LECTURE 02

01-10-2024

FANO ENCODING

ENTROPY AND NUMBER OF BITS

LABEL INVARIANCE

$H(f(x))$

JOINT ENTROPY

$H(X,Y) = H(X) + H(Y)$ FOR S.I. X Y

$H(X,Y) \leq H(X) + H(Y)$

CONDITIONAL ENTROPY

CHAIN RULE

MEANING OF $H(X|Y)$

$0 \leq H(X|Y) \leq H(X)$

INFORMATION GAIN

INTERPRETATION - FIGURE

FORMULAS

KULLBACK-LEIBLER DIVERGENCE

MEANING OF ENTROPY

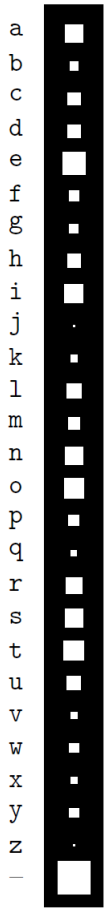
THE AMOUNT OF INFORMATION

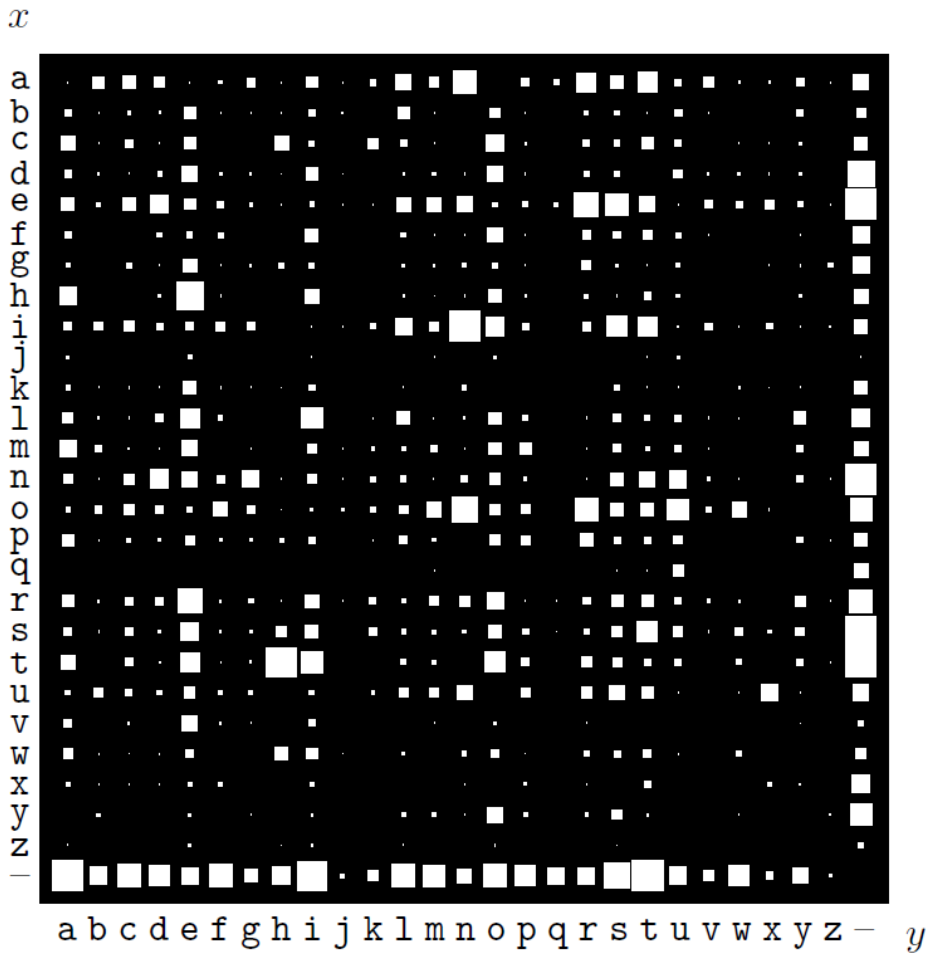
WE RECEIVE WHEN

WE OBSERVE AN EVENT

EXAMPLES OF PROBABILITY DISTRIBUTIONS

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928





SINGLE LETTER

TWO CONSECUTIVE LETTERS

FANO ENCODING

$X \qquad \Omega_X = \{ x_0 \quad x_i \quad x_n \}$

$P_X = \{ p_0 \quad p_i \quad p_n \}$

WE COULD USE FIXED-LENGTH BINARY CODES TO REPRESENT THE DIFFERENT OUTCOMES

EX $\Omega_X = \{ x_0 \quad x_1 \quad x_2 \quad x_3 \}$

$\qquad\qquad\quad 00 \quad 01 \quad 10 \quad 11$

IN THIS WAY WE DO NOT TAKE INTO ACCOUNT THEIR DIFFERENT PROBABILITIES.

IDEA: USE VARIABLE LENGTH CODES (BINARY CODES ASSOCIATED TO DIFFERENT OUTCOMES MAY HAVE DIFFERENT LENGTH)

THIS WAY WE COULD ASSOCIATE SHORT CODES TO OUTCOMES WITH HIGH PROBABILITIES AND LONG CODES TO THOSE WITH LOW PROBABILITIES.

GOAL: REDUCE THE AVERAGE NUMBER $\bar{n}$ OF BITS NECESSARY TO REPRESENT THE OUTCOMES → ADVANTAGE:

WE HAVE LESS BITS TO STORE, TRANSMIT, PROCESS..

THE OPTIMAL ENCODING ALGORITHM IS THE HUFFMAN ALGORITHM (SECTION 2).

