COMP6237 – Information Theory and Feature Selection

Shoaib Ehsan

s.ehsan@soton.ac.uk

Lecture slides available here:

http://comp6237.ecs.soton.ac.uk/

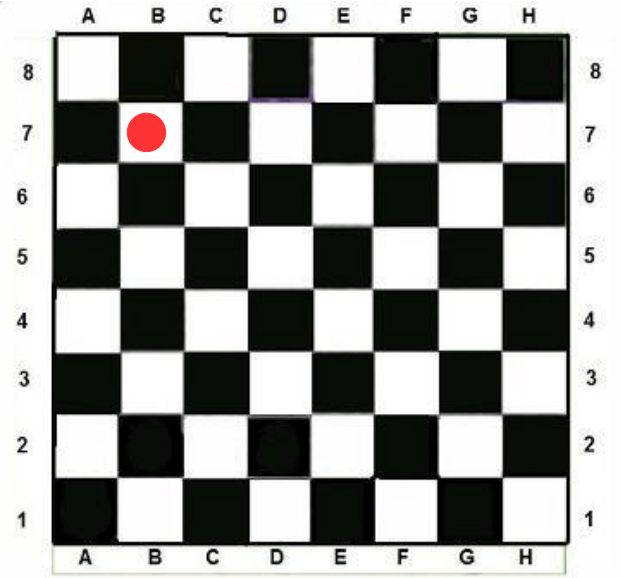
(I borrowed heavily from Cosma Shalizi to prepare this lecture)

Some Information Theory

- Why?
  - Understand Kullback-Leibler divergence
  - Useful in many other contexts in data mining
- Agenda:
  - Information
  - Entropy/Coding
  - Mutual information
  - Using information theory for feature selection
  - Summary
  - Problems

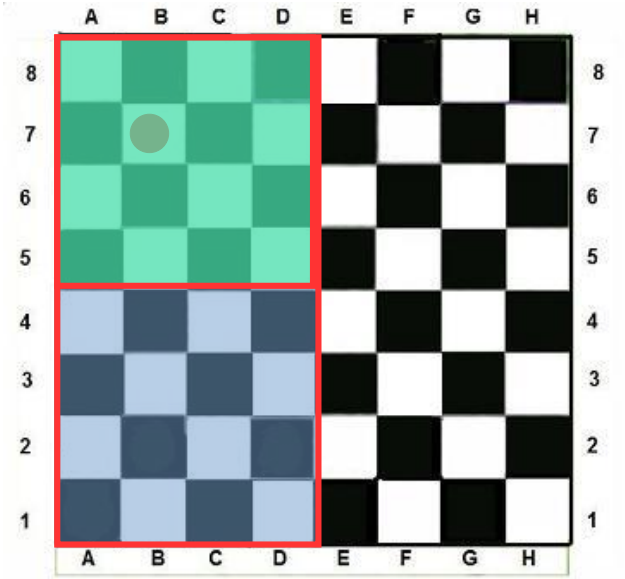
Information Theory

- To really understand this, we need to know some basic stuff from information theory
- … So: What is information?
- Imagine a single piece on a chess board; you don't know where it is. How much information is there in knowing its location?



Information Theory

- To really understand this, we need to know some basic stuff from information theory
- … So: What is information?
- Imagine a single piece on a chess board; you don't know where it is. How much information is there in knowing its location?



Can approach this as a sequence of YES/NO questions.

Want to ask the minimal number of such questions.

Divide remaining area of chess board into halves, and ask in which half the piece is. Iterate this until we find the piece.

We might identify the number of such questions with information.

Given that we have 64 squares, the number of such questions we have to ask is $\log_2 64=6$

Information Theory

- We can abstract this:
  - Assume there is some probability space W
  - And a pdf P(w) that assigns a likelihood to each member in W
  - Say one element has been sampled from P(w)
  - How much information is there in knowing what element it was (retrospectively)? → **Information**
    - A rare event will be difficult to figure out and thus will carry much information
  - Alternatively: when sampling an event from P, how surprised are we to find certain events? → **Uncertainty/Surprise**
    - Rare events will be unexpected and cause much surprise

Information Theory (2)

- What if the probability of the piece being at some location is different from other locations?
- e.g., a random source emits one signal A,B,C, or D according to

$$\{Pr(x = A) = 1/2, Pr(x = B = 1/4), Pr(x = C) = 1/8, Pr(x = D) = 1/8\}$$

- What is the optimal set of questions to figure out what symbol it was?
- How many questions do we need for A,B,C?
- What is the information content in A,B,C?

Information Theory Detour (2)

- What if the probability of the piece being at some location is different from other locations?
- e.g., a random source emits one signal A,B,C, or D according to

$$\{Pr(x = A) = 1/2, Pr(x = B = 1/4), Pr(x = C) = 1/8, Pr(x = D) = 1/8\}$$

- What is the optimal set of yes/no questions to figure out what symbol it was?
  - Is it A? If not, is it B? If not, is it C(D)?
- How many questions do we need for A,B,C?
  - A: 1=$\log_2 2$, B: 2=$\log_2 4$, C and D: 3=$\log_2 8$
- What is the information content in A,B,C?
  - 1 bit, 2bits, 3 bits ...

