Relative entropy and clustering

CMSC423 Fall 2015

September 29, 2015

Question 1. Consider profile P:

$$\begin{array}{ll} \text{A}: & 0.0 \quad 0.0 \quad 0.0 \\ \text{C}: & 0.5 \quad 0.5 \quad 0.5 \\ \text{G}: & 0.2 \quad 0.3 \quad 0.5 \\ \text{T}: & 0.3 \quad 0.2 \quad 0.0 \end{array}$$

and background frequencies $b_A = 0.3$, $b_C = 0.3$, $b_G = 0.2$ and $b_T = 0.2$.

(a) What is the *entropy* of profile P?

(b) What is the *relative entropy* of profile P with respect to the background frequencies?

Question 2. In the soft k-means (EM) clustering algorithm:

(a) What are the parameters we want to estimate?

(b) What does HiddenMatrix$_{ij}$ correspond to in this algorithm? E.g., it corresponds to the probability that *fill-in-the-blank* j is generated from *fill-in-the-blank* i.

(c) Given HiddenMatrix$_{ij}$ and data points Data, how is the i-th center calculated on the algorithm's M-step?