Homework 5

Kevin Chao

November 2019

Question 1:

a. Let's organize our information first. The mean time = t hours, and the mean time can be found given the formula below

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The expected mean is $E[M_n] = t$. The variance of a single student is 3, which means that the variance of $M_n is$

$$M_n = \frac{1}{n}\sum_{i=1}^{n} Var(X_i)$$
$$M_n = \frac{3}{n}$$

Since the problem is looking for an n that guarantees .99 probability within 30 minutes of M_n, the comparison would be

$$P(t - .5 \leq M_n \leq t + .5) \geq .99$$
$$P(\mid M_n - t \mid > .5) \leq .01$$

And now, using Chebyshev's inequality:

$$P(\mid M_n - t \mid > .5) \leq \frac{3}{n(.5)^2}$$

Now, we can compare the two inequalities.

$$\frac{3}{n(.5)^2} \leq .01$$
$$\frac{3}{.01(.5)^2} \leq n$$
$$1200 \leq n$$

b. Since M_n is now a normal distribution, it can written like in (a)

$$P(|M_n - t| > .5) \leq .01$$

The standard deviation is $\sqrt{\dfrac{3}{n}}$, which means that it is now

$$P(|\tfrac{M_n - t}{\sqrt{3/n}}| \geq \tfrac{.5}{\sqrt{3/n}}) < .01$$
$$\approx P(|Z| \geq \tfrac{.5\sqrt{n}}{\sqrt{3}})$$

Now, we can find that

$$P(|Z| \geq \tfrac{.5\sqrt{n}}{\sqrt{3}}) = \Phi(-\tfrac{.5\sqrt{n}}{\sqrt{3}}) + 1 - \Phi(\tfrac{.5\sqrt{n}}{\sqrt{3}}) = 2\Phi(-\tfrac{.5\sqrt{n}}{\sqrt{3}})$$

$P(|Z| \geq \tfrac{.5\sqrt{n}}{\sqrt{3}}) = q$, and now it can be simplified to

$$-\tfrac{.5\sqrt{n}}{\sqrt{3}} = \Phi^{-1}(\tfrac{q}{2})$$
$$\sqrt{n} = -\tfrac{\sqrt{3}}{.5}\Phi^{-1}(\tfrac{q}{2})$$

Using scipy.stats.norm.ppf, the answer is n = 80.

c. CLT's approximation of trials is much lower than Chebyshev's inequality approximation of trials. The reason for this is that Chebyshev's works for any probability distribution, while CLT has to have stronger assumptions (note how you have to assume that M_n is normal).

Question 2:

a. Since X, Y, and C are all binary, this is a Bernoulli distribution and thus, can find the entropy with the equation:

$$H(X) = plog_2(\tfrac{1}{p}) + (1-p)log_2(\tfrac{1}{1-p})$$

So for X, P(X=1) = .6 and P(X=0) = .4 and is now

$$H(X) = .6log_2(\tfrac{1}{.6}) + .4log_2(\tfrac{1}{.4})$$
$$H(X) = .97095$$

And for Y, P(Y=1) = .65 and P(Y=0) = .35, so H(Y) is

$$H(Y) = .65log_2(\tfrac{1}{.65}) + .35log_2(\tfrac{1}{.35})\ H(Y) = .93406$$

Thus, X has a larger uncertainty from its larger entropy.

b. I(X;C) = H(C) - H(C—X), however it is symmetric and thus can be rewritten as

I(X;C) = H(X) - H(X—C)

Conditional entropy is $H(X|C) = \sum_{i=0}^{1} p(C = i)H(X|C = i)$, and $P(C = 1) = .4$ and $P(C = 0) = .6$ and so

$$\begin{gathered}H(X|C) = .6 * H(.3, .3) + .4 * H(.1, .3)\\ H(.3, .3) = -[.3log_2(.3) + .3log_2(.3)]\\ H(.3, .3) = -[-.52109 + -.52109]\\ H(.3, .3) = 1.0422\\ H(.1, .3) = -[.1log_{.1} + -.52109]\ H(.1, .3) = .85328\\ H(X|C) = .6 * 1.0422 + .4 * .85328\ H(X|C) = .96663\end{gathered}$$