Entropy

- If we have a probability distribution p(x) x from X we can assign information values to each x
- The **information** of observing x then is **$\log_2$ 1/p(x)**
- (like in the chess board example each square had chance 1/64; this gives the same value in this case)
- Shannon (1948):

$$H(p) = -\sum_x p\,(x)\log_2 p(x)$$

[bits]

- i.e., entropy of a distribution is the expectation of information of the distribution or the average surprise when sampling from the distribution

Example (1)

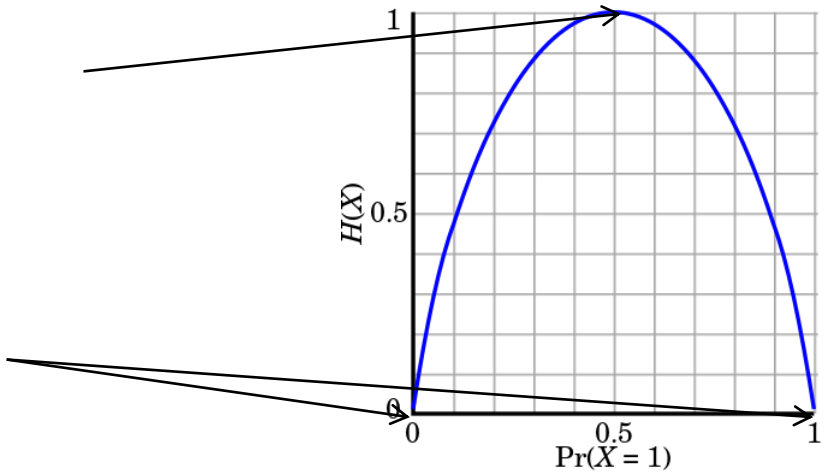
- Let's say we toss an unfair coin, heads appears with probability p.
- Entropy?

Example (1)

- Let's say we toss an unfair coin, heads appears with probability p.
- Entropy? $H(P) = -p\ln p - (1-p)\ln(1-p)$

Distribution has maximal information, we do not know what to expect.

Distribution has hardly any information (since we know what to expect, it's either almost always heads or tails)



Entropy is thus a measure of uncertainty.

Entropy (2)

- Suppose we have a data stream that emits symbols x with according to a probability distribution p(x)
- We are looking for an optimal (min. number of symbols) encoding of messages from that stream
- Roughly: the entropy of p determines the length of such a code through answers to optimal YES/NO questions, symbol length ~ -log
  - Frequent symbols get short codes, infrequent symbols get long codes

Example

- Back to the previous example:

$$\{Pr(x = A) = 1/2, Pr(x = B = 1/4), Pr(x = C) = 1/8, Pr(x = D) = 1/8\}$$

- What codes would we choose for a binary alphabet?
- Can do this by coding our answers to the yes/no
- questions.
  - A → 1
  - B → 01
  - C → 001, D → 000

Entropy Example

- Can use this to analyse text from newspapers
- e.g.,:
  - Somebody extracted all words from NYT articles for in 2004. Let's say we want to build a code based on these words and use it to encode articles, say for one issue in 2004 and one issue in 2005.
  - How can we extract information about these articles?
    - Crude approach: “bag of words” idea – count frequencies of all words and store them in some vector; can then interpret this as a probability vector
    - This ignores a lot of fine detail (e.g., correlations between words etc.)

Entropy Example

- OK, say we have one such vector
  - P(x) for the issue from 2004
  - Q(x) for the issue for 2005
- This allows us to evaluate the information content of both issues, e.g.

$$\sum_x P(x)\log\frac{1}{P(x)} = 12.94bits \qquad \sum_x Q(x)\log\frac{1}{Q(x)} = 12.77bits$$

- Can assign information content to words (-log 1/F(x)); most frequent words do not carry content, so we expect these non content words in equal proportions in 2004 and 2005

“non content words”

word	n.04	freq.04	bits.04	n.05	freq.05	bits.05
the	6375	0.0626	4	5783	0.0622	5
to	2777	0.0273	6	2543	0.0274	6
of	2708	0.0266	6	2365	0.0254	6
a	2557	0.0251	6	2497	0.0269	6
and	2338	0.0230	6	2137	0.0230	6
in	2248	0.0221	6	2107	0.0227	6
that	1377	0.0135	7	1315	0.0142	7
said	972	0.0096	7	1027	0.0111	7
for	952	0.0094	7	893	0.0096	7
he	901	0.0089	7	741	0.0090	7

Popularity Comparisons ...

