

Data Analysis & Modeling Techniques

Information Theory



Information Theory

- Information theory address the question how much information a particular event (or message) contains assuming no background knowledge
 - Information theory does not deal with the semantics (i.e. the meaning) of the event or the message.
 - It solely deals with the minimal amount of information required to represent its content in the absence of background information
 - How much information is required to perfectly remember the event/ message ?
 - What is the minimal size to which the event/message can be compressed?
- Information theory was established by Claude Shannon in the 1940s



Information Theory

- Information theory deals with a number of problems associated with information
 - How much information does an event provide us with ?
 - How far can we compress a message ?
 - What is the most compact encoding of a message ?
 - How much capacity does a transmission channel have if we use a particular encoding?
- Information theory disregards any semantics of the message or knowledge outside the actual message
 - Information is directly tied to the probability of the event or message
 - Events that are easily predictable contain less information than ones that are more difficult to predict
 - Events that are known with certainty to occur do not provide any information since we already knew their content beforehand



Information

- The Information content, I(p), of an event with likelihood p has to fulfill a number of properties
 - Information can not be negative

$$I(p) \ge 0$$

- No occurrence of an event can take away from a the information in previous events since we are not modeling their semantics but rather are only looking at remembering / representing the events
- An event that has probability 1, contains no information I(1) = 0
 - We obtain no information from observing such an event and need no information to remember it.
- The information provided by two independent events has to be the sum of the information from each one

$$I(p_1 * p_2) = I(p_1) + I(p_2)$$

 The occurrence one event can not provide us with any information about the occurrence of an independent event.

Information

 Based on the required properties for a measure of information and assuming that it is continuous we can derive a measure for information:

$$I(p^{x}) = x * I(p)$$
 for all $0 , $x > 0$
 $0 \le I(p)$
 $\Rightarrow I(p) = -\log_{b}(p)$$

- The base b selects the unit in which we measure the information but is not important for any of the calculations
 - b=2: bits
 - *b=3* : trits
 - b=e: nats
 - b=10: Hartleys



Entropy – The Mean Information

 For any random variable X with a probability mass function (pmf) p_i it is thus possible to measure the information content of each outcome

$$I(x_i) = I(p_i) = -\log_b(p_i)$$

 The expected information of an experiment (i.e. of the unknown value of X) can be computed

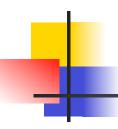
$$H(X) = E[I(X)] = \sum_{x \in X} p(x)I(x)$$
$$= -\sum_{x \in X} p(x)\log_b p(x)$$



Entropy

- Entropy can be interpreted in a number of ways and address a number of questions
 - What is the mean information we gain by querying a variable ?
 - How many bits do we need on average to represent the value of X in its most compressed form ?
 - How many (binary) questions do we need to ask on average to find the outcome?
- Entropy is a measure of disorder and uncertainty
 - Larger entropy represents more uncertainty and thus more information from finding out the result

© Manfred Huber 2017



Maximum Entropy

- We know that the minimum entropy is 0.
- Can we say something about the maximum entropy?
 - Maximum entropy corresponds to the highest information content in observing a random variable
- Is there a distribution for which the entropy is highest?

© Manfred Huber 2017

Maximum Entropy

- Bounded distributions are either finite or have a lower and upper bound
 - What distribution does have the highest entropy?
- The highest entropy distribution is the one that make sit the hardest to predict the outcome
 - The highest entropy for a left and right bounded distribution is that of the uniform distribution.
 - $H(P) = log_h(N)$ for discrete
 - $H(P) = log_b(upb lwb)$ for continuous



"Non-bounded" Maximum Entropy

- What unbounded, non-heavy tailed probability density function has the highest entropy?
 - i.e. μ , σ^2 are bounded
- The normal distribution $N(\mu, \sigma^2)$ has maximum entropy among all real-valued distributions with specified μ, σ^2
 - $H(X)=1/2 \log_b(2\pi \sigma^2)$



"One-bounded" Maximum Entropy

- What non-heavy tailed probability density function that is bounded in one direction has the highest entropy?
 - i.e. μ , σ^2 are again bounded
- The exponential distribution with mean $1/\lambda$ has maximum entropy among all realvalued, one tailed distributions defined over $[0..\infty)$ and mean $1/\lambda$
 - $H(X)=1-log_b(\lambda)$



Relative Entropy

 Relative Enthropy (Kullback-Leibler distance) is a measure of the difference of two distributions

$$D(p \parallel q) = E_p \Big[\log_b \Big(p(x) / q(x) \Big) \Big]$$
$$= \sum_{x} p(x) \log_b \Big(p(x) / q(x) \Big)$$

- Measures not the difference in the amount of information but difference in the information itself
- Both distributions have to be defined over the same domain
- Is always positive and zero only if the two distributions are identical

Joint Entropy

Joint Entropy of two variables is the entropy of the joint distribution

$$H(X,Y) = E\left[-\log_b P(X,Y)\right]$$
$$= -\sum_x \sum_y p(X = x, Y = y) \log(X = x, Y = y)$$

- Properties:
 - H(X,Y)≥max(H(X), H(Y)
 - H(X,Y)≤H(X)+H(Y)
 - H(X,Y)=H(X)+H(Y) if and only if X and Y are independent

Conditional Entropy

 Conditional Entropy (or conditional uncertainty) measures the information gained through Y if X is already known

$$H(Y|X) = E[-\log_b P(Y|X)] = E[-\log_b P(Y=y,X=x) + \log_b P(X=x)]$$

$$= \sum_{x} \sum_{y} P(X=x,Y=y) \left(-\log_b P(Y=y,X=x) + \log_b P(X=x)\right)$$

$$= -\sum_{x} \sum_{y} P(X=x,Y=y) \log_b P(Y=y,X=x) + \sum_{x} P(X=x) \log_b P(X=x)$$

$$= H(X,Y) - H(X)$$

© Manfred Huber 2017

4

Mutual Information

- aka transinformation
- The mutual information of two random variables X and Y is defined as:
 - I(X,Y) = H(X) H(X|Y)
 - I(X,Y)=H(Y)-H(Y|X)
- Alternatively:
 - I(X,Y) = H(X) + H(Y) H(X,Y)
- It answers the question: what is the uncertainty of X if we already know the outcome of Y (or viceversa).