HMM for Bioinformatics

Paweł Błażej
Department of Genomics, Faculty of Biotechnology,
blazej@smorfland.uni.wroc.pl

6 marca 2019

Conventions and notations

1. $\mathcal{X}, \mathcal{Y}$ – alphabet;

Conventions and notations

1. $\mathcal{X}, \mathcal{Y}$ – alphabet;
2. $|\mathcal{X}| = |\mathcal{Y}| = L$ – size of an alphabet;

Conventions and notations

1. $\mathcal{X}, \mathcal{Y}$ – alphabet;
2. $|\mathcal{X}| = |\mathcal{Y}| = L$ – size of an alphabet;
3. X, Y discrete random variables ($X : \Omega \rightarrow \mathcal{X}$);

Conventions and notations

1. $\mathcal{X}, \mathcal{Y}$ – alphabet;
2. $|\mathcal{X}| = |\mathcal{Y}| = L$ – size of an alphabet;
3. X, Y discrete random variables ($X : \Omega \rightarrow \mathcal{X}$);
4. $P_X(x_i)$ probability.

Entropy

We can think of entropy as the level of uncertainty associated with a random variable.

Entropy – Definition

We define the entropy of the random variable $X \sim P$ as:

$$H(P) = H(X) = -\sum_{i=1}^{L} P(x_i) log_2(P(x_i))$$

where $H(0) = 0$ (convention $0 \cdot log_2 0 = 0$).

Joint entropy

$$H(X,Y) = -\sum_x \sum_y P(x,y) log P(x,y)$$

conditional entropy

$$H(X|Y) = -\sum_x \sum_y P(x,y) log P(x|y).$$

Properties of the entropy function

1. $H(X) \geqslant 0$

Properties of the entropy function

1. $H(X) \geqslant 0$
2. $H(X) \leqslant log(L)$ with equality iff X is uniformly distributed i.e. $P(x_i) = 1/L$ for every i;

Properties of the entropy function

1. $H(X) \geqslant 0$
2. $H(X) \leqslant log(L)$ with equality iff X is uniformly distributed i.e. $P(x_i) = 1/L$ for every i;
3. $H(X, Y) \leqslant H(X) + H(Y)$ with equality iff X and Y are independent;

Properties of the entropy function

1. $H(X) \geqslant 0$
2. $H(X) \leqslant log(L)$ with equality iff X is uniformly distributed i.e. $P(x_i) = 1/L$ for every i;
3. $H(X, Y) \leqslant H(X) + H(Y)$ with equality iff X and Y are independent;
4. $H(X)$ is concave in X .

Chain Rule for entropy

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

mutual information

$$M(X, Y) = \sum_x \sum_y P(x, y) log \frac{P(x, y)}{P(x)P(y)}$$

mutual information

$$M(X, Y) = \sum_x \sum_y P(x, y) log \frac{P(x, y)}{P(x)P(y)}$$

Remark

$$M(X, Y) = H(H) + H(Y) - H(X, Y).$$

Example – Binary entropy function

$$\mathcal{X} = \{x_1, x_2\},\ P(x_1) = p,\ \ P(x_2) = 1 - p$$

$$H(p) = -plog_2 p - (1-p)log_2(1-p)$$

Example – Entropy of random DNA

1. $\mathcal{X} = \{A, T, G, C\}$;

Example – Entropy of random DNA

1. $\mathcal{X} = \{A, T, G, C\}$;
2. $p(X = A) = p(X = T) = \ldots = p(X = C) = 0.25$;

Example – Entropy of random DNA

1. $\mathcal{X} = \{A, T, G, C\}$;
2. $p(X = A) = p(X = T) = \ldots = p(X = C) = 0.25$;
3. $H(X) = (\frac{1}{4} log \frac{1}{4} + \ldots + \frac{1}{4} log \frac{1}{4}) = 2$;

Example – Entropy of random DNA

1. $\mathcal{X} = \{A, T, G, C\}$;
2. $p(X = A) = p(X = T) = \ldots = p(X = C) = 0.25$;
3. $H(X) = (\frac{1}{4} log \frac{1}{4} + \ldots + \frac{1}{4} log \frac{1}{4}) = 2$;
4. We can think in this case of the entropy as a number of yes/no questions needed to identify an autcome.

Definition

A code is called uniquely decodable if any string composed of finite number of symbols from $\mathcal{X}$ gets a unique codestring.

Entropy – Interpretation

Definition

A code is called uniquely decodable if any string composed of finite number of symbols from $\mathcal{X}$ gets a unique codestring.

Kraft inequality

A necessary and sufficient condition on the lengths of codestring l_i in a uniquely decodable code is:

$$\sum_{i=1}^{L} 2^{-l_i} \leqslant 1.$$

Entropy – Interpretation

The expected length of the optimal uniquely decodable code belongs to the interval $[H(X), H(X) + 1)$

relative entropy – Kullback distance

$$D(P_1|P_2) = \sum_{i=1}^{L} P_1(x_i) log \frac{P_1(x_i)}{P_2(x_i)}$$

$$D(P_1|P_2) = \sum_{i=1}^{L} P_1(x_i) log \frac{P_1(x_i)}{P_2(x_i)}$$

Convention

1. $0 \cdot log \frac{0}{P_2(x_i)} = 0;$
2. $P_2(x_i) log \frac{P_1(x_i)}{0} = \infty$

If X is a random variable with the probability P_1 on an alphabet of L symbols, and P_2 is the uniform distribution on this alphabet then

$$D(P_1|P_2) = log(L) - H(X).$$

Properties of relative entropy

Further reading

1. T. Inglot, „Teoria informacji a statystyka matematyczne" wykład wygłoszony na XXXVIII Konferencji Statystyka Matematyczna Wisła 2012;
2. L. Dębowski, "Information Theory and Statistics", Institute of Computer Science Polish Academy of Science, Warsaw, 2013.