Old news: Words that gained popularity

word	n.04	freq.04	bits.04	n.05	freq.05	bits.05
lebanon	1	1.96e-05	16	49	5.38e-04	11
lebanese	1	1.96e-05	16	47	5.16e-04	11
arts	0	9.82e-06	17	34	3.76e-04	12
bolton	0	9.82e-06	17	28	3.12e-04	12
hezbollah	1	1.96e-05	16	28	3.12e-04	12
march	30	3.04e-04	12	103	1.12e-03	10
prison	10	1.08e-04	14	27	3.01e-04	12
syria	9	9.82e-05	14	30	3.33e-04	12

Old news: Words that dropped in popularity

word	n.04	freq.04	bits.04	n.05	freq.05	bits.05
saatchi	41	4.12e-04	12	0	1.08e-05	17
dvd	32	3.24e-04	12	0	1.08e-05	17
cantalupo	32	3.24e-04	12	0	1.08e-05	17
april	111	1.10e-03	10	15	1.72e-04	13
bonds	57	5.69e-04	11	8	9.68e-05	14
kerry	43	4.32e-04	12	3	4.30e-05	15
tax	32	3.24e-04	12	8	9.68e-05	14
campaign	58	5.79e-04	11	26	2.90e-04	12

Example, cont.

- Let's quantify the difference:
  - Q(x) … prob. of x in 2004, P(x) in 2005

$$\log 1/Q\,(x) - \log 1/P\,(x) = \log \frac{P(x)}{Q(x)}$$

  - Averaging over the distribution of words of the 2005 paper the expected difference in code length is

$$\sum_x P\,(x) \log \frac{P(x)}{Q(x)}$$

    - If we use the code from 2004 to encode 2005 paper

$$\sum_x P\,(x) \log \frac{1}{Q(x)} = 13.29 bits$$

    - If we code using the frequencies from 2005:

$$\sum_x P\,(x) \log \frac{1}{P(x)} = 12.94 bits$$

K-L Divergence

- Given two probability distributions f(x) and g(x) for a random variable x, the K-L divergence (or relative entropy) is:

$$D(f \| g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

- Comments:
  - Compares the entropy of two distributions over the same random variable
  - Heuristically: number of additional bits encoding a random variable with distribution f(x) using g(x)

Cross Entropies (1)

- Suppose we want to measure the information content of some prob. distribution p(x) but measure it based on a code optimal for some other “artificial” q(x)
- → Cross entropy

$$H(p,q) = -\sum_x p\,(x)\log_2 q(x)$$

$$H(p) = -\sum_x p\,(x)\log_2 p(x)$$

$$D(p\|q) = \sum_x p\,(x)\log\frac{p(x)}{q(x)}$$

$$H(p,q) = H(p) + D(p\|q)$$

- →

Cross Entropies (2)

- Have seen before that we can see regression methods as trying to minimize K-L divergences
- When minimizing K-L against a fixed reference distribution p, minimizing K-L is equivalent to minimizing cross entropies (→ “Principle of minimum cross entropies”)
- Can easily see how this links to estimation problems

Cross Entropies + Estimation

- Suppose we have a training set in which the empirical frequency of occurrences of outcomes is N p_i and the estimated probability of outcome i is q_i
- Likelihood function then is
-

$$L(q_i; p_i) \propto \prod_i q_i^{Np_i}$$

$$1/N \log L(q_i; p_i) \sim \sum_i p_i \log q_i = H(p, q)$$

- Maximizing likelihood functions often equivalent to minimizing cross entropies

Conditional Entropy

- Let's say we have two random variables C and X which are not independent
- So, if we observe one feature in X this will change our knowledge about C, i.e., if we observe x our uncertainty about C changes by the **conditional entropy**

$$H(C|X=x) = -\sum_c Pr\,(C|X=x)\log_2 Pr(C|X=x)$$

- The difference between the entropy of H[C] and the conditional entropy H[C|X] is **realized information**

$$I[C;X=x] = H(C) - H(C|X=x)$$

(i.e. by how much did uncertainty change due to observing x)

Realized Information

$$I[C; X = x] = H(C) - H(C|X = x)$$

- Is not necessarily positive!
  - i.e. suppose C is “it rains today” and the probability that it rains is 1/7. Then H[C]=0.59 bits (check it!)
  - Suppose X=cloudy and the probability that it rains when it is cloudy is ½. Then H[C|X=cloudy]=1
  - Realized information from the observation of clouds is -0.41 bits, i.e., uncertainty has increased.

Mutual Information

- Mutual information is the expected information a feature gives us about a class

$$I[C;X] = H(C) - \sum_x Pr\,(X = x)H(C|X = x)$$

- Some remarks:
  - Mutual information is always positive
  - Is only zero if X and C are statistically independent
  - Is symmetric in X and C

Example: How much do words tell us about topics?

- Let's say we generate bag of words vectors and read all articles to classify them into two categories, articles about art and articles about music.
- Investigate the word “paint”. In how many articles in the arts or music categories is the word “paint” present

	Indicator X	
Class c	“paint”	“not paint”
art	12	45
music	0	45

(i.e., we have 57 articles about art and 45 about music, 12 art stories contain paint, no Music stories contain paint, etc. ...)