WE PRESENT THE SIMPLER FANO ALGORITHM

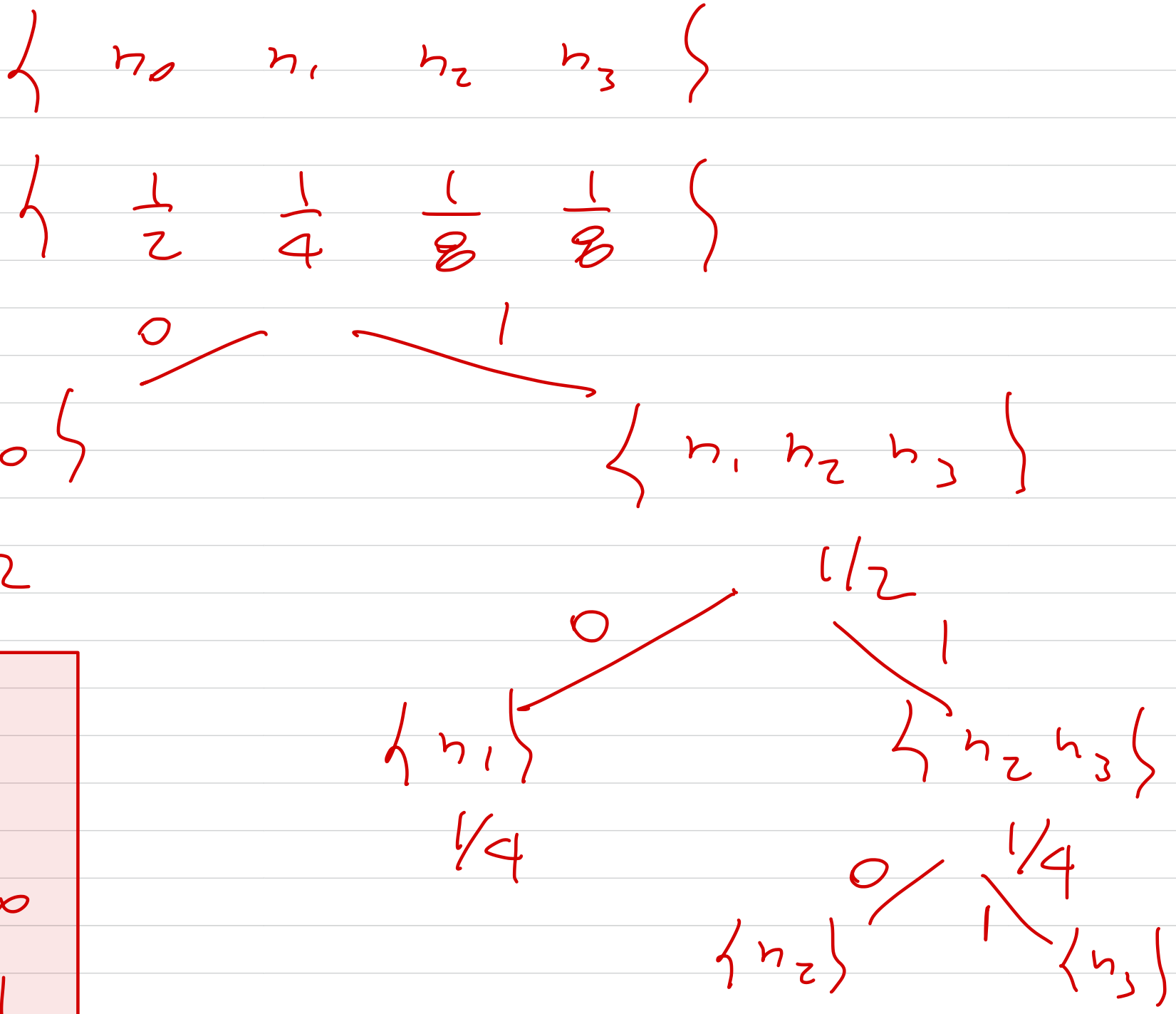
$$\Omega_x = \{ n_0 \quad n_1 \quad n_2 \quad n_3 \}$$

$$P_x = \{ \frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{8} \}$$

FANO'S ALGORITHM

ASSIGN THE BINARY CODES BY USING P_x

WE BUILD A TREE BY DIVIDING AT EACH STEP THE SET OF OUTCOMES INTO 2 SETS WITH CLOSE PROBABILITIES



n_0	0
n_1	1 0
n_2	1 1 0
n_3	1 1 1

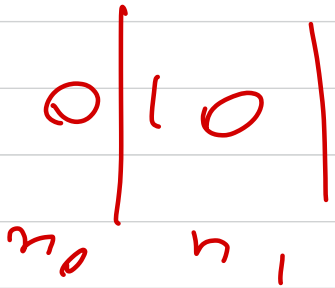
x_0 0 $P_0 = 1/2$

x_1 10 $P_1 = 1/4$

x_2 110 $P_2 = 1/8$

x_3 111 $P_3 = 1/8$

THE CODES SATISFY THE PREFIX RULE: A CODE IS NOT THE CONDITION FIRST PART OF ANY OTHER CODE



AT THE RX SIDE GIVEN THE BIT SEQUENCE IT IS POSSIBLE TO RECOVER THE SYMBOL SEQUENCE WITHOUT AMBIGUITY

n_0 0 $P_0 = 1/2$

n_1 10 $P_1 = 1/4$

n_2 110 $P_2 = 1/8$

n_3 111 $P_3 = 1/8$

AVERAGE NUMBER OF BITS

$$\bar{m} = \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{8}(3)$$

$1/2$ $1/2$ 0.75

$= 1.75$

ENTROPY

x_0	x_1	x_2	x_3
$1/2$	$1/4$	$1/8$	$1/8$

$$H(x) = \sum_i p_i \log_2 \frac{1}{p_i} =$$

$$= \frac{1}{2}\overbrace{\log_2 2}^{1} + \frac{1}{4}\overbrace{\log_2 4}^{2} + \frac{1}{8}\log_2 8 + \frac{1}{8}\underbrace{\log_2 8}_{3}$$

$$= 1.75$$

IN THIS SPECIFIC EXAMPLE

(WHERE THE PROBABILITIES HAVE THESE PARTICULAR VALUES) WE OBTAIN

THAT

$$\overline{m} = H(x)$$

IN GENERAL WE CAN OBTAIN A CODE WITH

$$H(x) \leq \overline{m} \leq H(x) + 1$$

IMPORTANT

So THE ENTROPY IS VERY CLOSE TO THE AVERAGE # OF BITS WE NEED TO REPRESENT THE OUTCOMES OF OUR RANDOM VARIABLE

LABEL INVARIANCE

$$X \qquad\qquad Y$$

$$\Omega_X = \{ x_1 \quad x_i \quad x_M \} \qquad \Omega_Y = \{ y_1 \quad y_i \quad y_M \}$$

$$P_X = \{ p_1 \quad p_i \quad p_M \} \qquad P_Y = \{ p_1 \quad p_i \quad p_M \}$$

$$I_X = I_Y$$

$$H = \sum_i p_i \log_2 \frac{1}{p_i}$$

TWO RANDOM VARIABLES WITH DIFF. ALPHABET BUT SAME PROB. HAVE SAME ENTROPY

FUNCTION OF A RANDOM VARIABLE

$$X \qquad \Omega_X = \{ x_1 \quad x_i \quad x_n \}$$

$$P_X = \{ p_1 \quad p_i \quad p_n \}$$

$$H(x) = \sum_i p_i \log_2 \frac{1}{p_i}$$

$$Y = f(x) \qquad \Omega_Y = \{ f(x_1) \quad f(x_i) \quad f(x_n) \}$$

$$H(Y) \stackrel{?}{=} H(X)$$

$n_1 \quad n_2$

$F(n_1) = F(n_2)$

IF F IS INJECTIVE (ONE-TO-ONE, NO COLLISIONS)

$$H(F(x)) = H(x)$$

OTHERWISE

$$H(F(x)) \leq H(x)$$

IF f IS NOT INJECTIVE $\quad H(X) \geq H(f(X))$

PROOF

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} \geq (p_1+p_2) \log_2 \frac{1}{p_1+p_2}$$

