INFORMATION THEORY
Master of Logic, Master AI, Master CS, University of Amsterdam, 2018
TEACHER: Christian Schaffner, TAs: Yfke Dulek, Esteban Landerreche, Kyah Smaal

Team homework set 4

Unless otherwise stated, you should provide exact answers rather than rounded numbers (e.g., $\log 3$ instead of 1.585) for non-programming exercises.

Problem 1: Calculation of the typical set (7pt)

To clarify the notion of a typical set $A_\varepsilon^{(n)}$ and the smallest set of high probability $B_\delta^{(n)}$, we will calculate these sets for a simple example. Consider a sequence of i.i.d. binary random variables $X_1, X_2, ...X_n$, where the probability that $P_X(1) = 0.6$ and $P_X(0) = 0.4$.

(a) **(3pt)** With $n = 25$ and $\varepsilon = 0.1$, which sequences fall in the typical set $A_\varepsilon^{(n)}$? What is the probability of the typical set (three decimals)? How many elements are there in the typical set? (This involves computation of a table of probabilities for sequences with k 1's, $0 \le k \le 25$, and finding those sequences that are in the typical set.)
Hint: Here is the table: `http://goo.gl/sQCPM0`

(b) **(2pt)** How many elements are there in the smallest set that has probability 0.9? In other words, what is $|B_\delta^{(n)}|$ for $n = 25$ and $\delta = 0.1$?

(c) **(2pt)** How many elements are there in the intersection $|A_\varepsilon^{(n)} \cap B_\delta^{(n)}|$ of the sets computed in parts (b) and (c)? What is the probability of this intersection (three decimals)?

Problem 2: Three random variables (3pt)

Let A, B, C be random variables such that

$$I(A;B) = 0$$
$$I(A;C|B) = I(A;B|C)$$
$$H(A|BC) = 0$$

Prove one of the following possible relations between $H(A)$ and $H(C)$: $=, \le, \ge, <,$ or $>$. If you show one of the weaker inequalities ($\le$ or $\ge$), also exhibit an example that shows why the inequality is not strict.

Problem 3: Piece of cake (3pt)

A big cake is repeatedly sliced into two pieces. At every step, the *smaller* part is discarded (or eaten). On the other part the process is continued. At every step, the remaining piece is cut randomly into two pieces with the following proportions:

$$\left(\frac{2}{3}, \frac{1}{3}\right) \text{ with probability } \frac{3}{4},$$
$$\left(\frac{3}{5}, \frac{2}{5}\right) \text{ with probability } \frac{1}{4}.$$

For example, three consecutive cuts might result in a piece of cake of size $\frac{3}{5} \cdot \frac{2}{3} \cdot \frac{2}{3}$. Let T_n be the fraction of the cake left after n cuts. Clearly, this fraction T_n decreases exponentially with n, i.e. $T_n \approx c^n$ for some real constant $0 < c < 1$. Determine c for large values of n.

Hint: Let C_i be a random variable describing the fraction of the cake that is cut (and kept) at the ith cut, i.e. $C_i = \frac{2}{3}$ with probability $\frac{3}{4}$, or $\frac{3}{5}$ otherwise. Then use the weak law of large numbers.

Problem 4: Programming (7pt)

In one of the quizzes for this week, you implemented the encryption and decryption procedure for the Vigenère cipher. (Note: we treat "`A`" as 1 instead of 0. That is, `K+A=L`)

(a) **(1pt)** Consider this cipher for a fixed message length m and a fixed key length k. Suppose the possible messages are uniformly distributed over all monocase messages (excluding spaces; there are 26 possible letters) of length m. What is $I(M;C)$? When is the Vigenère cipher perfectly secure (i.e., $I(M;C) = 0$)?

In reality, the possible messages are not uniformly distributed. We can use this fact to break a ciphertext that was encoded with the Vigenère cipher.

Recall the definition of collision probability from the first homework:

$$Coll(P) := \sum_{x \in \mathcal{X}} P(x)^2.$$

Let $Coll(P_M)$, $Coll(P_K)$, and $Coll(P_C)$ denote the collision probabilities of (sampling a single letter form) the plaintext, key, and ciphertext, respectively. Suppose you know that the message is in English. In the first homework, you estimated that $Coll(P_M) \approx 0.0655$.

(b) **(2pt)** Show that

$$Coll(P_C) \approx \frac{1}{k} Coll(P_M) + \frac{k-1}{k} Coll(P_K)$$

by assuming that:

1. the message is very long compared to the key (m is very large), and
2. the letters in the message are i.i.d. according to P_M.
3. the letters in the key are i.i.d. according to a uniform P_K.

(c) **(2pt)** Use (b) to find the most likely key length k in the following ciphertext:

```
QMIXFFKFMVPZSVYGOMAPOTTCUUBGIIPPOTPZEVNVJHIUBGSCDCIPXBKDBR
SIXFMVIUVJBCWYNPMVNSDFNDZDWQCVFLKZCGMXYJNJBAQPXISUEFVGGIT
FMQNOEIATBWG
```

(d) **(2pt)** The ciphertext in the previous subexercise was sent to Cleopatra by Caesar. That means it will very likely contain one or more of the following words: `CLEOPATRA`, `CAESAR`, `QUEEN`, `GRACE`, `ALEXANDRIA`. Use this side information, in combination with the key length you computed in (c), to figure out the original message. What was the encryption key?