Words, Topics, Information

	Indicator X	
Class c	“paint”	“not paint”
art	12	45
music	0	45

- Entropy of C? H[C]=0.99
- H[C|X=”paint”]=0
  - i.e., if we find paint, we can be certain that the story is about art
- H[C|X=”not paint”]=1.0
  - i.e., if “paint” is absent, we are as uncertain as we are about a fair coin flip (i.e., a bit more uncertain as we were before checking for paint with H[C]=0.99)
- I[C;X]=H[C]-Pr(X=1)H[C|X=1]-Pr(X=0)H[C|X=0]=0.99-12/102*0-90/102*1=0.11
  - The expected reduction in uncertainty when checking for the indicator X is fairly small (0.11 bits)

Finding Informative Features

- This leads to an idea for an information theoretic procedure to find important words:
  - Count how often each class c=1,…,K appears
  - For each word, build the K^x2 table of classes by word indicators
  - Compute the mutual information in each table
  - Return the m most informative words, i.e. those with the largest mutual information
- This might work as a first attempt, but ignores a number of important factors, e.g.:
  - That combinations of features might be useful
  - That some features might be redundant given others
  - To remedy these problems we need to look at **interactions**

Joint and Conditional Entropy

- Joint entropy:

$$H[X,Y] = -\sum_{x,y} Pr\,(X = x, Y = y)\log_2 Pr(X = x, Y = y)$$

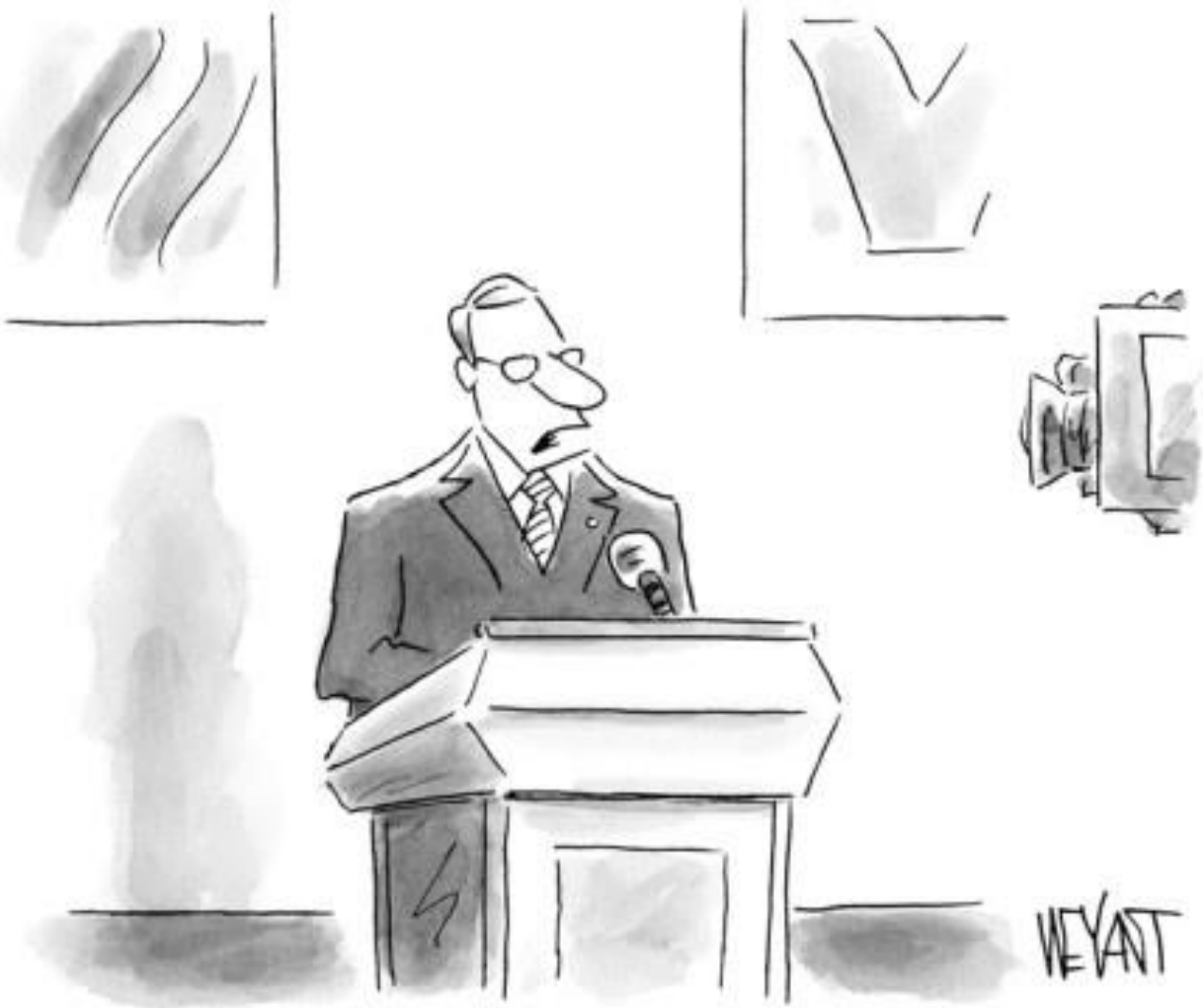
- Remarks:
  - This is sub-additive: $H[X,Y] \leq H[X] + H[Y]$
  - Mutual information: $I[X;Y] = H[X] + H[Y] - H[X,Y]$
  - Conditional entropy: $H[Y|X] = H[X,Y] - H[X]$
- Can also condition mutual information