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} - (p_1+p_2) \log_2 \frac{1}{p_1+p_2} =$$

$$= p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} - p_1 \log_2 \frac{1}{p_1+p_2}$$

$$- p_2 \log_2 \frac{1}{p_1+p_2} =$$

$$= p_1 \log_2 \frac{p_1+p_2}{p_1} + p_2 \log_2 \frac{p_1+p_2}{p_2}$$

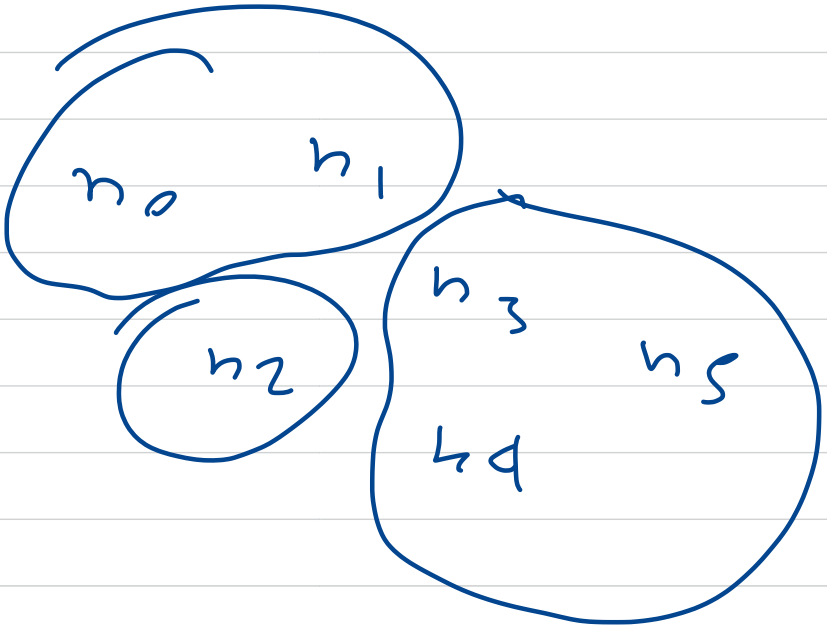
$$= p_1 \underbrace{\log_2 \frac{p_1+p_2}{p_1}}_{\geq 0} + \underbrace{p_2}_{\geq 0} \underbrace{\log_2 \frac{p_1+p_2}{p_2}}_{\geq 0} \geq 0$$

THEN

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} \geq (p_1+p_2) \log_2 \frac{1}{p_1+p_2}$$

CONSEQUENCES

IF THE FUNCTION IS NOT INJECTIVE THE ENTROPY (AMOUNT OF INFO.) DECREASES



EXAMPLE:

CLUSTERING ALWAYS DECREASES AMOUNT ON INFO.

JOINT ENTROPY

$X \quad \Omega_X$

$Y \quad \Omega_Y$

X \ Y	Temp < 25	Temp ≥ 25
Sunny	0.4	0.2
Cloudy	0.35	0.05

$(x, y) \quad \Omega_X \times \Omega_Y$

$$H(x,y) = \sum_{xy} p(x,y) \log_2 \frac{1}{p(x,y)}$$

RANGE

$$0 \leq H(X,Y) \leq \log_2 |\Omega_X||\Omega_Y|$$

EXAMPLES

(1)

0.4	0.2
0.35	0.05

$H(X,Y) =$

(2)

0.7	0.1
0.1	0.1

$H(X,Y) =$

(3)

0.25	0.25
0.25	0.25

$H(X,Y) =$

(4)

0.97	0.01
0.01	0.01

$H(X,Y) =$

(5)

0.5	0.5
0.0	0.0

$H(X,Y) =$

EXAMPLES

0.4	0.2
0.35	0.05

$H(X,Y) = 1.74$

0.7	0.1
0.1	0.1

$H(X,Y) = 1.35$

0.25	0.25
0.25	0.25

$H(X,Y) = 2$

0.97	0.01
0.01	0.01

$H(X,Y) = 0.24$

0.5	0.5
0.0	0.0

$H(X,Y) = 1$

LINK BETWEEN $H(X,Y)$ AND $H(X)$ $H(Y)$

X AND Y ARE STAT. IND. IFF $p(x,y) = p(x)\,p(y)$ $\forall$ PAIR (x,y)

IF X AND Y ARE S.I. $H(X,Y) = H(X) + H(Y)$

IF X AND Y ARE NOT S.I. $H(X,Y) \leq H(X) + H(Y)$

MEANING

IF X AND Y ARE NOT ST. IND.

THEY ARE LINKED SOME HOW

SO WHEN WE OBSERVE

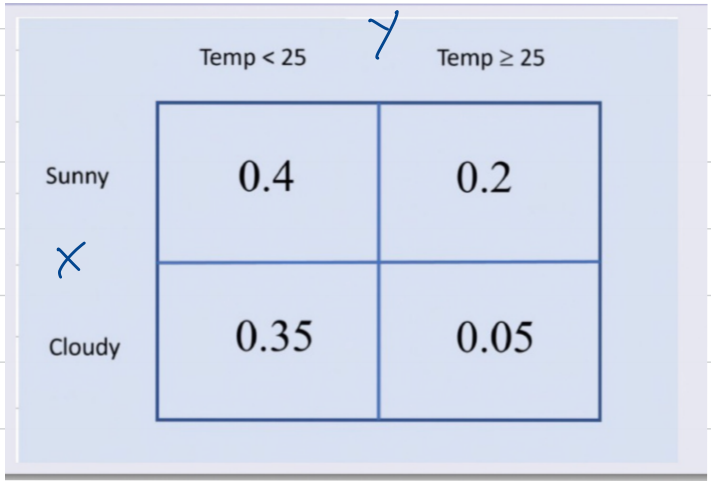
THEN TOGETHER THE

AMOUNT OF INF. IS

REDUCED

PROOF: MARGINALIZATION

$$\sum_{y} P(x, y) = P(x)$$



$$P(\text{SUNNY}) = P(\text{SUNNY}, \text{TEMP} < 25) + P(\text{SUNNY}, \text{TEMP} \geq 25)$$

$$= 0.4 + 0.2 = 0.6$$

$$P(\text{CLOUDY}) = P(\text{CLOUDY}, \text{TEMP} < 25) + P(\text{CLOUDY}, \text{TEMP} \geq 25) = 0.4$$

MARGINALIZATION PROPERTY

$$\sum_{x,y} p(x,y) \cdot f(x) = \sum_{x} f(x) \sum_{y} p(x,y)$$

$$= \sum_{x} f(x)\, p(x)$$

	Temp < 25	Temp ≥ 25
Sunny	0.4	0.2
Cloudy	0.35	0.05

$$p(S, T<25) \cdot \log_2 \frac{1}{p(S)} + P(S, T \geq 25) \log_2 \frac{1}{p(S)} +$$

$$+ P(C, T<25) \log_2 \frac{1}{p(C)} + P(C, T \geq 25) \log_2 \frac{1}{p(C)}$$

$$= \log_2 \frac{1}{p(S)} \cdot p(S) + \log_2 \frac{1}{p(C)} \cdot p(C)$$

X AND Y STATISTICALLY INDEPENDENT

$$H(x,y) = H(x) + H(y)$$

PROOF

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

$$H(x,y) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

$$= \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(x)\,p(y)}$$

$$= \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(x)} +$$

$$+ \sum_{x,y} p(x)\,p(y) \log_2 \frac{1}{p(y)}$$

→ MARGINALIZATION

$$= \sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_y p(y) \log_2 \frac{1}{p(y)}$$

