Local Patterns in Data

Esther Galbrun

Spring 2023

UNIVERSITY OF
EASTERN FINLAND

Last updated on 2022-12-22 16:18

Q2.1: Quilt entropy

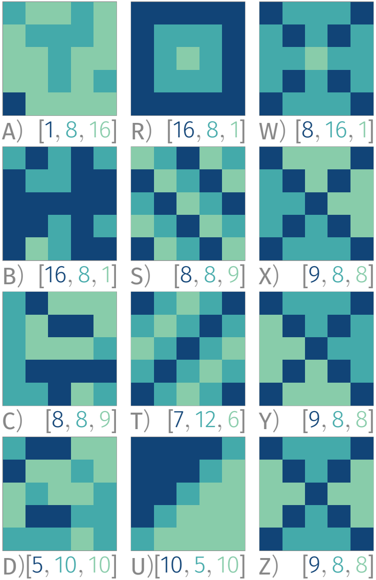
The colored quilt below can be thought of as a random variable taking one of three possible values in each cell

Let us ignore any possible dependencies between the cells

Compute the corresponding entropy

A)

Q2.2: Quilt entropies



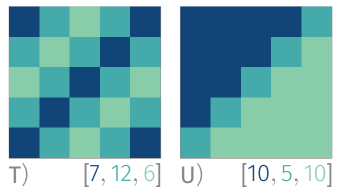
Associate each quilt to its entropy

$H(A) = ?$	$H(B) = ?$
$H(C) = ?$	$H(D) = ?$
$H(R) = ?$	$H(S) = ?$
$H(T) = ?$	$H(U) = ?$
$H(W) = ?$	$H(X) = ?$
$H(Y) = ?$	$H(Z) = ?$

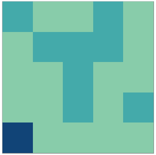
Q2.3: Conditional quilts

Let us consider the pairs of values appearing in corresponding positions of two quilts

Compute the corresponding joint entropy, conditional entropies and mutual information



Q2.4: Quilt inequalities (i)



A) [1, 8, 16]

B) [16, 8, 1]

C) [8, 8, 9]

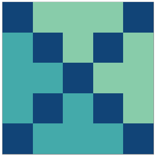
Fill in the following (in)equalities

$$H(A, B) = 2.078 \qquad H(A \mid B) = \ ?$$

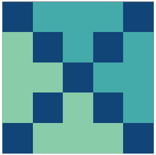
$$I(A; B) = \ ? \qquad H(B \mid A) = \ ?$$

$$I(A; C) \ ? \ H(A, C) \qquad H(A \mid C) \ ? \ H(C \mid A)$$

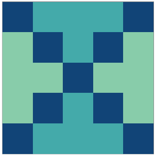
Q2.4: Quilt inequalities (ii)



X) $[9, 8, 8]$



Y) $[9, 8, 8]$



Z) $[9, 8, 8]$

Fill in the following (in)equalities

$$H(X \mid Y) = \ ? \quad I(X;Y) = \ ? \quad H(X \mid Z) \ ? \ H(Z \mid Y)$$

$$H(X,Y,Z) \ ? \ H(X,Z) \qquad H(X,Y,Z) \ ? \ H(Z \mid X,Y)$$

Q2.5: (De)code (i)

Consider the code $\mathcal{C}_1$ and bitstring $S =$`0010111011000110`

	a	b	c	d	e	f
$\mathcal{C}_1$	`11`	`10`	`01`	`001`	`010`	`0010`

What can you say about the following statements?

- ☐ The bitstring can be decoded uniquely with this code
- ☐ This is a prefix code
- ☐ There exists a prefix code with these code lengths

Q2.5: (De)code (ii)

Consider the code $\mathcal{C}_2$ and bitstring $S =$`0010111011000110`

	a	b	c	d	e	f
$\mathcal{C}_2$	`00`	`01`	`10`	`110`	`111`	`1001`

What can you say about the following statements?

- ☐ The bitstring can be decoded uniquely with this code
- ☐ This is a prefix code
- ☐ There exists a prefix code with these code lengths

Q2.5: (De)code (iii)

Consider the code $\mathcal{C}_3$ and bitstring $S =$`0010111011000110`

	a	b	c	d	e	f
$\mathcal{C}_3$	`11`	`10`	`01`	`001`	`0110`	`0101`

What can you say about the following statements?

- ☐ The bitstring can be decoded uniquely with this code
- ☐ This is a prefix code
- ☐ There exists a prefix code with these code lengths

Q2.5: (De)code (iv)

Consider the code $\mathcal{C}_4$ and bitstring $S =$`0010111011000110`

	a	b	c	d	e	f
$\mathcal{C}_4$	`00`	`01`	`110`	`111`	`101`	`1000`

What can you say about the following statements?

- ☐ The bitstring can be decoded uniquely with this code
- ☐ This is a prefix code
- ☐ There exists a prefix code with these code lengths

Consider the following message: badebfadebcdfabdefbcfcbc

and the two codes $\mathcal{C}_3$ and $\mathcal{C}_4$

	a	b	c	d	e	f
$\mathcal{C}_3$	`11`	`10`	`01`	`001`	`0110`	`0101`
$\mathcal{C}_4$	`00`	`01`	`110`	`111`	`101`	`1000`

What is the code length of this message, encoded with either of the two codes?

Q2.6: (Re)code (ii)

Consider the following message: badebfadebcdfabdefbcfcbc

and the two codes $\mathcal{C}_3$ and $\mathcal{C}_4$

	a	b	c	d	e	f
$\mathcal{C}_3$	`11`	`10`	`01`	`001`	`0110`	`0101`
$\mathcal{C}_4$	`00`	`01`	`110`	`111`	`101`	`1000`
nb. occs	3	6	4	4	3	4

Can you design a better code?

Q2.6: (Re)code (iii)

Consider the following message: badebfadebcdfabdefbcfcbc

	a	b	c	d	e	f
nb. occs	3	6	4	4	3	4

What is the theoretical optimal code length for this message?

Q2.7: (Trans)code (i)

Consider this transactional dataset, D

Let us ignore transaction delimiters
and consider ideal, not practical, codes

Let us encode the transactions
by assigning a codeword to each
distinct item, based on its frequency
We call this model M_0

tid	transaction
(1)	b
(2)	a d e
(3)	b f
(4)	a d e
(5)	b c d f
(6)	a b d e f
(7)	b c f
(8)	c
(9)	b c

What is the code length of the dataset encoded with M_0?

Q2.7: (Trans)code (ii)

Consider this transactional dataset, D

Let us ignore transaction delimiters
and consider ideal, not practical, codes

Let us encode the transactions
by assigning a codeword to
each distinct entire transaction,
based on its frequency
We call this model M_1

tid	transaction
(1)	b
(2)	a d e
(3)	b f
(4)	a d e
(5)	b c d f
(6)	a b d e f
(7)	b c f
(8)	c
(9)	b c

What is the code length of the dataset encoded with M_1?

Q2.7: (Trans)code (iii)

Consider this transactional dataset, D

Let us ignore transaction delimiters
and consider ideal, not practical, codes

Let us encode the transactions
by assigning codewords to itemsets
$\{\mathtt{ade}, \mathtt{bf}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, based on their frequency
We call this model M_2

tid	transaction
(1)	b
(2)	a d e
(3)	b f
(4)	a d e
(5)	b c d f
(6)	a b d e f
(7)	b c f
(8)	c
(9)	b c

What is the code length of the dataset encoded with M_2?

Q2.7: (Trans)code (iv)

Which is the best model for the dataset D?