A CRASH COURSE IN INFORMATION THEORY

Have a message x that can be one of $x_1, x_2, \ldots, x_n$ with some probability distribution.

How can we compress messages as well as possible, i.e. encode x_i as codeword c_i, so that expected # bits needed is minimized?

Ex 1 Horse race with 8 horses. Suppose a priori all horses equally likely to win.
Number horses from 0 to 7 and let c_i be binary expansion of i $\quad c_0 = 000, c_1 = 001, c_2 = 010$ etc
$c_7 = 111$

3 bits needed — seems hard to do better

Ex 2 Suppose we know more about how probable horses are to win

Horse	Winning prob
0	$1/2 = p_0$
1	$1/4 = p_1$
2	$1/8$
3	$1/16$
4	$1/64$
5	$1/64$
6	$1/64$
7	$1/64 = p_7$

Use this to encode horses

c_0	0
c_1	10
c_2	110
c_3	1110
c_4	111100
c_5	111101
c_6	111110
c_7	111111

Now the expected length is $\sum_{i=0}^{7} p_i \cdot \text{length}(c_i)$

$$= \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + 4\cdot\frac{1}{64}\cdot 6 = 2$$

One bit less! Is this optimal?
Need to measure "amount of randomness" in x
"How much information" x conveys

Some notation:

X random variable

Ranges over universe $\mathcal{U} = \{x_1, x_2, \ldots, x_n\}$

$p(x) = \Pr[X = x]$ probability that the message is $x \in \mathcal{U}$ (w.r.t some probability distribution that should be clear from context

Def 3 (SHANNON) ENTROPY

$$\mathcal{H}(X) = \sum_{x \in \mathcal{U}} p(x) \cdot \log\left(\frac{1}{p(x)}\right)$$

$$= -\sum_{x \in \mathcal{U}} p(x) \log p(x)$$

Convention $0 \log 0 = 0$ (makes sense by continuity)

Note that entropy is function of distribution only

Doesn't depend on actual values in $\mathcal{U}$, only on probabilities

Thm 4 Source coding theorem [Shannon '48]

Let l_i be length of codeword c_i in best encoding of X and let $\mathcal{L} = \sum_{i=1}^{n} p(x_i) \cdot l_i$ be the expected length of the encoding of a message.

Then

$$\mathcal{H}(X) \leq \mathcal{L} < \mathcal{H}(X) + 1$$

~~EXERCISE FROM THIS~~

So in particular solutions in Ex 1 & 2 are optimal.

Skip the proof of Thm 4

Ex 1 $8 \cdot -\left(\frac{1}{8} \log \frac{1}{8}\right) = \log 8 = 3$

Ex 2 $\frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{16} \log 16 + 4 \cdot \frac{1}{64} \log 64 = 2$

Observations 5

(a) $H(X) \geq 0$

(b) $H(X) = 0$ iff X is constant [no uncertainty]

(c) $H(X)$ is maximized for X uniformly distributed over $\mathcal{U}$.

[5(c) requires a proof, of course. Return to this later.]

Ex 6 An important special case is

$$X = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}$$

Then $H(X) = -p \log p - (1-p)\log(1-p)$

This quantity is denoted $H(p)$

Maximized when $p = 1/2$, i.e., when X is an unbiased coin flip.

Obs 7 The entropy of X can be interpreted as the expected value of the random variable $Z = \log \frac{1}{p(X)}$. That is

$$H(X) = E\left[\log \frac{1}{p(X)}\right] = E\left[-\log p(X)\right]$$

Proof Immediate from definition.

For a pair of (discrete) random variables X, Y over $\mathcal{U}_X$ and $\mathcal{U}_Y$ we write

$$p(x,y) = \Pr[X = x \text{ and } Y = y]$$

Def 8 JOINT ENTROPY

The joint entropy of X and Y is the entropy of their joint distribution, i.e.

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

By Obs 7 we have

$$\mathcal{H}(X,Y) = \mathbb{E}\left[-\log p(X,Y)\right]$$

Conditional entropy measures expected value of entropy conditioning on some other random variable

Def 9 CONDITIONAL ENTROPY

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

$$= -\sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

$$= -\sum_x \sum_y p(x,y) \log p(y|x)$$

$$= \mathbb{E}\left[-\log p(Y|X)\right]$$

[Note that $p(Y|X)$ is just the random variable that takes value $p(y|x) = \Pr[Y=y \mid X=x]$ with probability $p(x,y) = \Pr[X=x \text{ and } Y=y]$.]

