Week 3 Information Measures - entropy

07 July 2020 17:57

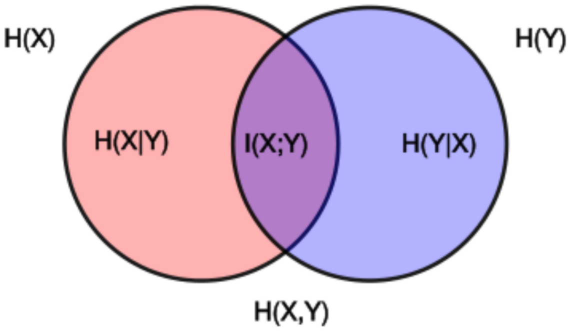
Entropy

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i) \log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

- where $I_X(x_i)$ is the self-information associated with particular outcome; I_X is the self-information of the random variable X in general, treated as a new derived random variable; and $\mathrm{E}[I_X]$ is the expected value of this new random variable, equal to the sum of the self-information of each outcome, weighted by the probability of each outcome occurring[3]; and b, the base of the logarithm, is a new parameter that can be set different ways to determine the choice of units for information entropy.
    - usual unit: bit (b=2)
    - information content of an event E: I(E)=-log2(p(E))
        - low probability -> high information content
- The self-information quantifies the level of information or surprise associated with one particular outcome or event of a random variable, whereas the entropy quantifies how "informative" or "surprising" the entire random variable is, averaged on all its possible outcomes.
    - higher entropy: more unpredictable
- Properties
    - determined entirely by the probability distribution of the data source, it is additive for independent sources
    - maximized at the uniform distribution = logn
    - minimized (and equal to zero) when there is 100% probability of only one event occurring/when outcome is known
    - obeys a certain derived version of the chain rule of probability
- Conditional entropy:
    - The conditional entropy of Y given X is defined as

      $$\boxed{\mathrm{H}(Y|X) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)} \qquad \text{(Eq.1)}}$$

      where $\mathcal{X}$ and $\mathcal{Y}$ denote the support sets of X and Y.
- mutual information: a measure of the mutual dependence between the two variables.
    - 

    - X, Y independence iff H(X,Y) = 0

Monty Hall problem

- $X \quad P(1/3, 1/3, 1/3) \quad H(X)$

  $\frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 + \frac{1}{3}\log_2 3 = \log_2 3$

  1.585 bits

  Y - what Monty told us

  $H(X|Y) \quad H(1/3, 2/3) = \frac{1}{3}\log_2 \frac{3}{1} + \frac{2}{3}\log_2 \frac{3}{2}$

  .918 bits

  $H(X) - H(X|Y) = I(X;Y) = 1.585 - .918 = .667$ bits
- information gain by opening one door = I(X;Y) = H(X) - H(X|Y)

Bayes thm

- transform fact -> knowledge
- Bayes does the work and entropy keeps the score
- probability distributions are carriers of information.
- Information gained in one toss of coin
  - Know that a coin is fair with probability 1/2, crooked (0.4 head, 0.6 tail)with probability half. One toss get head.
  - by bayes, P(fair coin | first toss is head) = 5/9
  - H(0.5, 0.5) - H(0.556, 0.444) = 0.0091 -information gain

Entropy in confusion matrix

Confusion Matrix				Test Classification **Y**			
				[optical scanner on assembly line]			
				"Positive"		"Negative"	
				0.3	c	0.7	d
Condition **X**	"+"	0.2	a	0.1	e	0.1	f
[defective computer chip]							
	"-"	0.8	b	0.2	g	0.6	h

H(X)	= a*log(1/a)	+ b*log(1/b)
0.7219	0.4644	0.2575

I(X;Y) =	H(X)	- H(X|Y)
0.0323	0.7219	0.6897

I(X;Y) =	H(Y)	- H(Y|X)
0.0323	0.8813	0.8490

H(Y)	= c*log(1/c)	+ d*log(1/d)
0.8813	0.5211	0.3602

I(X;Y) =	H(X)	+ H(Y)	- H(X,Y)
0.0323	0.7219	0.8813	1.5710

H(X,Y)	= e*log(1/e)	+ f*log(1/f)	+ g*Log(1/g)	+ h*log(1/h)
1.5710	0.3322	0.3322	0.4644	0.4422

0.1	e	0.10	f	0.20	g	0.60	h
0.06	ac	0.14	ad	0.24	bc	0.56	bd

Mutual Information I(X:Y) = Relative Entropy of Joint and Product Distributions --- D(p(X,Y||p(X)p(Y))							
= e*log(e/ac)	0.073696559	+ f*log(f/ad)	-0.048542683	+ g*log(g/bc)	-0.0526	+ h*log(h/bd)	0.05972
0.0323							

H(Y|X)	=	(a	*H(e/a, f/a))	+	(b	*H(g/b, h/b)
		0.2000	1.0000		0.8000	0.8113
0.8490						

H(X|Y)	=	(c	*H(e/c, g/c))	+	(d	*H(f/d, h/d)
		0.3000	0.9183		0.7000	0.5917
0.6897						