$$I[C;Y|X] = H[C|X] - H[C|Y,X]$$

  - i.e., we ask how much information does Y contain about C if we "control" for X

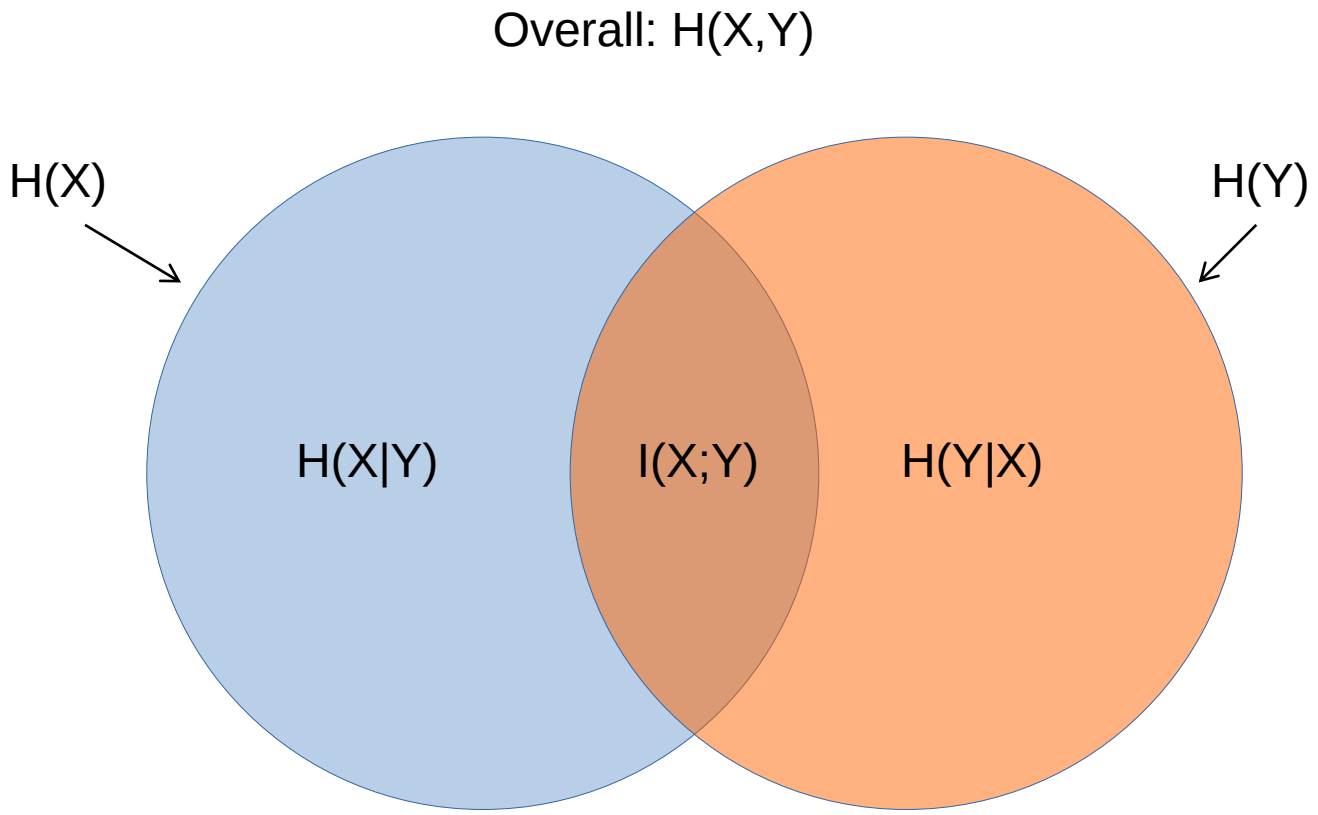
Interaction

- Conditional mutual information I[C;Y|X] is positive
  - But might be smaller/larger/equal to I[C;Y]
  - If I[C;Y|X] =I[C;Y]: C and Y are **conditionally independent** given X; otherwise there is an **interaction** between X and Y (regarding their information about C)
  - I[C;Y|X] <I[C;Y]: Some of the information in Y about C is **redundant** given X
  - Use this to define **interaction information**
  - I(C;Y;X)=I[C;Y]-I[C;Y|X]