The entropy of a pair of random variables is the entropy of one plus conditional entropy of the other

Thm 10 Chain rule

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

Proof $p(x,y) = p(x) \cdot p(y|x)$

Taking logarithms

$$\log p(x,y) = \log p(x) + \log p(y|x)$$

But then $$E[\log p(x,y)] = E[\log p(x)] + E[\log p(y|x)]$$

or $$-H(X,Y) = -H(X) - H(Y|X)$$ ∎

Cor 11

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

The mutual information of two random variables is the reduction ~~of~~ in the uncertainty of one variable due to the knowledge of the other.

Def 12 MUTUAL INFORMATION

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

Intuitively, $I(X;Y)$ measures how much Y tells us about X.

Prop 13

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

(b) $I(X;Y) = I(Y;X)$

Proof: 13(a) follows by the chain rule, and the symmetry in 13(b) is an immediate consequence of 13(a)

The relative entropy ~~##~~ is a measure of the distance between two distributions. Also known as:

- Kullback-Leibler divergence
- Kullback-Leibler distance
- information divergence

DEF 14 RELATIVE ENTROPY or KULLBACK-LEIBLER DIVERGENCE [Kullback & Leibler '51]

p and q two distributions / probability mass functions over $\mathcal{U}$.

$$D(p\|q) = \sum_{x \in \mathcal{U}} p(x) \log \frac{p(x)}{q(x)}$$

$$= E\left[\log \frac{p(x)}{q(x)}\right]$$

Conventions: $0 \log 0/0 = 0$

$0 \log 0/q = 0$

$p \log p/0 = \infty$ if $p > \infty$

Hence, if there is an x s.t. $p(x) > 0$, $q(x) = 0$, then $D(p\|q) = \infty$

Will show later:

$D(p\|q)$ always non-negative

$D(p\|q) = 0$ iff $p = q$

<u>But</u> not symmetric

does not satisfy triangle inequality

Hence not a "distance" in the mathematical sense.

The mutual information of X and Y is the KL-divergence between the joint distribution (X, Y) and the product of the marginal distributions.

<u>Intuition</u>: The further away X and Y are from being independent, the more info they provide about each other.

<u>Thm 15</u> Let $p(x,y)$ be joint distribution on X, Y with marginal distributions $p(x)$ and $p(y)$. Then

$$D(p(x,y) \| p(x)p(y)) = I(X;Y).$$

<u>Proof</u>

skip in class

$$D(p(x,y)\|p(x)p(y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y)\right)$$

$$= H(X) - H(X|Y) = I(X;Y)$$ ☐

The conditional mutual info $I(X;Y|Z)$ is the reduction in uncertainty of X due to knowledge of Y when Z is given

Def 16 CONDITIONAL MUTUAL INFORMATION

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

Obs 17

$$I(X;Y|Z) = \mathop{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right]$$

Thm 18 Chain rule for entropy (generalized)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \ldots + H(X_n | X_{n-1}, \ldots, X_1)$$

Proof Exercise.

Thm 19 Chain rule for information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)$$

Proof Expand $I(X_1, X_2, \ldots, X_n; Y)$ according to definition and then use chain rule in Thm 18

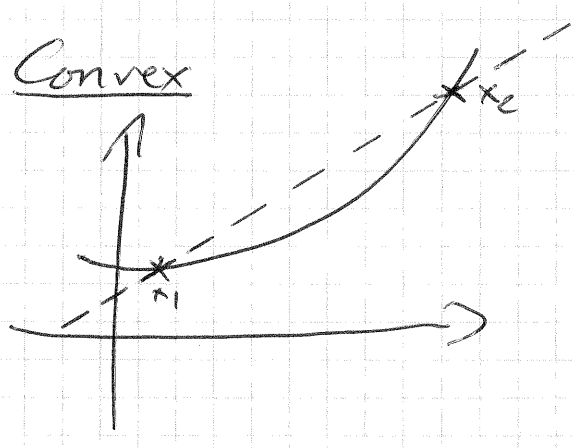
Def 20 $f: \mathbb{R} \to \mathbb{R}$ is CONVEX over interval (a,b) if for all $x_1, x_2 \in (a,b)$, $0 \leq \lambda \leq 1$ it holds that

