INFORMATION THEORY
University of Amsterdam, 2018
TEAM NAME: Latin Square

Homework set 2

Contents

Question 1

(a) The possible outcomes of the experiment are (for A, B):

(1,1)	(1,3)	(1,5)
(2,2)	(2,4)	(2, 6)
(3, 1)	(3, 3)	(3, 5)
(4, 2)	(4, 4)	(4, 6)
(5, 1)	(5, 3)	(5, 5)
(6, 2)	(6, 4)	(6,6)

They are uniformly distributed e.g. they all have the same probability:

$$P_{AB}(ab) = \begin{cases} \frac{1}{18} & \text{if } a \in \{1,3,5\} \text{ and } b \in \{1,3,5\} \\ \frac{1}{18} & \text{if } a \in \{2,4,6\} \text{ and } b \in \{2,4,6\} \\ 0 & \text{otherwise} \end{cases}$$

(b)

- The outcome of X is always equal to zero, therefore $H(X) = 0$.
- For Y the probability distribution is:

  $$P_Y(y) = \begin{cases} \frac{1}{2} & \text{if } y = 0 \\ \frac{1}{2} & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

  Thus, $H(Y) = \frac{1}{2}\text{log}(2) + \frac{1}{2}\text{log}(2) = 1$
- For Z the probability distribution is:

  $$P_Z(z) = \begin{cases} \frac{4}{18} & \text{if } z = 4 \\ \frac{6}{18} & \text{if } z = 0 \\ \frac{8}{18} & \text{if } z = 2 \\ 0 & \text{otherwise} \end{cases}$$

  Thus, $H(Z) = \frac{4}{18}\log(\frac{18}{4}) + \frac{6}{18}\log(\frac{18}{6}) + \frac{8}{18}\log(\frac{18}{8})$

(c) We first compute the different probabilities:

- $P(Z = 0|A = 1) = \frac{P(A=1 \cap Z=0)}{P(A=1)} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}$
- $P(Z = 2|A = 1) = \frac{P(A=1 \cap Z=2)}{P(A=1)} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}$
- $P(Z = 4|A = 1) = \frac{P(A=1 \cap Z=4)}{P(A=1)} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}$

$H(Z|A = 1) = \frac{1}{3}\log 3 + \frac{1}{3}\log 3 + \frac{1}{3}\log 3 = \log(3)$

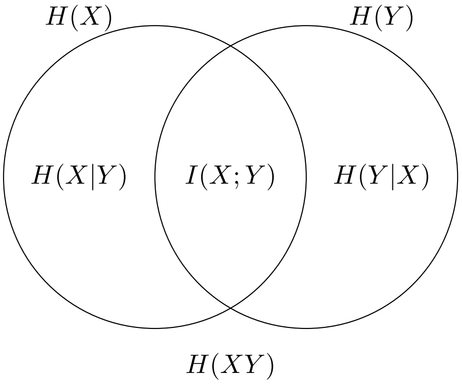
(d) Since AB is uniformly distributed we can make use of $H(X) = \log|\mathcal{X}|$, so $H(AB) = \log(18)$

(e) $$P_M(m) = \begin{cases} \frac{4}{18} & \text{if } m < 5 \\ \frac{5}{18} & \text{if } 5 < m < 7 \\ \frac{9}{18} & \text{if } m > 7 \\ 0 & \text{otherwise} \end{cases}$$

Thus, $H(M) = \frac{4}{18}\log(\frac{18}{4}) + \frac{5}{18}\log(\frac{18}{5}) + \frac{9}{18}\log(\frac{18}{9})$

Question 2

We have a joint pdf of X and Y. We can quickly use the table to work out an entropy diagram.



Where $H(X) = h(\frac{4}{7}) = H(Y) = h(\frac{4}{7}) = 0.985$, rounded at the third decimal.

