EE2340/EE5847: Information Sciences/Information Theory **2022**

Homework 1: 12th Jan 2022

Instructor: Shashank Vatedka

Instructions: *You are encouraged to discuss and collaborate with your classmates. However, you* must explicitly *mention at the top of your submission who you collaborated with, and all external resources (websites, books) you used, if any. Copying is NOT permitted, and solutions must be written independently and in your own words.*

Please scan a copy of your handwritten assignment as pdf with filename <your ID>_HW<homework no>.pdf. Example: EE19BTECH00000_HW1.pdf.

For programming questions, create separate files. Please use the naming convention <your ID>_HW<homework no>_problem<problem no>.. Example: EE19BTECH00000_HW1_problem1.c.* You may upload c,cpp,py or m files only. No other format will be allowed.

Exercise 1.1 (Probability refresher, nothing to submit)**.** Review the following concepts from your probability course:

- Probability mass function
- Probability density function, probability distribution function
- Random variables: continuous, discrete
- Functions of random variables
- Common random variables: Bernoulli, Binomial, Poisson, Exponential, Gaussian
- Bayes theorem
- Expectation, expected values
- Moments, moment generating function
- Markov's inequality, Chebyshev's inequality, Chernoff bounds

Also go through the lecture notes for the notation. Whenever you have to use a Chernoff bound, see if you can use the simplified version for Bernoulli random sequences.

Exercise 1.2 (Conditional probability, 25pts)**.** Let us assume that the RT-PCR test for COVID has the following performance metrics:

- If a patient has COVID, then the test returns positive 67% of the time
- If a patient does not have COVID, then the test returns negative 99% of the time.

Similarly, assume that the Rapid Antigen Test (RAT) has the following metrics:

- If a patient has COVID, then the test returns positive 50% of the time

- If a patient does not have COVID, then the test returns negative 95% of the time.

Now assume that 5% of the population has COVID. If a random person is picked from the population and

(a) tested positive using RT-PCR,

(b) tested positive using RAT,

(c) tested positive with both RT-PCR and RAT,

then what is the probability that he/she is actually infected? You may assume that the test results of RT-PCR and RAT are statistically conditionally independent of each other given the actual COVID status of the person.

(d) How would the answers change if the following change: RT-PCR returns negative 90% of the time if the patient does not have COVID and RAT returns positive 60% of the time if the patient has COVID

(e) If we sample 100 people at random (assume that the population is large, and we are sampling with replacement to simplify things), then what is the expected number of positives we would see if we tested only using RT-PCR with the original figures? How does this compare to the ground truth (that 5% has COVID)?

Note: The above numbers are not accurate and our assumptions are simplistic, so one must not look too much into the final result. However, it is very important to understand conditional probability and its implications as this problem suggests.

Hint: You may look at example 2.3 in the book by David Mackay

Exercise 1.3 (The empirical frequency is close to the true probability if the number of samples is large, 17pts)**.** Given a sequence $x^n = (x_1, x_2, \ldots, x_n)$, where each $x_i \in \mathcal{X}$, the relative frequency/empirical frequency of occurrence of $a \in \mathcal{X}$ is defined as

$$\mu_a(x^n) = \frac{\text{Number of times } a \text{ occurs in } x^n}{n}$$

For example, if $x^n = (a, b, a, a, b, c)$, then $\mu_a(x^n) = 1/2$, $\mu_b(x^n) = 1/3$, and $\mu_c(x^n) = 1/6$. Observe that $(\mu_a(x^n) : a \in \mathcal{X})$ forms a valid probability mass function. This is also called the empirical pmf.

Consider a sequence of n i.i.d. random variables $X_1, X_2, \ldots, X_n$ over the alphabet $\mathcal{X}$ and having distribution p_X.

1. What is $\mathbb{E}\mu_a(X^n)$ for any $a \in \mathcal{X}$?
   Hint:: If it helps, define Y_i to be a random variable which is equal to 1 if $X = a$ and zero otherwise. Can you express μ_a in terms of this?

