Topics in Data Science, Data Protection — Spring 2020
CS 4364 — L. Longpré

Class notes 2/20/2020

1 Entropy

We consider a source that issues a sequence of messages from a finite set of possible messages. In our examples, "messages" are characters or bytes, and we use these different views interchangeably.
Let $X_1, ..., X_n$ be n possible messages occurring with probabilities $p(X_1), ..., p(X_n)$, where $\sum_{i=1}^{n} p(X_i) = 1$. The *entropy* of a given message is defined by the weighted average:

$$H(X) = -\sum_{i=1}^{n} p(X_i) \lg p(X_i) = \sum_{i=1}^{n} p(X_i) \lg \left(\frac{1}{p(X_i)} \right).$$

Entropy is a theoretical limit on how a sequence of characters can be compressed.
Examples for entropy

1. {Male, Female} with equal probability: 1.
2. {a, b, c} with probabilities $1/2, 1/4, 1/4$: 1.5.
3. Set of size n with each probability $1/n$: $\lg n$.
4. {a} with probability 1: 0.

Given a message Y in the set $\{Y_1, ..., Y_m\}$, where $\sum_{i=1}^{m} p(Y_i) = 1$, let $p_Y(X)$ be the conditional probability of message X given message Y.

$$p_Y(X) = p(X \wedge Y)/p(Y)$$

Example: X = probability of UTEP student is female. Y = probability UTEP student is CS major.
Conditional entropy of X given Y:

$$H_Y(X) = \sum_{X,Y} p_Y(X) p(Y) \lg \left(\frac{1}{p_Y(X)} \right). \quad (1)$$

$$= \sum_{Y} p(Y) \sum_{X} p_Y(X) \lg \left(\frac{1}{p_Y(X)} \right). \quad (2)$$

Example:

$$p(X \wedge Y) = 0.01 \quad (3)$$
$$p(X \wedge \overline{Y}) = 0.03 \quad (4)$$
$$p(\overline{X} \wedge Y) = 0.48 \quad (5)$$
$$p(\overline{X} \wedge \overline{Y}) = 0.48 \quad (6)$$

Exercise 1

What is $H_Y(X)$?

2 Huffman codes

As seen in class, with a:45, b:13, c:12, d:16, e:9, f:5. Huffman code: a:0, b:101, c:100, d:111, e:1101, f:1100

Exercise 2

Redraw the tree

Entropy was about 2.22. Huffman code average length of code: 2.24.

Exercise 3:

Using Huffman code to compress: how to represent the optimal code with the fewest number of bits?

Exercise 4:

The following is a representation of a code using the algorithm seen in class. The possible characters are a, b, c, d, e, f, g, each represented by 3 bits. What is the code?
00100010110010010110111100101011100