HMM for Bioinformatics 2.

Definitions

Definition 1 *Let us define the entropy of the finite valued random variable* $X \sim p$ *as*

$$H(p) = H(x) = -\sum_{i=1}^{L} p(x_i) log_2(p(x_i)),\ H(0) = 0.$$

Definition 2 *Suppose* X *and* Y *are random variables assuming values in finite sets. Then the quantity*

$$H(X|Y) = \sum_{i=1}^{L} H(X|Y = y_i)p(y_i)$$

is called the conditional entropy of the random variable X *with respect to the random variable* X*.*

Definition 3 *The typical set* $A_\epsilon^n(X)$ *is a set of sequences* $x_1 x_2 \dots x_n$ *such that*

$$e^{-nH(X)+\epsilon} \geq p_{X_1,X_2,\dots,X_n}(x_1 x_2 \dots x_n) \geq e^{-nH(X)-\epsilon}$$

1. Prove that:

a) $H(p) \geq 0$ for all probability distribution p;

b) $H(p) = 0$ if and only if p is degenerate distribution;

c) $H(p) \leq log_2 L$ with equality if and only if p is the uniform distribution.

d) $H(X|Y) \leq H(X)$;

e) $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

2. Let X be a random variable having the probability distribution

$$p(x) = \frac{1}{2^x},\ x = 1, 2, \dots$$

Compute the entropy $H(X)$.

3. Select an integer $X \in U\{1, 2, \dots, n\}$ and an integer $Y|X = x \in U\{1, 2, \dots, n\}$. Compute the joint entropy $H(X,Y)$ without using the joint distribution of the random variable (X,Y).

4. Prove the following preposition:
If $X_1, X_2, \ldots, X_n$ are independent and identically distributed wit the distribution p, then

$$-\frac{1}{n} log_2 p_{X_1,X_2,\ldots,X_n}(X_1, X_2, \ldots, X_n) \to H(X)$$

in probability.

5. Verify three properties of the typical set:

a) if $(x_1 x_2 \ldots x_n) \in A_\epsilon^n$, then

$$P(X_1 = x_1, \ldots, X_n = x_n) \propto e^{-nH(X)\mp\epsilon}$$

b) $P(A_\epsilon^n) > 1 - \epsilon$ for n sufficiently large;

1. $|A_\epsilon^n| \geq e^{-nH(X)+\epsilon}$.

6. Write an algorithm to compute the entropy of a DNA sequence. Implement the algorithm in your favorite programming language. Obtain a DNA sequence and use your program to compute the entropy of the sequence. Report the result.

7. Prove that the Kullback-Leibler divergence is always nonnegative.

8. Prove that the cross entropy is minimal for identical distributions.

9. Prove that the mutual information $M(X, Y)$ is zero whenever X and Y are independent.

10. Prove that the mutual information $M(X, Y)$ equals the entropy H_X when X and Y are perfectly dependent (i.e. when they always evaluate to the same value).

11. Explain the practical differences between relative entropy and mutual information. Give an example where one is applicable and the other is not, and vice versa.