$$= H(x) + H(y)$$

→ $$H(x, y) = H(x) + H(y)$$

NO H. STATISTICALLY INDEPENDENT

$$H(x,y) \leq H(x) + H(y)$$

PROOF

$$H(x,y) - H(x) - H(y) =$$

$$= \sum_{xy} p(x,y) \log_2 \frac{1}{p(x,y)} - \sum_x p(x) \log_2 \frac{1}{p(x)}$$

$$- \sum_y p(y) \log_2 \frac{1}{p(y)}$$

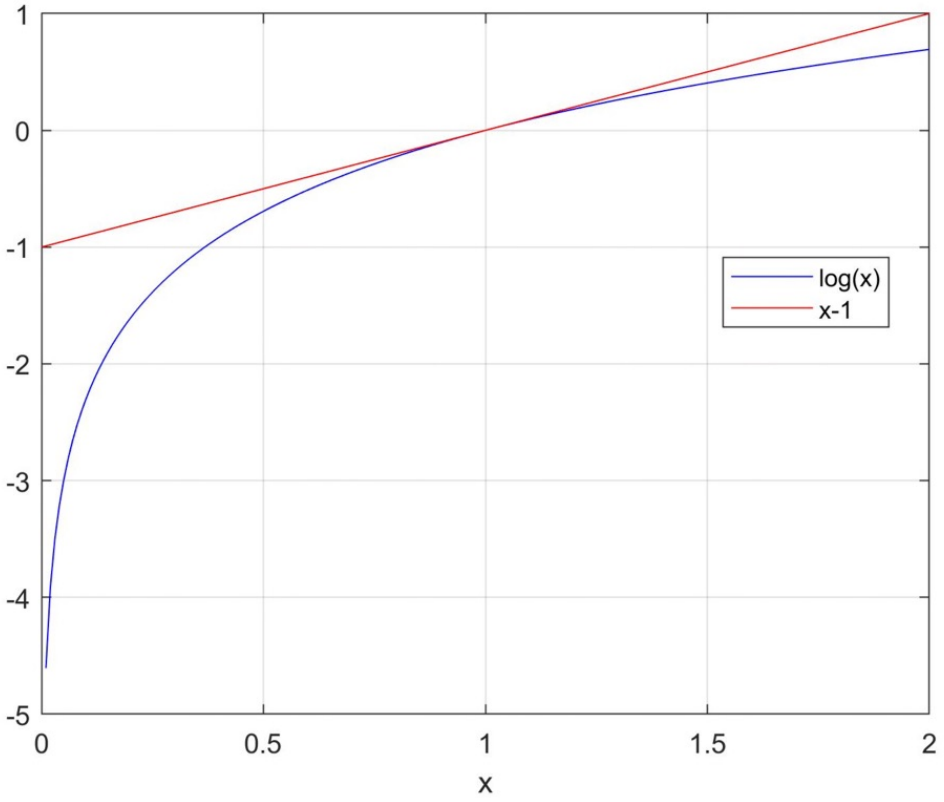
$$= \sum_{xy} p(x,y) \log_2 \frac{1}{p(x,y)} - \sum_{xy} p(x,y) \log_2 \frac{1}{p(x)}$$

$$- \sum_{xy} p(x,y) \log_2 \frac{1}{p(y)}$$

$$= \sum_{xy} p(x,y) \log_2 \frac{p(x)\,p(y)}{p(x,y)}$$

LOG INEQUALITY

$$\log_e x \leq x - 1$$



$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x)\,p(y)}{p(x,y)} \leq$$

$$\leq \log_2 e \sum_{x,y} p(x,y) \left[\frac{p(x)\,p(y)}{p(x,y)} - 1 \right]$$

$$= \log_2 e \left[\sum_{x,y} p(x)\,p(y) - \sum_{x,y} p(x,y) \right]$$

$$= \log_2 e \left[\sum_{x} p(x) \sum_{y} p(y) - \sum_{x,y} p(x,y) \right]$$

$$= \log_2 e \left[1 \cdot 1 - 1 \right] = 0$$

$$H(X,Y) \leq H(X) + H(Y)$$

(X, Y) EXAMPLES

	y_1	y_2
x_1	0.4	0.2
x_2	0.35	0.05

	X
x_1	0.6
x_2	0.4

	Y
y_1	0.75
y_2	0.25

$H(X,Y) = 1.74$ $\quad H(X) = 0.97$ $\quad H(Y) = 0.81$

$H(X,Y) \leq H(X) + H(Y) = 1.78$

STATISTICALLY IND: $p(x,y) = p(x)\,p(y)$

0.45	0.15
0.30	0.10

$H(X,Y) = 1.78 = H(X) + H(Y)$

RECAP: $H(X)$ $H(X,Y)$

$$H(X,Y) \leq H(X) + H(Y)$$

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

IFF X Y

ARE ST. IND.

CONDITIONAL ENTROPY

$$H(X,Y) = \sum_{xy} p(x,y) \log_2 \frac{1}{p(x,y)} = (*)$$

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

$$(*) = \sum_{xy} p(x,y) \log_2 \frac{1}{p(x|y)\, p(y)}$$

$$= \sum_{xy} p(x,y) \log_2 \frac{1}{p(x|y)\,p(y)} =$$

$$= \boxed{\sum_{xy} p(x,y) \log_2 \frac{1}{p(x|y)}} \longleftarrow H(X|Y)$$

CONDITIONAL ENTROPY

$$+ \sum_{xy} p(x,y) \log_2 \frac{1}{p(y)}$$

$$\sum_{y} p(y) \log_2 \frac{1}{p(y)} = H(Y)$$

MEANING (IMPORTANT)

$$H(x,y) = H(x|y) + H(y)$$

RESIDUAL UNCERTAINTY ABOUT X WHEN Y OUTCOME IS REVEALED

$H(x)$ = ORIGINAL UNCERTAINTY ABOUT X

CASE 1: X AND Y STATISTICALLY INDEPENDENT

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

WE OBSERVE Y BUT

X AND Y ARE NOT LINKED

→ WE GAIN NOTHING

X Y STATISTICALLY INDEPENDENT

$$H(x|y) = H(x)$$

PROOF

$$H(x,y) = H(x|y) + H(y)$$

IF X AND Y ARE STATISTICALLY IND.

$$H(x,y) = H(x) + H(y)$$

$$\rightarrow \quad H(x|y) + H(y) = H(x) + H(y)$$

$$\rightarrow \quad H(x|y) = H(x)$$

CASE 2: X AND Y NOT STATISTICALLY INDEPENDENT

$$H(x|y) \leq H(x)$$

WHEN WE OBSERVE Y

THE AMOUNT OF UNCERTAINTY

ABOUT X IS REDUCED

X Y NOT STATISTICALLY INDEPEND.

