

“Information is the resolution of uncertainty.”

Shannon

Erfan Mirzaei

Information Theory Mini-Course

Sep 2022

Session2

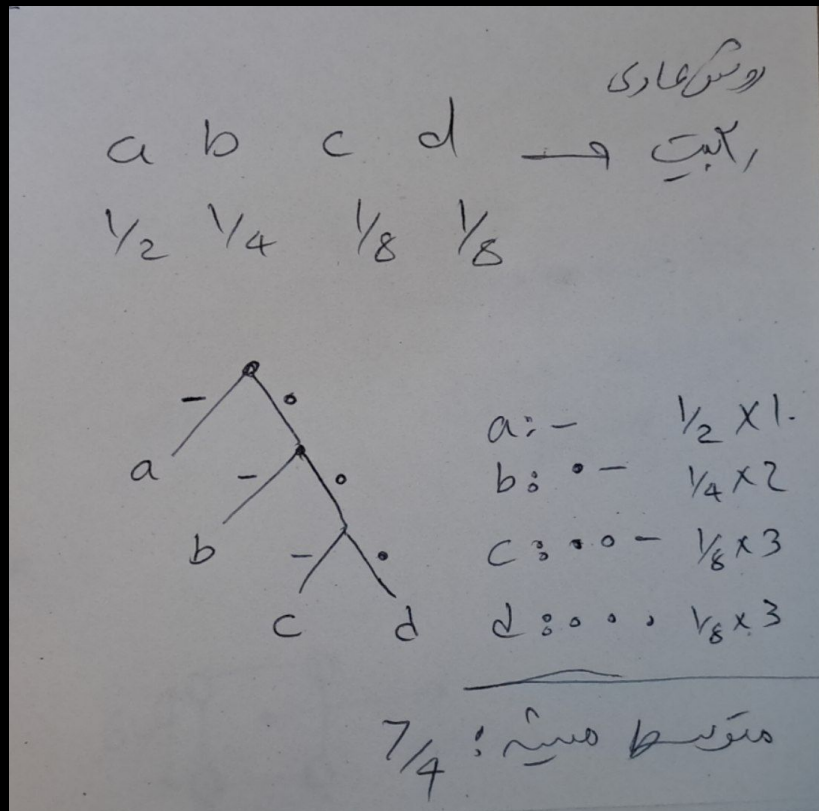
Joint, Conditional, Cross Entropy
Mutual Information

Recap

- Imagine a 4 sided dice and you want guess the outcome
Just by asking Y/N questions:
- A trivial but not efficient way is to ask exhaustively.
This takes $10/4$ questions in average.
- Another approach is that you assume that the toss is fair
and ask search for the answer by binary search.
This takes 2 questions in average. This is equal to entropy when the toss
is fair which is the max entropy among all random variables with the
same outcome space size.
- But imagine you know that the probability of numbers is as follows: $[1/2, 1/4, 1/8, 1/8]$
With knowing this and choosing good strategy for asking questions you
could find the answer with $7/4$ questions in average. [Equal to Entropy.]



Recap



characters should be sent =
stones should be thrown =
questions should be asked

This is the beauty of information Theory.

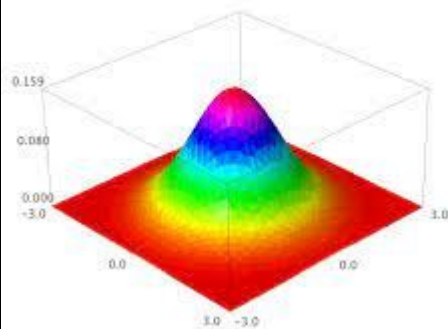
Joint Entropy

Have two random variables X and Y with joint probability distribution $P(X = x, Y = y) = P(x, y)$

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y),$$

$$H(X, Y) = -E \log p(X, Y).$$

$$p_{X,Y}(x, y) = P(X = x \text{ and } Y = y)$$



		X		
		1	2	3
Y	1	0	1/6	1/6
	2	1/6	0	1/6
	3	1/6	1/6	0

Conditional Entropy

Have two random variables X and Y with joint probability distribution $P(X = x, Y = y) = P(x, y)$

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X). \end{aligned}$$

First Chain rule

The naturalness of definition of joint and conditional entropy is exhibited:

(Chain rule)

$$H(X, Y) = H(X) + H(Y|X).$$

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

Corollary

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \quad (2.21)$$

Proof

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \quad (2.15)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \quad (2.16)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \quad (2.17)$$

$$= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \quad (2.18)$$

$$= H(X) + H(Y|X). \quad (2.19)$$

Equivalently, we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X) \quad (2.20)$$

and take the expectation of both sides of the equation to obtain the theorem. \square

What is the intuition?