Note that we have used h, the binary entropy function.
We have $H(XY) = \frac{3}{7}\log_2(\frac{3}{7}) + \frac{1}{7}\log_2(\frac{1}{7}) + \frac{1}{7}\log_2(\frac{1}{7}) + \frac{2}{7}\log_2(\frac{2}{7}) = 1.842$

The conditionals can be obtained by using the chain rule, which corresponds to the graphical representation above
$H(X|Y) = H(X,Y) - H(Y) = 1.842 - 0.985 = 0.857$ $H(Y|X) = H(X,Y) - H(X) = 1.842 - 0.985 = 0.857$
This makes sense since $H(X) = H(Y)$. Finally, we have $I(X;Y) = H(X) - H(X|Y) = 0.985 - 0.857 = 0.128$.

		Y	
		0	1
X	0	$\frac{3}{7}$	$\frac{1}{7}$
	1	$\frac{1}{7}$	$\frac{2}{7}$

Question 3

a We write down the conditional entropy $H(f(X)|X)$

$$H(f(X)|X) = \sum_{x \in \mathcal{X}} p(x) \sum_{x \in \mathcal{X}} p(f(x)|x) \log p(f(x)|x) \tag{1}$$

Now, since f is a function, it deterministically maps inputs s to values $f(x)$. Thus, we can consider the probability $p(f(x)|x) = 1$ for every $x \in \mathcal{X}$. This causes every log term to be zero, which makes the whole conditional entropy be indeed equal to zero.

b From the chain rule, we have that $H(X,Y) = H(X) + H(Y|X) = H(Y) = H(X|Y)$. We can use this fact together with the above proved $H(f(X)|X) = 0$ to prove the inequality. We have

$$\begin{aligned} H(X) + H(f(X)|X) &= H(f(X)) + H(X|f(X)) \\ H(X) &= H(f(X)) + H(X|f(X)) \\ H(X) &\geq H(f(X)) \end{aligned}$$

As the conditional entropy $H(X|f(X))$ is always non negative.

Question 4

a We apply Jensen's inequality to the negative relative entropy $-D(P||Q)$.

$$\begin{aligned}-D(P||Q) = -\sum_{x\in\mathcal{X},P(x)>0} P(x)\log\frac{P(x)}{Q(x)} &= \sum_{x\in\mathcal{X},P(x)>0} P(x)\big(-\log\frac{P(x)}{Q(x)}\big) = \sum_{x\in\mathcal{X},P(x)>0} P(x)\log\frac{Q(x)}{P(x)}\\ \le \log\big(\sum_{x\in\mathcal{X},P(x)>0} P(x)\frac{Q(x)}{P(x)}\big) &= \log\big(\sum_{x\in\mathcal{X},P(x)>0} Q(x)\big) \le \log\big(\sum_{x\in\mathcal{X}} Q(x)\big) = \log 1 = 0\end{aligned}$$

We have found that $-D(P||Q) \le 0$, which means that $D(P||Q) \ge 0$

We want to show now that $D(P||Q) = 0$ if and only if $P = Q$.

If $D(P||Q) = 0$, then the second inequality must be an equality, so $\{x \in \mathcal{X}, P(x) > 0\}$ is the same as $\{x \in \mathcal{X}\}$ which means that for all $x \in \mathcal{X}, P(x) > 0 \implies Q(x) > 0$. This is due to the fact that Q has to sum up to one, and means that P and Q have the same support. Also, Jensen's inequality must be an equality in this case, and it follows that $\frac{Q(x)}{P(x)} = c$ for all $x \in \mathcal{X}, P(x) > 0$. So, we must have that for each element, $Q(x) = cP(x)$. Summing over all elements of the support, we get that as P and Q are probability distributions (in particular, they sum up to one), c must be 1, so $Q(x) = P(x)$ for all $x \in \mathcal{X}, P(x) > 0$. This, together with the fact that the two distributions have the same support, proves that indeed $P = Q$.