and with H(X) in (a), then I(X;C) is

$$I(X; C) = .97095 - .96663 = .00432$$

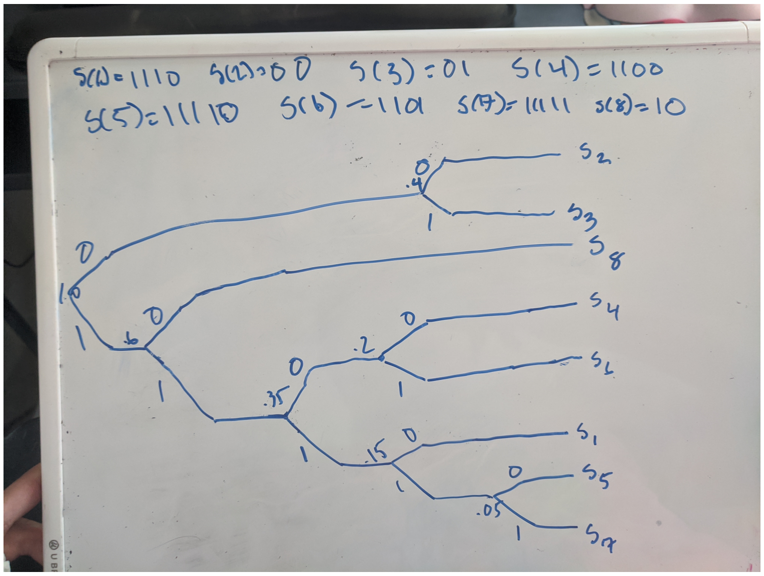
With I(Y;C), it is similar to above

$$\begin{aligned}H(Y|C) = .6 * H(.3, .3) + .4(.05, .3)\\ H(.05, .3) = -[.05 * log_2(.05) + .3 * log_2(.3)]\\ H(.05, .3) = .21610 + .52109\\ H(.05, .3) = .73719\\ H(Y|C) = .6 * 1.0422 + .4 * .73719\\ H(Y|C) = .92020\\ I(Y; C) = .93406 - .92020\\ I(Y; C) = .01386\end{aligned}$$

c. When mutual information(X,Y) = 0, that means that X and Y are independent. With that, Y would have a stronger dependence on C since it has a higher deviation from 0 than X in terms of mutual information (This is given from (b) where I(Y;C) ¿ I(X;C))

Question 3:

a. Using a fixed length code for S, with 3 variables and each variable being binary, then the codeword length is 3, as it could be S(1) = 000, S(5) = 001

b. 

S(1)=1110 S(2)=00 S(3)=01 S(4)=1100

S(5)=11110 S(6)=1101 S(7)=11111 S(8)=10

c. L(C) = (.2 + .25 + .2)*2 + (.1 + .1 + .1)*4 + .05*5 = 2.75

d. To find entropy, we do p*log(p) for each S_i, and then add.

$$H(S) = .1log(.1) * 3 + .2log(.2) * 2 + .25log(.25) + .05log(.05)$$

$$H(S) = 2.6414$$

The expected length, L(C) is typically bigger than the entropy, and as such it followed that $H(S) \leq L(C)$, even for the smallest possible expected length.

Question 4:

a. The entropy from the tortoise and the hare characters is 4.2847. The code for getEntropy() is below.

```
def getEntropy(freqs_dict):
    pmf = np.array(list(freqs_dict.values()))
    pmf = pmf/np.sum(pmf)
    entropy = 0
    for i in range(0,len(pmf)):
        entropyI = pmf[i] * math.log2(pmf[i])
        entropy = entropy + entropyI
    entropy = entropy * -1
    return entropy
```

b. The average length of bits for the tale was 4.32739. The average code length aka the expected codeword length is slightly bigger than the entropy of the text, which does follow $H(X) \leq L(C)$, which is more than likely one of the more optimal expected length given how close H(X) is to L(C).

c.

```
def decode(binaryS, huffmanTree):
    ls = []
    cur = huffmanTree
    text = binaryS + '2'
    for i in text:
        if cur.item != None:
            ls.append(cur.item)
            cur = huffmanTree
            if i == '1':
                cur = cur.right
            elif i == '0':
                cur = cur.left
        else:
            if i == '0':
                cur = cur.left
            elif i == '1':
                cur = cur.right
            else:
                continue
```

d. The decoded message says ”peter anteater”

e. Applying getEntropy to webFreqs, the entropy is 2.3562

f. The decode output for the web traffic data was [1, 4, 5, 62, 99, 34]

g. The code to find expected once the tree and codes have been set is as shown:

```
expected = 0
for i in range(1,101):
    length = len(webCodes[i])
    prob = webProbs[i-1]
    expected = expected + (length * prob)
```

The expected length is 2.3966. Again, this follows the same principle as (b)'s Huffman codes; $L(C) \geq H(X)$. It is very close to optimal (if not, then it is optimal), as it is extremely close to the entropy (the closer the expected/average length of codeword is to the entropy, the more optimal it is).