When you have uncertainty about two random variables you can think about it in a sequential manner:

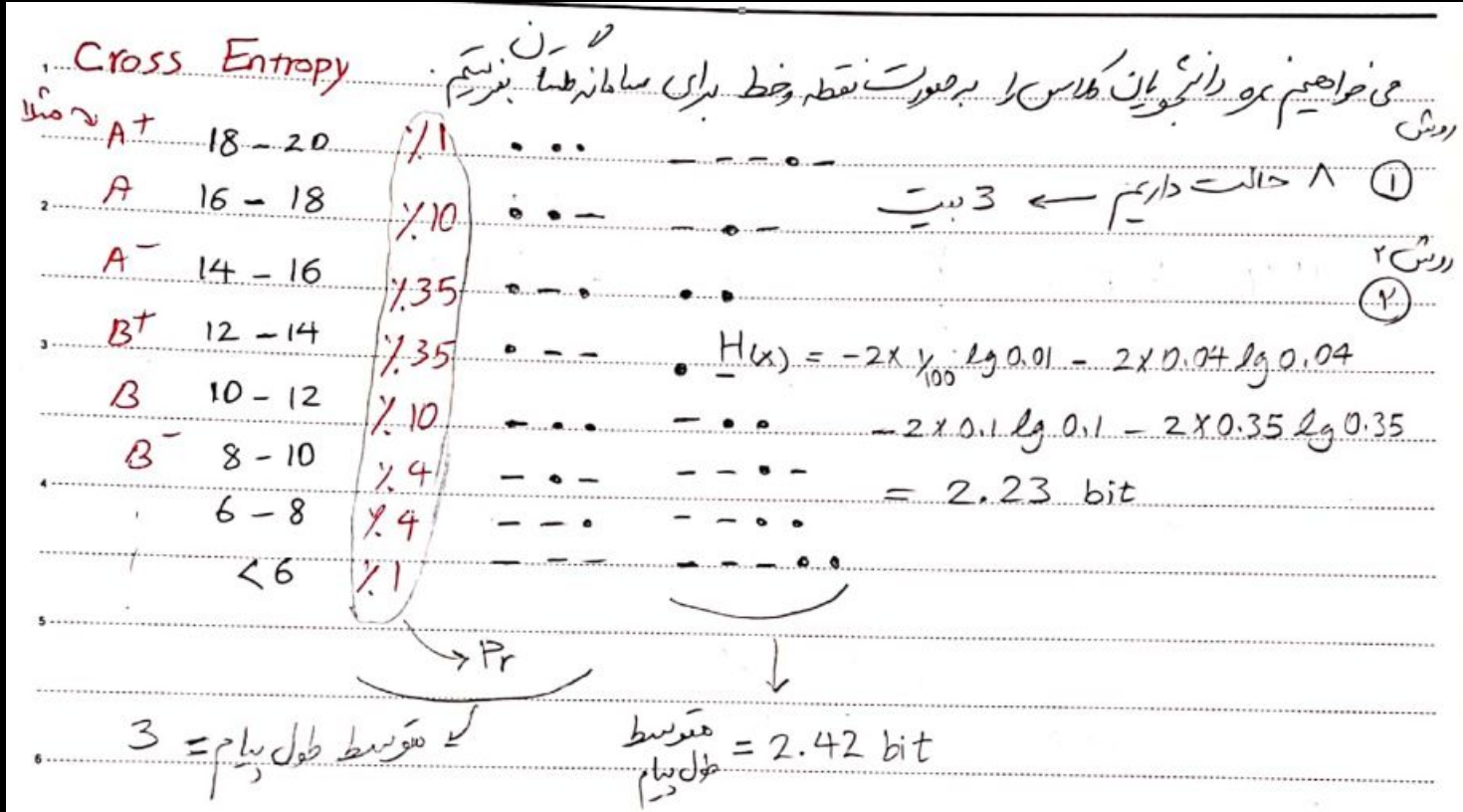
- $H(Y | X) \leq H(Y)$

When we know X , we can not have more uncertainty about Y than we know nothing. [Or revealing doesn't add more info]