2. Use the Chebyshev inequality and Chernoff bound (simplified version for Bernoulli random variables) to obtain upper bounds on
   $$\Pr[|\mu_a(X^n) - p_X(a)| > \delta p_X(a)]$$
   for a given $a \in \mathcal{X}$.

3. Using this, obtain upper bounds on
   $$\Pr[|\mu_a(X^n) - p_X(a)| > \delta p_X(a)]$$

4. If X^n is stationary but not iid, then what can you say about $\mu_a(X^n)$? Are the bounds you obtained previously valid for this case?

Exercise 1.4 (Typical Hamming distance, 20pts)**.** Consider a binary sequence $x^n = (x_1, x_2, \ldots, x_n)$. Let Y^n be obtained by flipping each x_i with probability p. In other words,

$$Y_i = \begin{cases} x_i & \text{with probabiliy } 1-p \\ \bar{x}_i & \text{with probability } p \end{cases}$$

where $\bar{x}_i$ is 1 if $x_i = 0$ and 0 if $x_i = 1$. Let $d_H(x^n, Y^n)$ be equal to the number of locations i in which $x_i \neq Y_i$. This is called the Hamming distance between the sequences.

1. What is $\mathbb{E}[d_H(x^n, Y^n)]$?
2. Can you bound $\Pr[d_H(x^n, Y^n) \geqslant np(1+\epsilon)]$ for any $\epsilon > 0$? What happens to this as $n \to \infty$? (You can use Chernoff bound)
3. Obtain an exact expression (can be a summation of terms involving n, p, ϵ) for $\Pr[d_H(x^n, Y^n) \geqslant np(1+\epsilon)]$.
4. Evaluate (2) and (3) for $n = 100$, $p = 0.3$, $\epsilon = 0.2$.

Exercise 1.5 (Bounds on the binomial coefficient, 22pts)**.** Let us prove some results that will be useful in the rest of the course.

1. Show that for all positive integers (n, k, l) with $n > k > l$, we have $\frac{n-l}{k-l} \geqslant \frac{n}{k}$.
2. Prove the following bounds for the binomial coefficient:

$$\left(\frac{n}{k}\right)^k \leqslant \binom{n}{k} \leqslant \left(\frac{n^k}{k!}\right).$$

3. Prove that for all positive integers $n > k > 0$, we have

$$\frac{\sqrt{2\pi}}{e^2} \frac{1}{\sqrt{np(1-p)}} 2^{nH_2(p)} \leqslant \binom{n}{k} \leqslant \frac{e}{2\pi} \frac{1}{\sqrt{np(1-p)}} 2^{nH_2(p)}$$

   where $p = k/n$ and $H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$.The term $H_2(p)$ is called the binary entropy of p.

   You may use (without proof) Stirling's approximation:

$$\sqrt{2\pi} n^{n+1/2} e^{-n} \leqslant n! \leqslant e n^{n+1/2} e^{-n}.$$

4. Use the above to conclude for any $0 < p \leqslant 1/2$,

$$\lim_{n\to\infty} \frac{1}{n} \log_2 \binom{n}{\lfloor np \rfloor} = H_2(p).$$

Exercise 1.6 (Simple variable-length compressor, 16pts)**.** You will now implement a very simple compressor for some randomly generated files (without actually going through the design).

The file given to you contains a sequence consisting of a, b, c, d, e. For the file given to you, compute the empirical pmf $\mu_x(x^n)$, and use this to calculate the empirical entropy.

$$H_{\text{emp}} = \sum_{x \in \{a,b,c,d,e\}} \mu_x(x^n)$$

The compressor performs the following one-to-one map:

$$a \mapsto 0$$
$$b \mapsto 10$$
$$c \mapsto 110$$
$$d \mapsto 1110$$
$$e \mapsto 1111.$$

For example, the compressed sequence for *abaac* is 01000110.

Find the compressed binary sequence for the file given to you. Submit this in a separate text file with the specified filename. What is the length of the compressed sequence?

Now compress the original file using zip. What is the compressed filesize in bits? How do both these quantities compare to the entropy?

(Bonus) Can you think about how to decompress a file encoded using the above technique? Do you think any one-to-one mapping from $\mathcal{X}$ to unique binary sequences can be decompressed?