If $P = Q$, then the equality condition of Jensen's inequality tells us that in this case we have an equality. For the second inequality, we have

$$\log\big(\sum_{x\in\mathcal{X},P(x)>0} Q(x)\big) = \log\big(\sum_{x\in\mathcal{X},P(x)>0} P(x)\big) = \log 1$$

which means that also the second inequality is an equality in this case.

Thus, we have found that equality $D(p||q) = 0$ holds if and only if $P = Q$

b First, we know that $I(X;Y) = D(P_{XY}||P_XP_Y)$, which implies $I(X;Y) \ge 0$. Then, we know that $I(X;Y) = H(X) - H(X|Y)$. So, we have that

$$\begin{aligned}H(X) - H(X|Y) &= I(X;Y)\\ H(X) - H(X|Y) &\ge 0\\ H(X) &\ge H(X|Y)\end{aligned}$$

c

$$\begin{aligned}I(X;Y) = D(P_{XY}||P_XP_Y) &= \sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\\ &= \sum_x\sum_y p(x,y)\log p(x,y) - \sum_x\sum_y p(x,y)\log[p(x)p(y)]\\ &= -H(X,Y) + H_C(P_{XY}, P_XP_Y)\\ &= H_C(P_{XY}, P_XP_Y) - H(X) - H(Y|X)\\ &= H_C(P_{XY}, P_XP_Y) - H(Y) - H(X|Y)\end{aligned}$$

Question 5

(a) We will expect higher entropy compare to sampling a single letter, because there are a lot of more words then letters. Moreover some words are very unusual so, when we see them, we are very "surprised". To sum up, there are more surprising words then characters, so from this we can conclude that entropy of the words will be higher.

(b) $H(X_{words}) = 8.341112484463014$

(c) To count letters, which are first, we count all letters except the last one (because the last letter doesn't have follower, so it can't be first letter).

Similarly, to count the second letters, we count all of them except the first one (because this letter doesn't have predecessor).

If we compare the probability distributions, they are same, because one letter difference doesn't make a big influence.

Why can we count all the middle characters (expect the first one and the last one for X and Y)? Each one of these characters belongs to X, because it can be first character and also it belongs Y, because it can be the second character.

(d) $H_C(P_{XY}, P_X.P_Y) = 8.194720759582626$

(e) We hypothesise that X and Y are independent. This gives us distribution of pair letters $= P_X.P_Y$. We would like to compare dist distribution the to real world two letters in row distribution P_{XY}. So we calculate $H_C(P_{XY}, P_X.P_Y)$.

So, it tells us how surprised we are, if we expect that first and second letter are independent compare to the real data.

Code explanation: We chose Haskell again. because of his easy implementation of the entropy function and possibility to abstract these entropy function so they can be used over Hashmaps, Lists, Trees .. (everything, which is Foldable and has Functor instance.)

Firstly, we needed to deal with counting characters, double characters and words. Before counting, we lowered whole text and filter out not ASCII characters and keeping spaces and newlines. Secondly, we created 3 functions.

- countWords: this function has counted number of words. (This was easy, we just used **word** function)
- countDoubles: this function count occurrences of two characters in the row. Here we had to filter our new line characters
- countCharacters: this function counts number of characters. Also, it filters out the newline charaters.

All these function takes as preformat text, which was obtain from previous formating text function.

To calculate $P_X.P_Y$, we calculate Cartesian product between X and X. (We created a new Map from X's Map, where keys are join characters and probabilities are multiplied.) Example: X = [P('a') = 0.5, P('b') = 0.5], so the Cartesian product is : [P('aa') = 0.25, P('ab') = 0.25, P('ba') = 0.25, P('bb') = 0.25]

Secondly, we needed to verify our two functions, cross_entropy and entropy. To verify entropy, we verify that entropy of uniform distribution is equal to $log_2(N)$, where N is number of elements.

Lastly, we verified the cross_entropy function, knowing that $H(P_x) = H_C(P_x, P_x)$, where P_x is some distribution.