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

PROOF

$$H(X,Y) \leq H(X) + H(Y)$$

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

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

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

CASE 3: $H(x|y) = 0$

Y CONTAINS ALL INFO ABOUT X

IF $y = f(x)$ WITH f INJECTIVE

$$H(x|y) = 0$$

$x \quad \Omega_x = \{ x_1 \quad x_i \quad x_n \}$

$y = f(x) \quad \Omega_y = \{ f(x_1) \quad f(x_i) \quad f(x_n) \}$

IF WE OBSERVE $f(x_i)$

WE AUTOMATICALLY

KNOW x_i

SUMMARY AND MEANING

$$0 \leq H(x|y) \leq H(x)$$

CHAIN RULE OF ENTROPY

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

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

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

$$+ H(Z)$$

VERY IMPORTANT: INFORMATION GAIN

X Y

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

$I(X;Y)$: REDUCTION OF UNCERTAINTY ABOUT X OBTAINED BY OBSERVING Y

$\equiv$

INFORMATION GAIN ABOUT X OBTAINED BY OBSERVING Y

$H(X)$: ORIGINAL UNCERT. ABOUT X

$H(X|Y)$: RESIDUAL UNCERT. ABOUT X AFTER OBSERVING Y

$$0 \leq I(X;Y) \leq H(X)$$

WHEN X AND Y ARE ST. IND.

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

GAIN = 0

WHEN Y CONTAINS ALL INFO ABOUT X

GAIN = ENTIRE $H(X)$

TWO ALTERNATIVE FORMULATIONS FOR $I(X;Y)$

$$I(x;y) = H(x) + H(y) - H(x,y)$$

$$I(x;y) = H(x) - H(x|y)$$

BUT $$H(x,y) = H(x|y) + H(y)$$

$$\rightarrow H(x|y) = H(x,y) - H(y)$$

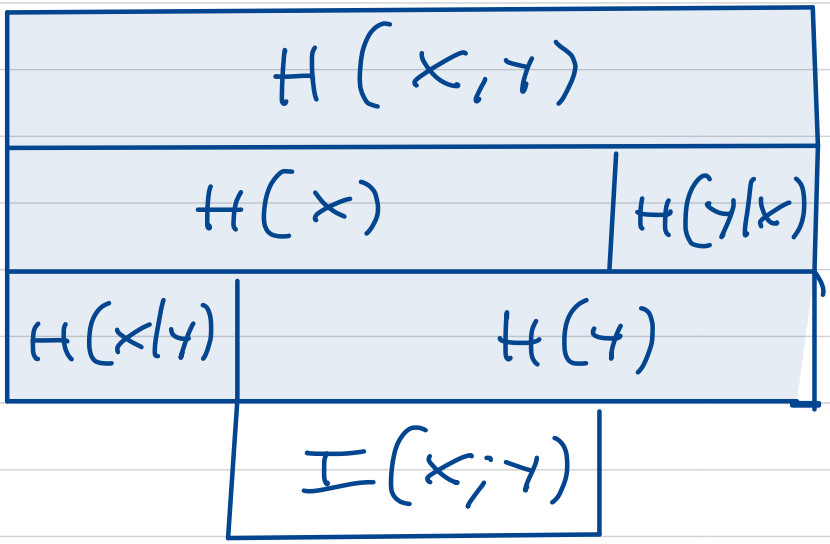
$$H(x,y) = H(y|x) + H(x)$$

$$I(x;y) = H(y) - H(y|x)$$

$$I(x;y) = H(x) - H(x|y)$$

SYMMETRIC !!

IMPORTANT FIGURE



$$I(x;y) = H(x) - H(x|y)$$

$$I(x;y) = H(y) - H(y|x)$$

$$I(x;y) = H(x) + H(y) - H(x,y)$$

INFORMATION GAIN RATIO

$$0 \leq I(x;y) = H(x) - H(x|y) \leq H(x)$$

$$0 \leq I'(x;y) = \frac{I(x;y)}{H(x)} \leq 1$$

FOR THE DATA SET OF ASSIGNMENT 2

$X_1 \quad X_2 \quad X_3 \mid X$ ← TARGET

$$\overline{I}(X;X_i) = \frac{I(X;X_i)}{H(X_i)} = \frac{H(X) - H(X|X_i)}{H(X_i)}$$

$1 \leq i \leq 3$

KULLBACK - LEIBLER DIVERGENCE

$$X \qquad \Omega_x = \{ x_1 \quad x_i \quad x_n \}$$

$$P = \{ p_1 \quad p_i \quad p_n \}$$

$$Q = \{ q_1 \quad q_i \quad q_n \}$$

$$D_{KL}(P|Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

IT'S USED TO COMPARE 2 PROB. DISTRIBUTIONS

TYPICALLY $P \equiv$ OBSERVED, MEASURED

$Q \equiv$ ANALYTICAL FIT

$$D_{KL}(P \| Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

IF $p_i = q_i$ $\forall i$ $\quad D_{KL}(P \| Q) = 0$

NOT A DISTANCE

$$D_{KL}(P|Q) = D_{KL}(Q|P)$$

$$D(P|Q) \geq 0$$

PROOF $D \geq 0$

$$D = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

$$-D = \sum_i p_i \log_2 \frac{q_i}{p_i}$$

$$\leq \sum_i p_i \left(\frac{q_i}{p_i} - 1 \right) = \sum_i q_i - \sum_i p_i =$$

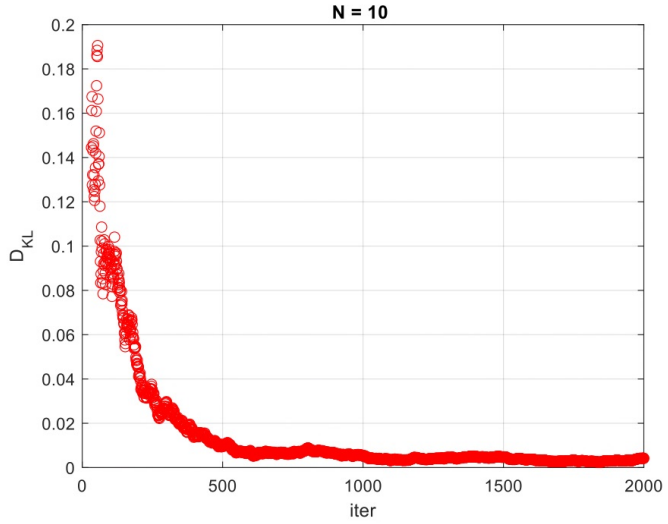
$$= 1 - 1 = 0$$

$$-D \leq 0 \quad \rightarrow \quad D \geq 0$$

UNIFORM DISTRIB. OVER 10 ELEM

$$q_i = \frac{1}{10} \quad \forall i$$

COMPUTATION OF $D_{KL}(P|Q)$ FOR INCREASING # OF ITERATIONS



LINK BETWEEN $I(X;Y)$ AND D

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

$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

$$= D\left(P(X,Y) \,\middle|\, P(X)P(Y) \right)$$

COMMON INTERPRETATION

$$D_{KL}(P/Q) = \sum_i P_i \log_2 \frac{P_i}{Q_i}$$

P_i = OBSERVED PROB. DISTR.

Q_i = MODEL/ANALYTICAL DISTR.

