

050.680 — Learning Theory
Problem Set 2
Due 23 February 2007

Problem 1

Prove that $D(p||\hat{p}) = H(p, \hat{p}) - H(p)$.

Problem 2

The frequency p_n of the n th most frequent word in English is roughly approximated by

$$p_n \approx \begin{cases} \frac{0.1}{n} & n \in [1, 12367] \\ 0 & n > 12367 \end{cases}$$

This remarkable $1/n$ law is known as Zipf's law, and applies to the word frequencies of many languages (Zipf, 1949). Let us assume that English is generated by picking words at random according to this distribution. Using Matlab, compute the entropy of English.

Problem 3

Consider the following joint probability distribution for random variables X and Y :

$p(x, y)$		x			
		1	2	3	4
1		1/16	1/8	1/32	1/32
y 2		1/8	1/16	1/32	1/32
3		1/16	1/16	1/16	1/16
4		1/4	0	0	0

Represent this joint distribution as a 4×4 matrix in Matlab, and write Matlab expressions to answer the following questions:

1. What are the marginal probability distributions for X and Y ?
2. What is the joint entropy $H(X, Y)$?
3. What are the entropies $H(X)$ and $H(Y)$?
4. For each value of y , what is the conditional entropy $H(X|y)$?
5. What is the conditional entropy $H(X|Y)$?
6. What is the conditional entropy of Y given X ?
7. What is the mutual information between X and Y ?

Problem 4

In class we talked about the relationship between the entropy of a distribution and the average code-length per symbol in an optimal code. In the problem, you will explore this connection experimentally.

Part A: Find an optimal binary prefix code for the following distribution over an 8 symbol alphabet, where the code-length for each symbol is equal to its contribution to the entropy.

symbol	p	$-p \log_2 p$
a	.25	2
b	.25	2
c	.0625	4
d	.0625	4
e	.0625	4
f	.0625	4
g	.125	3
h	.125	3

What is the expected value for the average code-length per symbol under this code?

Part B: Now, exhibit a (non-optimal) code for the same 8 symbol alphabet, where the code-lengths are as follows:

symbol	p	code-length
a	.25	2
b	.25	2
c	.0625	3
d	.0625	4
e	.0625	5
f	.0625	6
g	.125	7
h	.125	2

What is the expected value for the average code-length per symbol under this code?

Part C: Finally, write a short Matlab program that does the following:

1. Generates a random string over this alphabet respecting the probability distributions given above for each string length in the range of 1 to 1000.
2. For each string, determine the average code-length per symbol.
3. Plot the results as a graph of string length vs. average code-length per symbol.

Use your program to compare the average code-length per symbol for the codes from parts A and B as string length increases. How does these values compare to the theoretical values you computed above?