ISTA 311 Written Homework 5

Due: Thursday, November 14

Complete the problems below. "LN" means the lecture notes. Submit this homework handwritten on a separate sheet or sheets of paper.

If you collaborate with another student on this assignment, please note the name of your collaborator(s) on your paper.

1. You are told that when a pair of fair dice are rolled, the sum of the dice was (a) 2 (b) 7. How much information is there in each of the two messages?

2. We have three cards. One is red on both sides; one is blue on both sides; one is red on one side and blue on the other side. The cards are shuffled and their orientation randomized. One card is drawn and placed on the table; the side facing up is blue.
   (a) Compute the information content of this observation (that the side facing up is blue.)
   (b) Find the prior and posterior probability distributions of the color of the bottom side of the card. (Use Bayes' theorem to calculate the posterior probability distributions.)
   (c) Find the entropy of each of the probability distributions you calculated in part (a).

3. In class we discussed a simplified version of the game *Battleship* where the only target on the board is a single 1×1 "submarine".

   Now, consider the version where a single 1×2 "destroyer" is placed on the board. The destroyer may be placed horizontally or vertically. Assume that the position of the destroyer is chosen uniformly at random from all legal positions.
   (a) Let P be the random variable representing the position of the destroyer. Find the entropy $H(P)$ of this random variable.
   (b) Label the coordinate grid by the letters A-H (columns) and the numbers 1-8 (rows). Suppose your first shot is $C3$ and your second shot is $D3$; both hit (so you win the game in two turns). Calculate the information content of this pair of events.
   (c) Now suppose that your first shot is $F1$ and your second shot is $F2$; both hit. Calculate the information content of this pair of events.

4. The frequency of letters in written English is given approximately by the following table:

Letter	Freq.	Letter	Freq.
`a`	0.0575	`n`	0.0596
`b`	0.0128	`o`	0.0689
`c`	0.0263	`p`	0.0192
`d`	0.0285	`q`	0.0008
`e`	0.0913	`r`	0.0508
`f`	0.0173	`s`	0.0567
`g`	0.0133	`t`	0.0706
`h`	0.0313	`u`	0.0334
`i`	0.0599	`v`	0.0069
`j`	0.0006	`w`	0.0119
`k`	0.0084	`x`	0.0073
`l`	0.0335	`y`	0.0164
`m`	0.0235	`z`	0.0007
`space`	0.1928		

I expect you to do the following calculations on a computer, so you don't need to show the details; just include the important numbers. To save you typing, there is a file on D2L called `letterdict.py` containing the above probability distribution as a dictionary.

(a) Compute the entropy per character of a random source generating characters selected independently from the above distribution.[1]

(b) Shannon's *source coding theorem* states (informally) that an information source with entropy H bits/symbol can be encoded into a binary sequence with an average of H bits/symbol without losing information. ASCII text uses a fixed 8 bits/symbol, but your answer to part (a) should be substantially less than 8. This implies that if the distribution above is a good model for English, a text file could be compressed to require a smaller amount of space.

Based on your answer to part (a), by what factor could a typical text file be compressed?

(c) Download the file `hp1_cleaned.txt` from D2L; this file contains the text of *Harry Potter and the Philosopher's Stone*, with all characters other than letters and punctuation removed and all letters made lowercase, so that it may be modeled by the above distribution.

Use a file compression tool such as `zip` to compress this file, and note the difference in size between the plain text and the zipped file. Does the size reduction agree with your answer to part (b)? If not, can you propose a reason for the discrepancy?

[1] Obviously, this shouldn't be taken as a good model of English text.

Solutions

1. (a) The probability of rolling a 2 is 1/36, so the information content is $-\log_2(1/36) = \log_2(36) \approx 5.170$.

   (b) The probability of rolling a 7 is 6/36, so the information content is $-\log_2(1/6) = \log_2(6) \approx 2.585$.

2. (a) The probability that the visible face is blue is 1/2, so the information content is 1 bit.

   (b) The prior probability of the bottom face being blue is 1/2. Considering three hypotheses, R/R; R/B; B/B, corresponding to the identity of the card, we find that the posterior probabilities are

$$P(B/B|\text{observed blue face up}) = \frac{P(\text{observed blue}|B/B)P(B/B)}{P(\text{observed blue}|B/B)P(B/B) + P(\text{observed blue}|R/B)P(R/B)} = \frac{2}{3}$$

$$P(R/B|\text{observed blue face up}) = \frac{P(\text{observed blue}|R/B)P(R/B)}{P(\text{observed blue}|B/B)P(B/B) + P(\text{observed blue}|R/B)P(R/B)} = \frac{1}{3}$$

   Since the bottom face is blue if we have the blue/blue card and red otherwise, the posterior probability distribution of the bottom face color is $P(B) = 2/3, P(R) = 1/3$.

   (c) The entropy of the prior distribution is 1 bit, and the entropy of the posterior distribution is $\frac{1}{3}\log_2(3) + \frac{2}{3}\log_2(3/2) \approx 0.918$ bits.

3. (a) There are 112 legal positions, all equally likely, so the entropy is $\log_2(112) \approx 6.807$. To see this, consider vertical and horizontal alignments separately. For vertical alignments, identify a position by its upper square; for horizontal, identify a position by its left square. Then, for vertical squares, all but the bottom row are valid locations to place the ship; for horizontal, all but the rightmost column. So, there are $64-8 = 56$ vertical placements and the same number of horizontal placements; $56 + 56 = 112$.

   (b) The probability that the first shot hits is 4/112, since there are 4 legal placements that intersect the space $C3$. For the same reason, the probability that the second shot hits is 1/4, because $D3$ is one of the four equally likely possibilities for the other square of the ship. Therefore, the information is

$$\log_2(112/4) + \log_2(4) = \log_2(112) \approx 6.807$$

   (c) There are three legal placements intersecting $F1$, so the probability that the first shot hits is 3/112. For the same reason, the probability that the second shot hits is 1/3. Therefore, the information content is

$$\log_2(112/3) + \log_2(3) = \log_2(112) \approx 6.807$$