D = INF. GAIN WHEN WE USE MEASURED P_i INSTEAD OF MODEL Q_i

D = INF. LOSS WHEN WE USE MODEL Q_i INSTEAD OF MEASURED P_i

Information Theory for Data Science

Assignment 1

Introduction to Information Theory and application to Classifiers

Draft version 0.2

Exercises:

1. Entropy of a random variable with 3 outcomes (pt. 3)
2. Entropy of a random variable from a data series (pt. 4)
3. Application of the principle of maximum entropy (pt. 4)

4. Exercise on Information Gain Ratio (pt. 4)
5. Kullback Leibler divergence from empirical distribution (pt. 3)
6. Permutation entropy for anomaly detection (pt. 5)

7. Information Gain Ratio and Classifier (pt. 7) (not yet available)

Exercise 1 - Entropy of a random variable with 3 outcomes (pt. 3)

Exercise 1 - Entropy of a random variable with 3 outcomes

1. Given a random variable with 3 outcomes, write a program to plot the entropy as a function of all possible probability vectors
2. Start with a probability vector where one of the elements is significantly higher than the others. Apply an iterative averaging procedure (for example, replace each element with the average of itself and its neighbors, followed by normalization). For each updated vector, compute the entropy and plot its value on the figure generated in step 1. Show that as the probability distribution approaches the uniform distribution, the entropy approaches its maximum value. Finally, discuss the results.

Exercise 2 - Entropy of a random variable from a data series (pt. 4)

Exercise 2 - Entropy of a random variable from a data series

1. Identify a data series and estimate the probabilities of the outcomes based on their occurrences, updating the probabilities at each time step.
2. At each time step, compute the entropy, plot its behavior, and discuss the results

Note: In the presentation, include a link to the source of the data series

Exercise 3 - Application of the principle of maximum entropy (pt. 4)

Exercise 3.a

1. Invent an exercise where you have a random variable X with an alphabet Ω_X with 2 outcomes with integer values.
2. Show some examples of the the probability distribution P(X) for different values of the mean value.
3. Discuss the results

Exercise 3.b

1. Invent an exercise where you have a random variable X with alphabet Ω_X with at least 4 outcomes, where each outcome has an inteher value (“cost”).
2. Fix the mean value bigger than the arithmetic average of the costs, and apply the principle of maximum entropy to find the probability distribution P(X)
3. Plot P(X)
4. Repeat with a mean value equal to the arithmetic average and plot the result
5. Repeat with other values of the mean value and plot the results
6. Comment the results

You must numerically solve the equation generated by the Lagrange optimization.

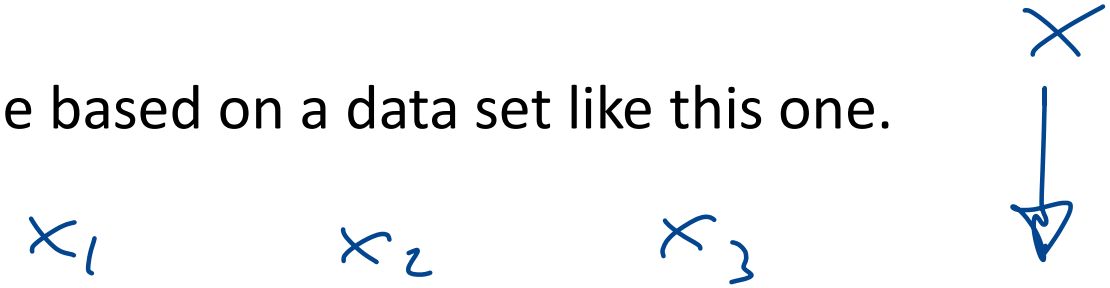
As an example , for Matlab you can use

```
syms x
eqn = ( . . .  )*mu == ( . . . );
V = vpasolve(eqn,x,[0 10])
```

Exercise 4 – Exercise on Information Gain Ratio (pt. 4)

Exercise 4 – Exercise on Information Gain Ratio

1. Invent an exercise based on a data set like this one.



Athlete	Training Hours	Rest Hours	Gym Workouts	Performance
A	High	Low	Low	Lose
B	Medium	High	Medium	Win
C	Low	Medium	High	Lose
D	High	Medium	Medium	Win
E	Medium	High	Low	Win
F	Low	Low	High	Lose
G	High	Low	Medium	Win
H	Medium	Low	Low	Lose
I	Low	High	Medium	Win
J	High	High	High	Win

The last column represents your target variable X.

$x_1 \quad x_2 \quad x_j$

2. Compute the Information Gain Ratio with respect to all the other variables and select the one that provides more information about X.

$$\overline{I}(x; x_i) = \frac{I(x; x_i)}{H(x_i)} = \frac{H(x) - H(x|x_i)}{H(x_i)}$$

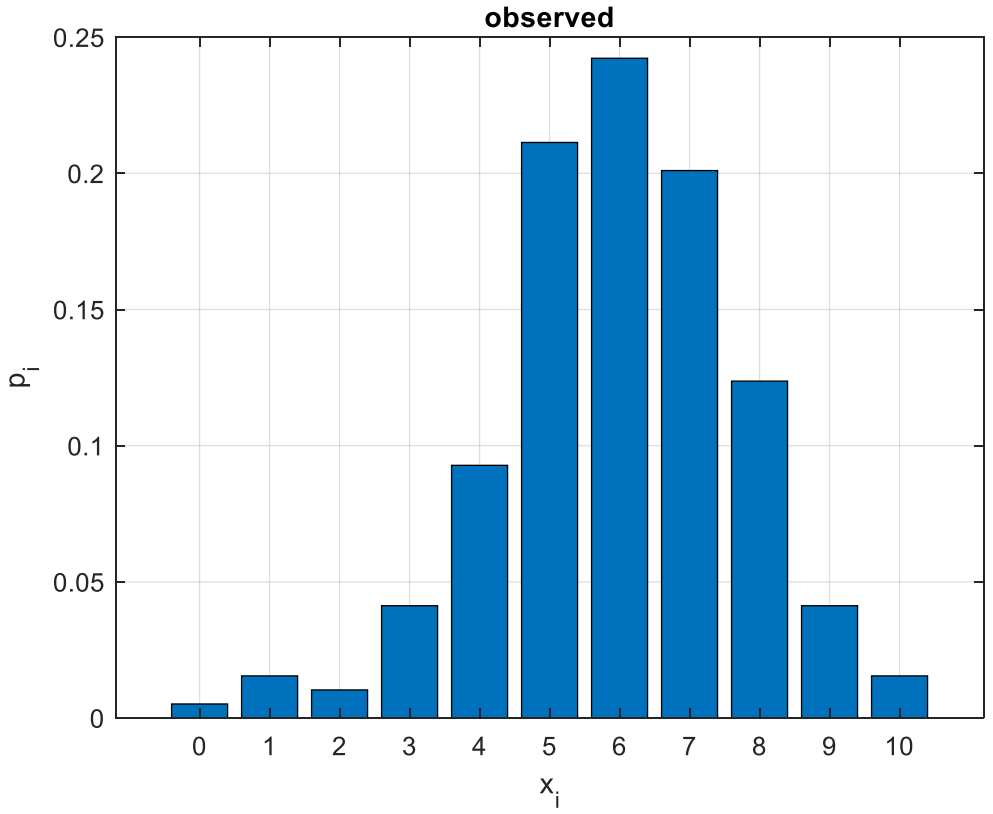
Exercise 5 - Kullback-Leibler distance from empirical distribution (pt. 3)

