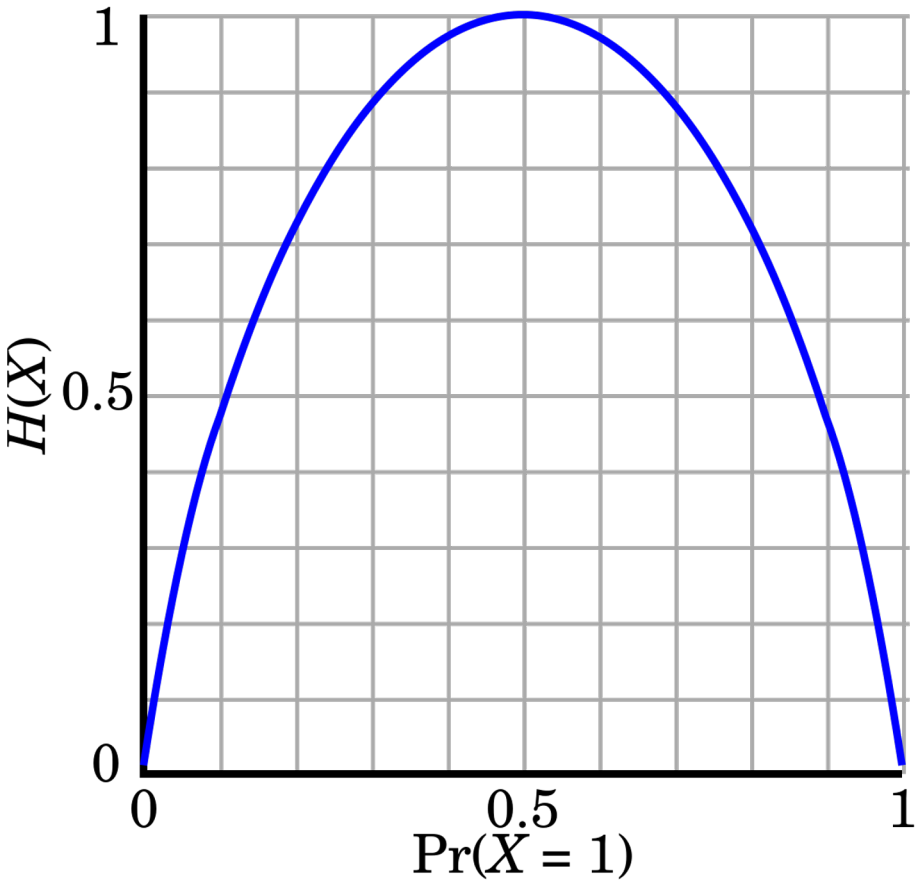
Entropy

- More formally, if X is a random variable with {x1, x2 … xn} possible values, the entropy of X is the expected "information" of an event, where "information" is defined as –log(P(X)).

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

http://en.wikipedia.org/wiki/Entropy_(information_theory)

Binary Entropy Function



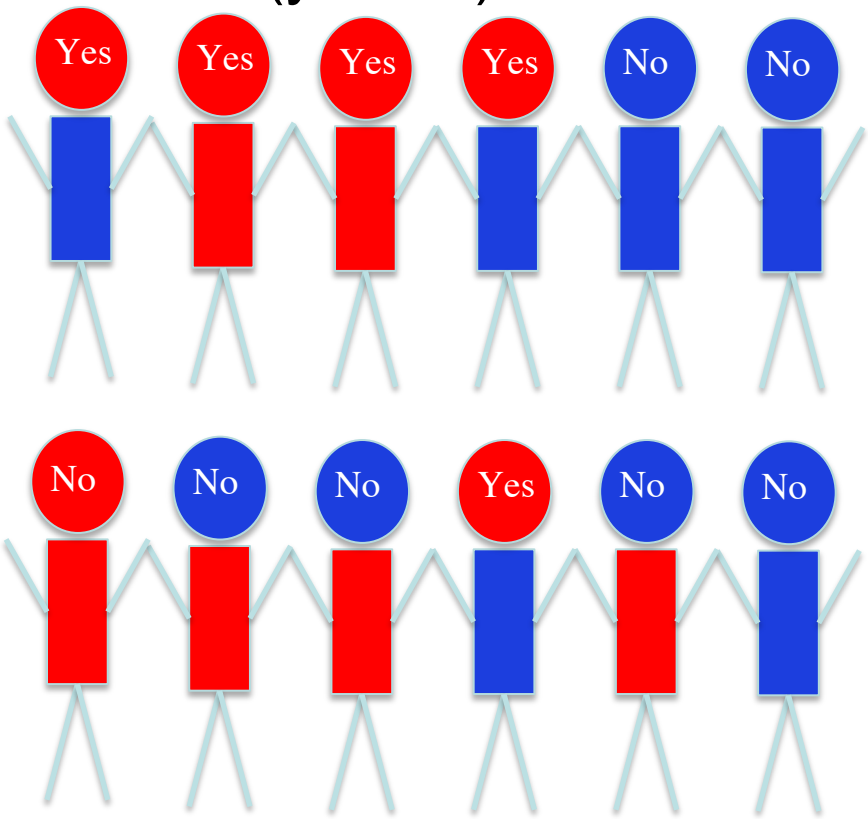
https://en.wikipedia.org/wiki/Binary_entropy_function

Bernoulli process with probability *p* of one of two values (e.g., head or tail).

Here, we use **log2**.

Conditional Entropy

We want to explore whether the features “head color” and “body color” give us more information about our target variable (yes/no).



We use **Conditional Entropy**

$$H(Y|X) \equiv \sum_{x \in \mathcal{X}} p(x)\, H(Y|X=x)$$

where,

- Y is the binary target variable
- X is the feature
- x is a value of a feature
- p(x) is the probability X=x
- H(Y|X=x) is the entropy of Y where X=x

$$H(Y|X=x_i) = -\sum_{yi \in \{yes,no\}} P(Y=yi|X=x_i) log(P(Y=y_i|X=x_i))$$

Example: Conditional Entropy

To compute conditional entropy of an attribute: $H(Y|X) \equiv \sum_{x \in \mathcal{X}} p(x)\, H(Y|X=x)$

Compute $P(X_i)$, $P(Y|X_i)$, $H(Y|X_i)$ for each attribute value X_i:

Attribute	Value	P(Xi)	P(Y=yes|X=Xi)	P(Y=no|X=Xi)	H(Y|X=Xi)
Body	Red				
Body	Blue				
Head	Red				
Head	Blue				

Now apply the Conditional Entropy formula to the given attributes:

H(Y|Body) =

H(Y|Head) =

For this exercise, **use log2** (logarithm with base 2)

Which variable helps in the prediction of the target the most?