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

Cross Entropy



Cross Entropy

اگر از کدش ۲ برای نماد ۵ پیام به از این کلاس استفاده کنیم

$\frac{1}{32}$	18 - 20	35%	5
$\frac{1}{8}$	16 - 18	4%	3
$\frac{1}{4}$	14 - 16	4%	2
$\frac{1}{4}$	12 - 14	1%	2
$\frac{1}{8}$	10 - 12	1%	3
$\frac{1}{6}$	8 - 10	10%	4
$\frac{1}{16}$	6 - 8	10%	4
$\frac{1}{32}$	< 6	35%	5

به طور متوسط هر بیت 4.58 بیت فشرده که از میان 2.23 بیت آن اطلاعات مفید است.

احتمال رخ دادن = $\frac{\text{تعداد کاراکتر در کدها}}{2}$

چون می‌دانیم مسئله را به این صورت در نظر بگیریم که ارسال نشده ی پیام باور اولی زهی این نیست به هر کلاس دانش و بدون منظور از این کدش استفاده کرده بنابراین احتمال هر یک از حالت ها در ذهنی آنها یک هم چنین چیزی بوده (جمع احتمال ها یک معانی است نشه)

Cross Entropy

Cross Entropy

P : توزیع واقعی

q : توزیع مفروض

$$H(p, q) = - \sum_{x \in X} p(x) \log q(x)$$

اگر توزیع مفروض با توزیع واقعی برابر باشد
Cross Entropy = Entropy

$$H(p) = H(x) \rightarrow 2.23 \text{ bit}$$

$$H(p, q) \rightarrow 4.58 \text{ bit}$$

خطای مدل در تخمین احتمالات (یعنی q به جای p)
ما یک سده که representation نویسی داشته باشیم
و این باعث افزایش مابقی (اطلاعات) سده.

Cross Entropy Loss

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

“Minimizing this KL divergence corresponds exactly to minimizing the cross-entropy between the distributions.”

Page 132, Deep Learning, 2016

[Reference](#)

Any loss consisting of a negative log-likelihood is a cross-entropy between the empirical distribution defined by the training set and the probability distribution defined by model. For example, mean squared error is the cross-entropy between the empirical distribution and a Gaussian model.

Relative Entropy OR KL Divergence

$$\begin{aligned} H(p||q) - H(p) &= - \sum p(x) \lg q(x) + \sum p(x) \lg p(x) \\ &= \sum p(x) \lg \frac{p(x)}{q(x)} = D(p||q) \text{ Relative Entropy} \\ &\quad \downarrow \\ &\quad \text{Kullback-Liebler Divergence (KL divergence)} \end{aligned}$$

اختلاف بین انتروپی و اِنترنل

$$D(p||q) \neq D(q||p)$$

- How much p and q as two probability distribution is similar to each other?
- How much we had more uncertainty just because we don't know the probability distribution. We sent longer messages. And revealing the outcome was more surprising to us and gave us more information.