Exercise 5 - Kullback-Leibler distance from empirical distribution

Consider the observed data

$$x_i = [0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10]$$

Number of observed x_i = [1 3 2 8 18 41 47 39 24 8 3]


observed
0.25
0.2
0.15
0.1
0.05
0
p_i
0
1
2
3
4
5
6
7
8
9
10
x_i

Compare with uniform pmf

1. Plot the two observed and the uniform pmfs
2. In the title write the KL divergence value

Compare with binomial pmf with $0 < p < 1$ (step=0.001)

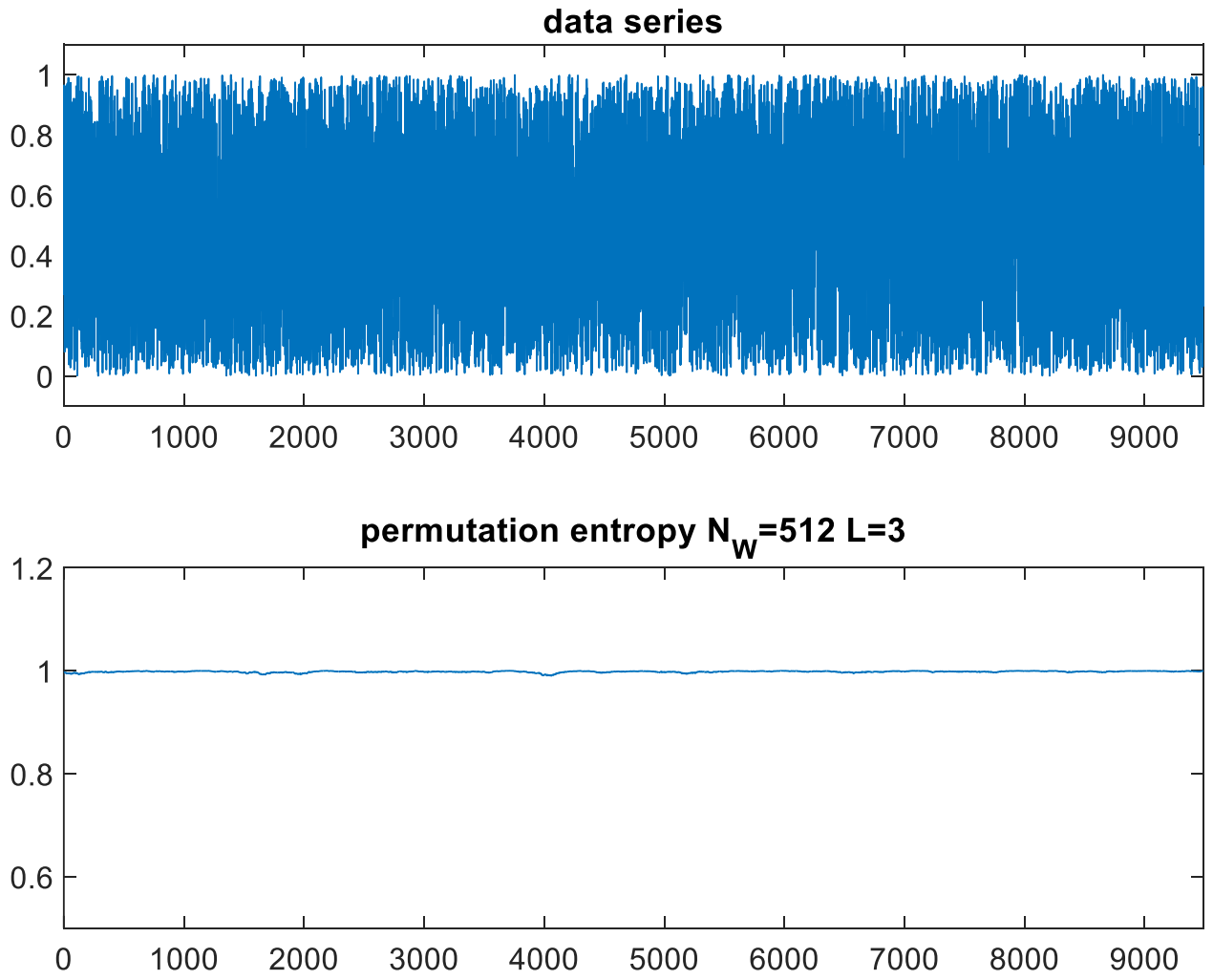
3. Identify the pmf at minimum KL divergence
4. Plot the observed and the best binomial pmfs
5. In the title write the value of p and KL divergence

Exercise 6 - Permutation Entropy for Time Series Anomaly Detection (pt. 5)

Exercise 6 - Permutation Entropy for Time Series Anomaly Detection

Write a program that:

- Generates a data series made by 10,000 random symbols.
- Computes the permutation entropy with a sliding window of $N_W = 512$ symbols. (For example, choose an order m = 3.)
- Plot the data and the entropy.


data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
permutation entropy N_W=512 L=3
1.2
1
0.8
0.6
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