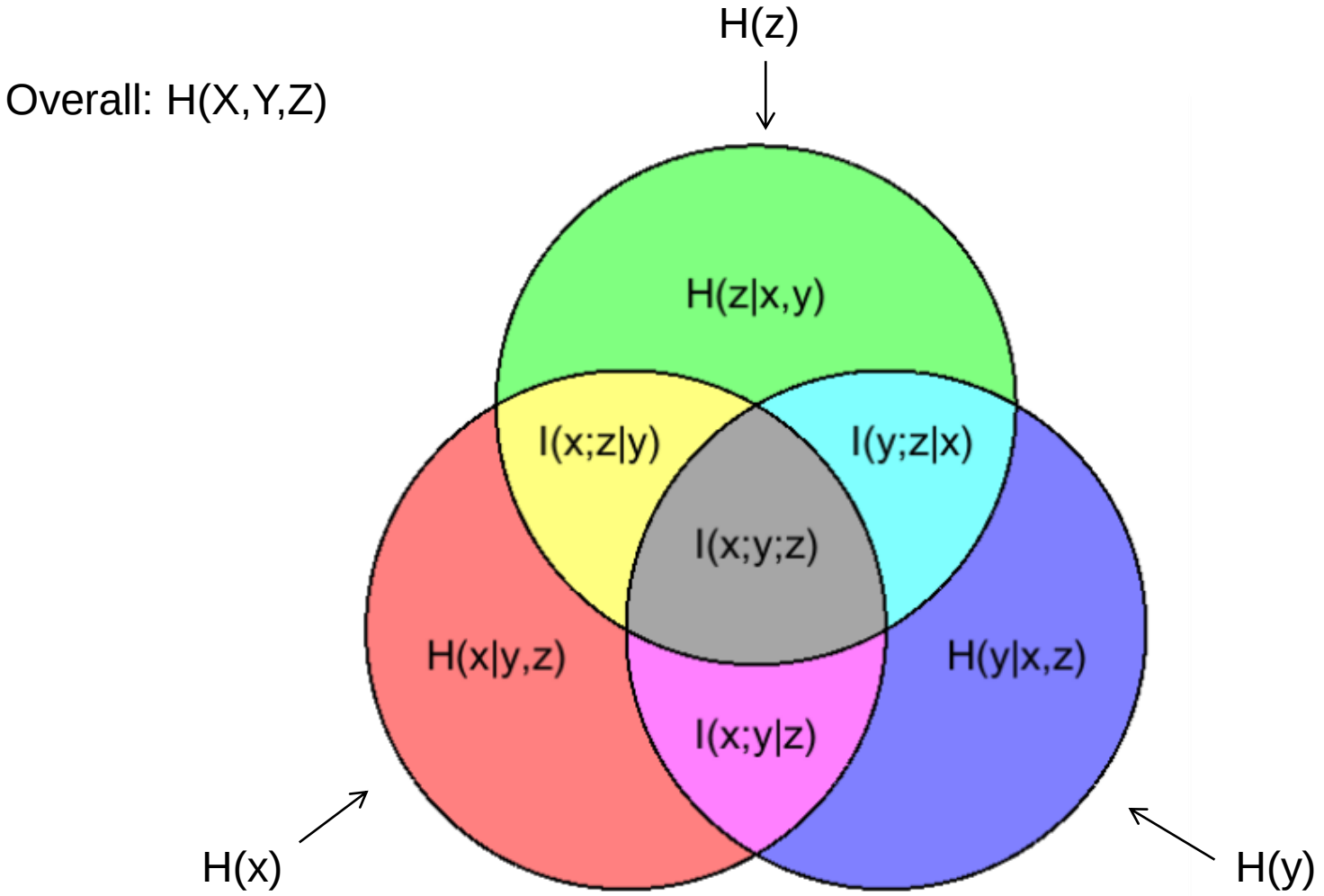
"I regret that my poor choice of words caused some people to understand what I was saying."