Mutual Information

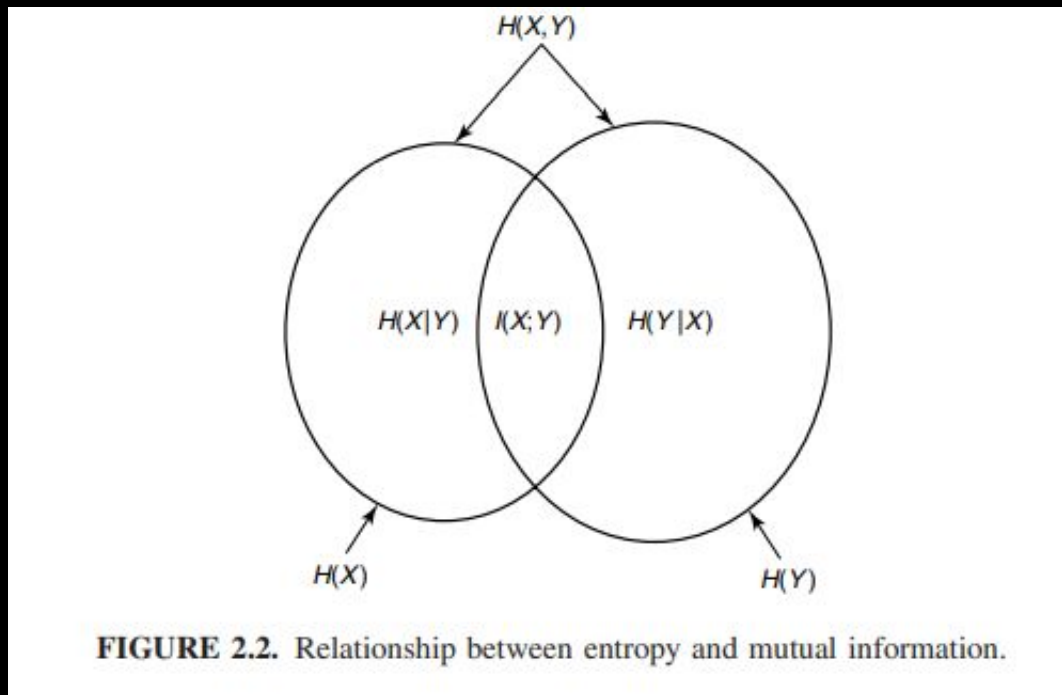
Have two random variables X and Y with joint probability distribution $P(X = x, Y = y) = P(x, y)$

- Relative Entropy or KL-Divergence between $P(x,y)$ and $P(x)*P(y)$

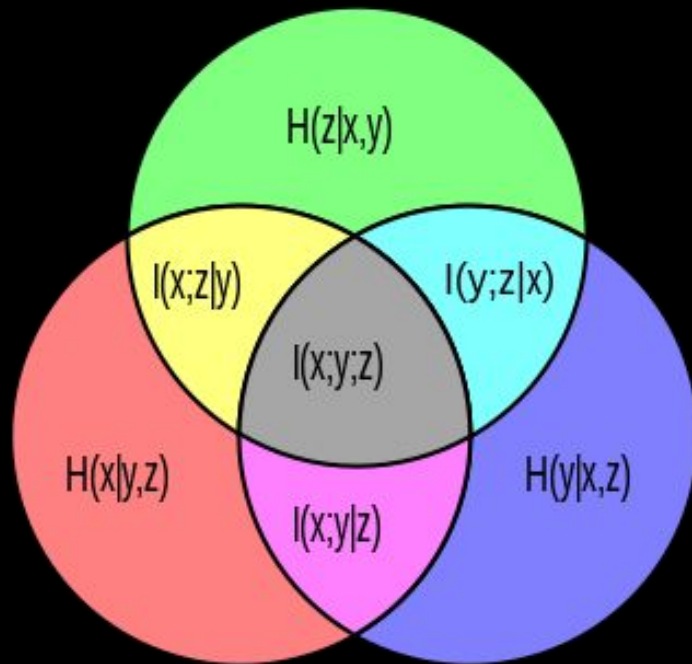
$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= D(p(x, y) || p(x)p(y)) \\ &= E_{p(x,y)} \log \frac{p(X, Y)}{p(X)p(Y)}. \end{aligned}$$

- $I(X ; Y) = I(Y ; X)$, $I(X ; X) = H(X)$
- $I(X ; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) = H(X) + H(Y) - H(X, Y)$

Mutual Information and Entropy



Mutual Information and Entropy



Chain Rules

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1) \quad (2.52)$$

$$= \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1). \quad \square \quad (2.53)$$

$$I(X_1, X_2, \dots, X_n; Y) = H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n|Y) \quad (2.63)$$

$$\begin{aligned} &= \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1) - \sum_{i=1}^n H(X_i|X_{i-1}, \dots, X_1, Y) \\ &= \sum_{i=1}^n I(X_i; Y|X_1, X_2, \dots, X_{i-1}). \quad \square \quad (2.64) \end{aligned}$$

Thanks for your attention

Erfan Mirzaei
erfunmirzaei@gmail.com