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

f is STRICTLY CONVEX if equality holds only for $\lambda = 0$ and $\lambda = 1$.

f is (STRICTLY) CONCAVE if $-f$ is (strictly) convex

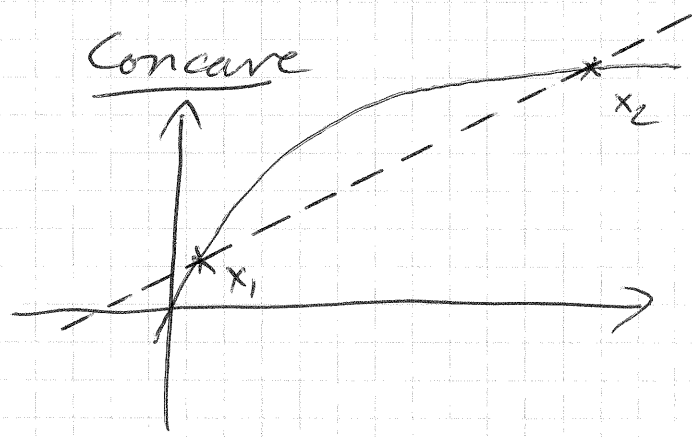
Convex



$f(x) = x^2$

$f(x) = e^x$

Concave



$f(x) = \log(x)$ $x > 0$

$f(x) = \sqrt{x}$ $x \geq 0$

Lem 21

If f has second derivative f'' that is non-negative (strictly positive) over (a,b), then f is convex (strictly convex) over (a,b)

Proof Exercise in standard calculus using Taylor expansion. Won't do it now.

Thm 22 JENSEN'S INEQUALITY

If f is a convex function and X is a random variable, then

$$\mathcal{E}[f(X)] \geq f(\mathcal{E}[X]) \qquad (*)$$

Moreover, if f is strictly convex, then equality in $(*)$ implies that $X = \mathcal{E}[X]$ with probability 1, i.e., X is constant.

Obviously get analogous version with inequality in $(*)$ turned around for concave functions.

Proof. By induction on # elements in distribution.

(skip in class) Suppose $|\mathcal{U}| = 2$, $\mathcal{U} = \{x_1, x_2\}$. Then $(*)$ becomes

$$p(x_1) f(x_1) + p(x_2) f(x_2) \geq f(p(x_1)\cdot x_1 + p(x_2) x_2)$$

which follows by def of convexity.

Write $p(x_i) = p_i$ for brevity and assume $(*)$ holds for distributions with $k-1$ elements.

Let $p_i' = p_i / (1-p_k)$ for $i = 1, 2, \ldots, k-1$.

Then

$$\mathcal{E}[f(X)] = \sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i)$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \quad \text{[by IH]}$$

$$\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) \quad \text{[by convexity]}$$

$$= f\left(\sum_{i=1}^{k} p_i x_i\right)$$

Strictly convex part left to the reader.

Remark Jensen's inequality also holds for continuous distributions.

Now we will use Jensen's inequality to prove (or claim) a number of important facts

Thm 23 (Information inequality)
Let $p(x)$, $q(x)$, $x \in \mathcal{U}$, be two distributions/probability mass functions. Then

$$D(p \| q) \geq 0$$

with equality iff $p = q$.

Proof Let $A = \{x \mid p(x) > 0\}$ be the support of p. Then

(skip in class)

$$-D(p\|q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \quad [\text{by def}]$$

$$= \sum p(x) \log\left(\frac{q(x)}{p(x)}\right) \quad \left[= \mathbb{E}\left[\log \frac{q(X)}{p(X)}\right]\right]$$

$$(\dagger) \leq \log\left(\sum p(x) \frac{q(x)}{p(x)}\right) \quad [\text{Jensen on concave function}]$$

$$= \log\left(\sum_{x \in A} q(x)\right)$$

$$(\ddagger) \leq \log\left(\sum_{x \in \mathcal{U}} q(x)\right)$$

$$= \log(1)$$

$$= 0$$

Since $\log(t)$ is strictly concave for $t > 0$

~~Hat~~

The inequality in (†) is strict unless $q(x)/p(x)$ is constant i.e. $c\,q(x) = p(x)$

$$\sum_{x\in A} q(x) = c\sum_{x\in A} p(x) = c$$

Equality in (†) only if $\sum_{x\in A} q(x) = \sum_{x\in U} q(x) = 1$, which implies $c = 1$ ∎

** **Corollary 24** Nonnegativity of mutual information

$I(X;Y) \geq 0$ with equality iff X and Y are independent.