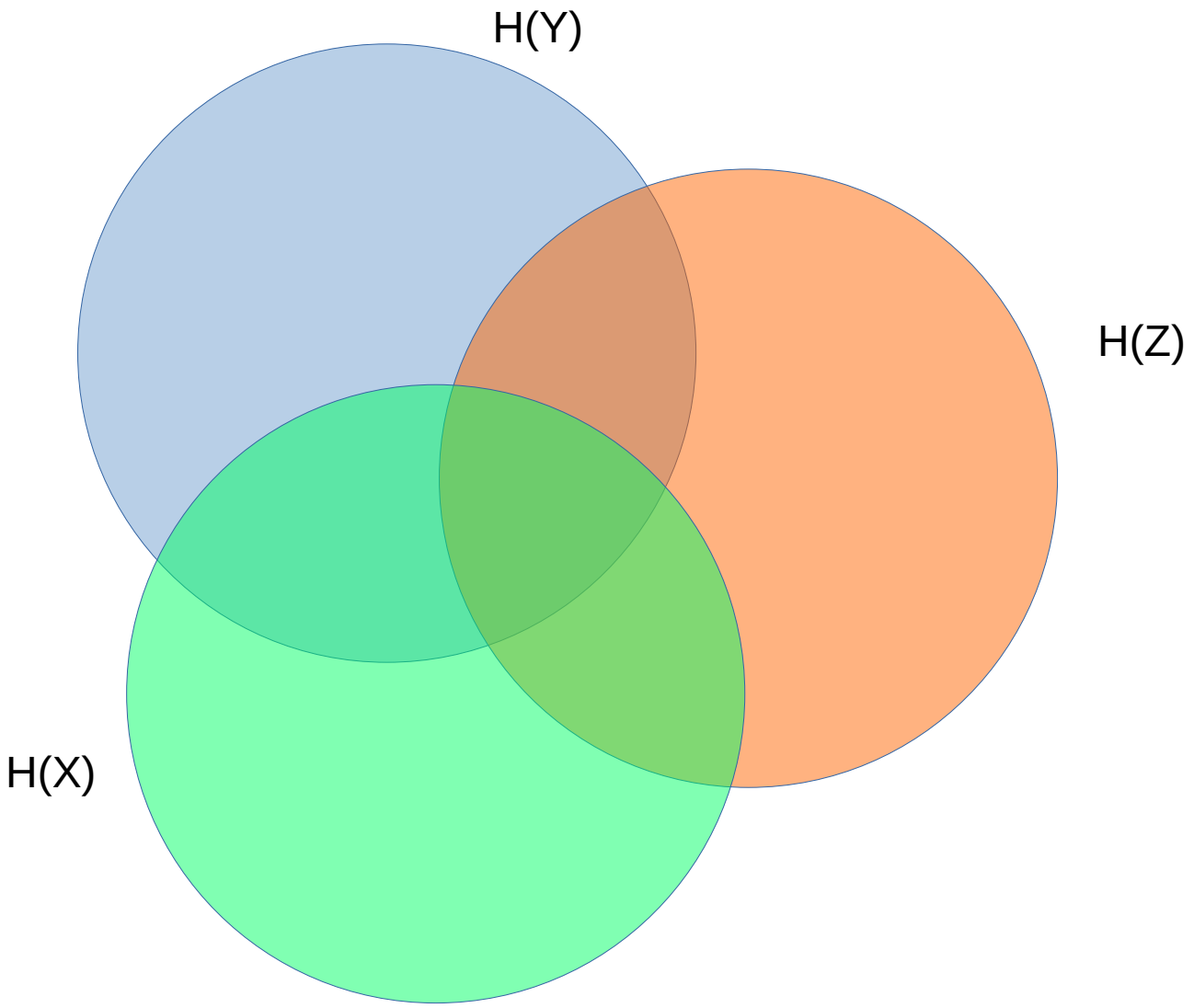
Venn Diagram for Information Content of 2 Random Variables



⟶ I(X;Y)=H(Y)-H(Y|X)=H(X)-H(X|Y) or
H(X,Y)<=H(X)+H(Y) etc.