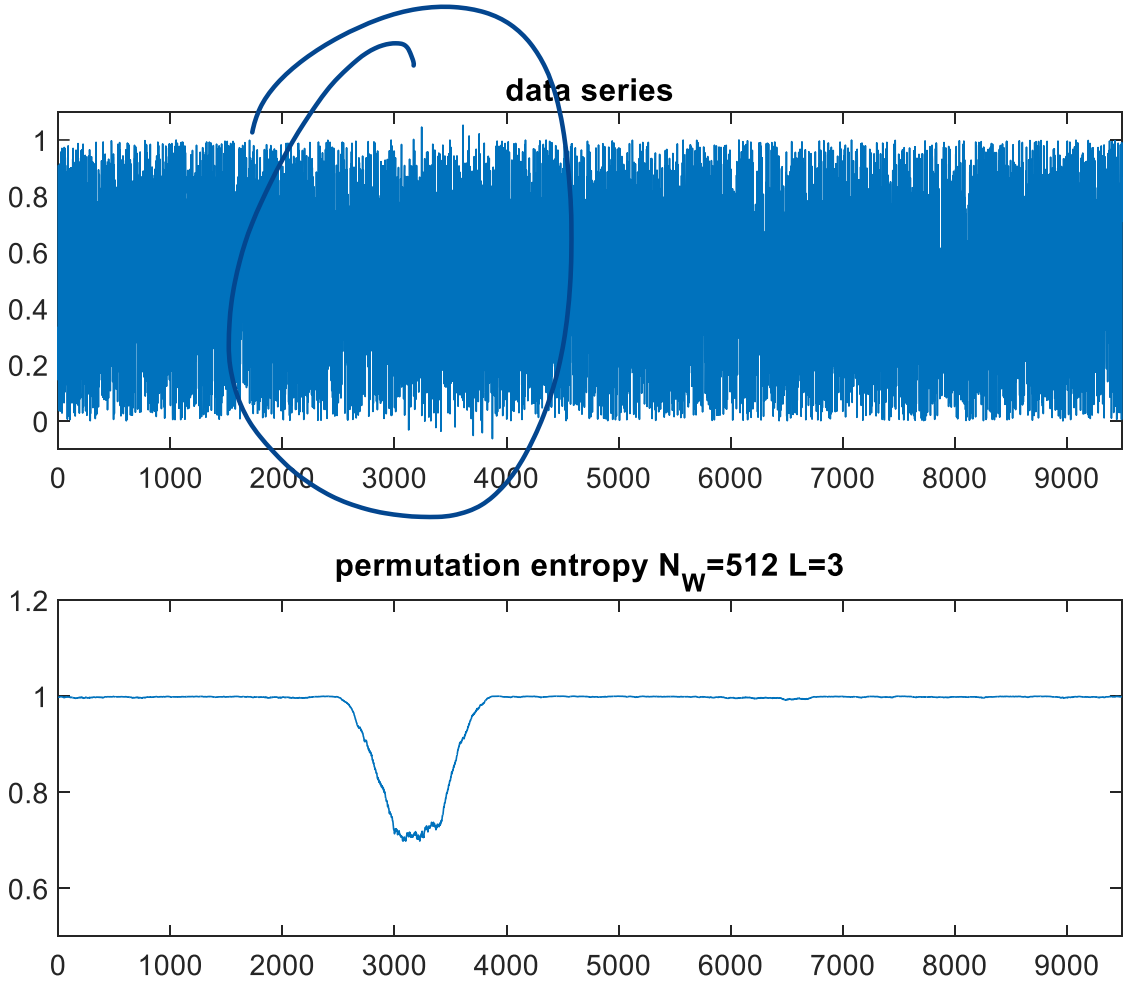
- Insert a pattern between 3000 and 4000 made by correlated data (with about the same mean value and variance of the original random data).

- **Describe how you generated the correlated data**

- Compute the permutation entropy with the same N_W and m used before.

- Plot the data and the entropy.

- Comment the results


data series
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
permutation entropy N_W=512 L=3
1.2
1
0.8
0.6
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

Some Matlab functions that might be useful

```
sort
perms
```

Exercise 7 - Information Gain Ratio and Classifier (pt. 7)

Important

Final version assigned on 08/10/2024 (**to be confirmed**)

Delivery by

- 23/10/2024, 11.59 PM: +2 points
- 30/10/2024, 11.59 PM: +1 point
- 21/02/2025, 11.59 PM: 0 points
- **Later: not accepted**

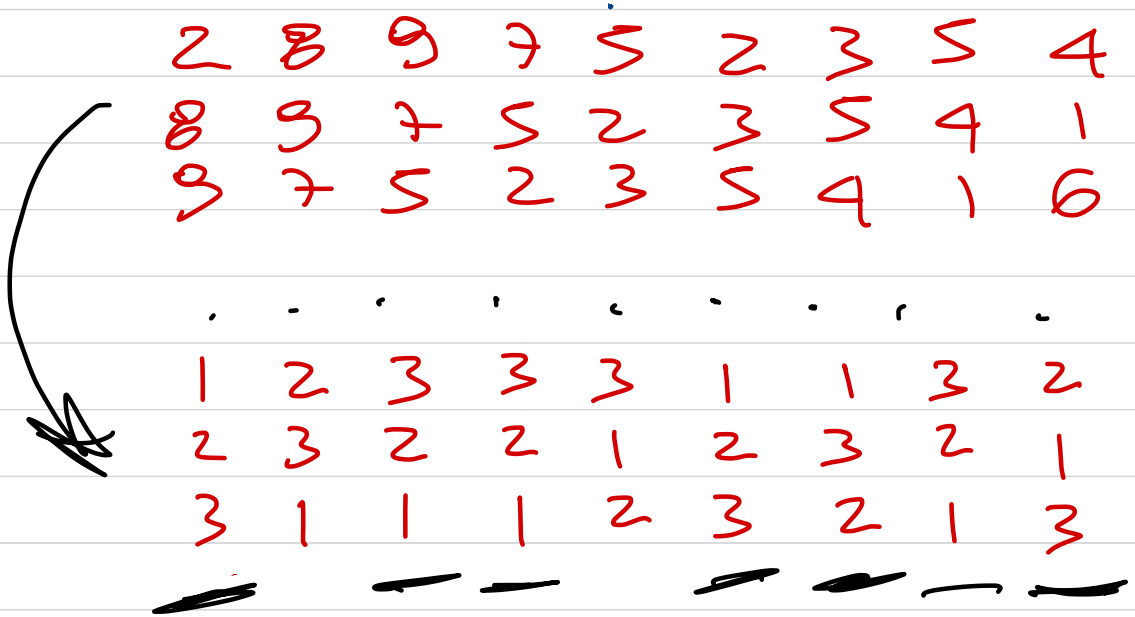
PERMUTATION ENTROPY

2 8 9 7 5 2 3 5 4 1 6

ORDER $m = 3$

WE ANALYZE VECTORS OF LENGTH 3 INSIDE THE SEQUENCE

2 8 9 7 5 2 3 5 4
8 9 7 5 2 3 5 4 1
9 7 5 2 3 5 4 1 6



$m = 3$

1 1 2 2 3 3
2 3 1 3 1 2
3 2 3 1 2 1

2 1 1 1 1 3 | 9

$\frac{2}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{1}{9}\ \frac{3}{9}$

$P_1\ P_2\ P_3\ P_4\ P_5\ P_6$

$$H = \sum_{i=1}^{6} P_i \log_2 \frac{1}{P_i}$$

$$0 \leq H \leq \log_2 m!$$

$$0 \leq \bar{H} = \frac{\sum p_i \log_2 \frac{1}{p_i}}{\log_2 m!} \leq 1$$

RANDOM DATA $\rightarrow$ $\bar{H}$ CLOSE TO 1

CORRELATED DATA $\rightarrow$ WE OBSERVE A DROP OF $\bar{H}$

$a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8\ a_9\ a_{10}$

- $N = 10$
- $N_w = 5$
- $m = 3$

$a_1\ a_2\ a_3$
$a_2\ a_3\ a_4$ → 1 VALUE OF P.E.
$a_3\ a_4\ a_5$

$a_2\ a_3\ a_4$
$a_3\ a_4\ a_5$ → 1 VALUE OF P.E.
$a_4\ a_5\ a_6$

⋮

$a_6\ a_7\ a_8$
$a_7\ a_8\ a_9$
$a_8\ a_9\ a_{10}$

$N - N_w + 1$ VALUES OF P.E.

Some Matlab functions that might be useful

```
sort
perms
```

M = MATRIX

$a_0\ a_1\ a_2 \ldots$
$a_1\ a_2\ a_3 \ldots$
$a_2\ a_3\ a_4 \ldots$

[B, P] = sort(M)

CONTAINS PERMUTAT.