Cor 25

$$I(X;Y|Z) \geq 0.$$

Thm 26 $H(X) \leq \log(|\mathcal{U}|)$ with equality iff X is uniformly distributed.

Proof Let $u(x) = 1/|\mathcal{U}|$ uniform distribution over $\mathcal{U}$. Then (skip in class)

$$D(p\|u) = \sum p(x) \log \frac{p(x)}{u(x)}$$

$$= \sum p(x)\log p(x) - \sum p(x) \log\left(\frac{1}{1/|\mathcal{U}|}\right)$$

$$= -H(X) + \log|\mathcal{U}|$$

Since $D(p\|u) \geq 0$ we get $H(X) \leq \log|\mathcal{U}|$, and equality holds only if $p = u$

) **Proof of Cor 24 $I(X;Y) = D(p(x,y)\,\|\,p(x)p(y)) \geq 0$ with equality if $p(x,y) = p(x)p(y)$ i.e. independence

Thm 27 Conditioning reduces entropy
(a.k.a. Information can't hurt)

$H(X|Y) \leq H(X)$ with equality iff X and Y are independent

Proof: $H(X) - H(X|Y) = I(X;Y) \geq 0$.

Note that Thm 27 is true only on average. For a specific y, we can have

$$H(X|Y=y) > H(X),$$

but on average

$$H(X|Y) = \sum_y p(y) H(X|Y=y)$$
$$\leq H(X).$$

Thm 28 Subadditivity of entropy

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$$

with equality iff all the X_i are independent.

Proof By the chain rule and Thm 27 we have

$$H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \ldots, X_1)$$
$$\leq \sum_{i=1}^n H(X_i)$$

The fact that log is a concave function implies an inequality that is useful for proving some further results about entropy. We won't prove them now, but state the inequality for reference.

Thm 29 Log sum inequality

For nonnegative $a_1, \ldots, a_n, b_1, \ldots, b_n$ it holds that

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \right)$$

with equality iff all a_i / b_i are constant.

Proof Assume w log $a_i > 0$ $b_i > 0$.

(skip in class) Let $f(t) = t \log t$. Then $f''(t) = \frac{1}{t} \log e > 0$ for $t > 0$ so f is strictly convex.

By Jensen's inequality

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i \right) \quad (***)$$

for $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$.

Let $\alpha_i = b_i / \sum_{j=1}^{n} b_j$ and $t_i = \frac{a_i}{b_i}$.

Then

$$\sum_i \frac{b_i}{\sum_{j=1}^{n} b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_i \frac{a_i}{\sum b_j} \log \frac{a_i}{b_i} \geq \left[\text{by } *** \right]$$

$$\geq \sum_i \frac{a_i}{\sum b_j} \log \sum_i \frac{a_i}{\sum b_j}$$

~~Dividing~~ Multiplying by $\sum_{j=1}^{n} b_j$ yields the log sum inequality. ☐

DATA PROCESSING INEQUALITY

Suppose we have an input X
We code this as a message Y, which is intended to convey information about X.

If then someone does a clever transformation of Y into Z based on Y only and not on X, this can never increase the information.

Thm 30 Data processing inequality

If random variables X and Z are conditionally independent given Y, then

$$I(X;Y|Z) \leq I(X;Y)$$
$$I(X;Z) \leq I(X;Y)$$

Proof By the chain rule applied twice, we have

$$I(X;(Y,Z)) = I(X;Z) + I(X;Y|Z) \quad (1)$$

$$I(X;(Y,Z)) = I(X;Y) + I(X;Z|Y) \quad (2)$$

Since $I(X;Z|Y) = 0$ by assumption, putting (1) and (2) together we get

$$I(X;Y) = I(X;Z) + I(X;Y|Z)$$

Since information is non-negative, the theorem follows.