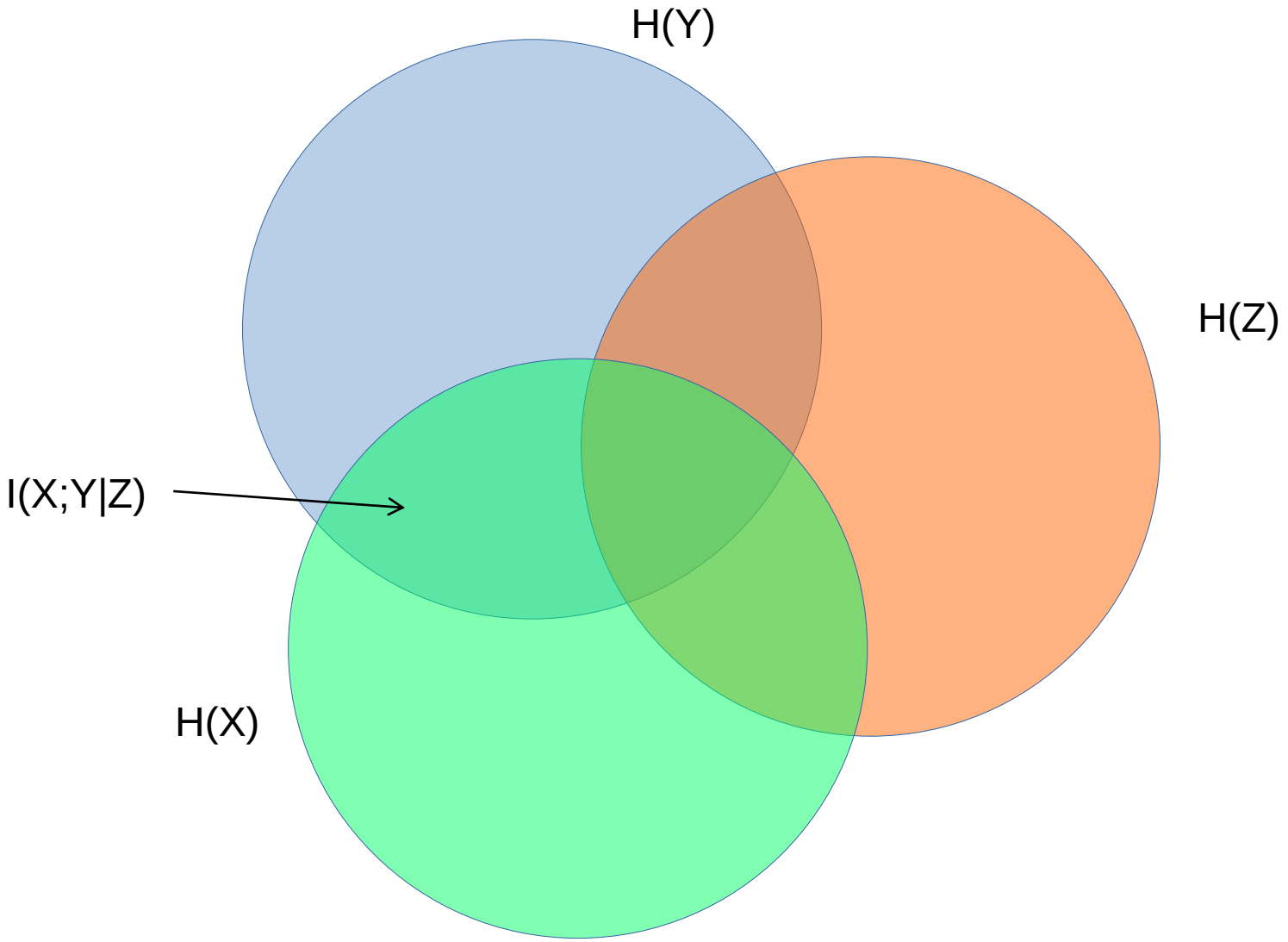
Venn Diagram for Information Content of 3 Random Variables





I(X;Y|Z)=?

This representation is useful to remember relationships between information theoretic measures for correlated (=”overlapping”) variables.



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

Finding Informative Features (2)

- Can use this to improve the algorithm from earlier on, i.e., have p features Xi and want to use it to predict C
  - Find I[C;Xi]. Select feature with most mutual information with C, say X1.
  - Given k selected features, calculate I[C;Xi|X1,…,Xk] for all non selected variables I
  - Select Xk+1 as the feature with most conditional mutual information and iterate.
- This is a **greedy** algorithm, so it does not necessarily come up with the best combination of features
- We need to impose a stopping condition, e.g., a threshold for I[C;Xi|X1,…,Xk] or a maximum number of features

Summary

- Important to remember:
  - How can we quantify information?
  - Entropy/Mutual information … and ideally a bit more information theory
  - Be able to apply these concepts in basic settings (try the problems in the next slides)
  - Idea of feature selection using information theory.
- Further reading:
  - An easily accesible primer on information theory:
  - http://alum.mit.edu/www/toms/papers/primer/primer.pdf
  - A more detailed and technical paper:
  - http://arxiv.org/pdf/cs/0308002v3.pdf

Problem (1)

- Prove that the information measure (slide 8) is additive: that the information gained from observing the combination of N independent events, whose probabilities are pi for i = 1....N, is the sum of the information gained from observing each one of these events separately and in any order.

Problem (2)

- Consider two independent integer-valued random variables, X and Y. Variable X takes on only the values of the eight integers {1, 2, ..., 8} and does so with uniform probability. Variable Y may take the value of any positive integer k, with probabilities $P\{Y = k\} = 2^{-k}$, k = 1, 2, 3, …
  - Which random variable has greater uncertainty? Calculate both entropies H(X) and H(Y).
  - What is the joint entropy H(X,Y) of these random variables, and what is their mutual information I(X;Y)?

Problem (3)

- Assume that we have some random source that emits one of M symbols with equal likelihood. What is the entropy?
- Assume a source is restricted to emitting one of M symbols at a time. What is the distribution of probabilities over these symbols that maximises the average uncertainty of the receiver?

Problem (4)

- Polynesian languages are famous for their small alphabets. Assume a language with the following letters and relative frequencies:
  - p (1/8), t (¼), k (1/8), a(1/4) i (1/8), u (1/8)
  - What is the per-character entropy for this language?
  - Design an (optimal, i.e., short) code to transmit a letter.

Problem (5)

- Find an example for three random variables X,Y,Z with
  - Negative interaction $I(X;Y|Z)<I(X;Y)$ and one for
  - Positive interaction $I(X;Y|Z)>I(X;Y)$