Homework Assignment #1

Xueqiang Wang
Machine Learning in Bioinformatics

January 22, 2018

Questions 1)

(a).

$$\begin{aligned}\theta_{MLE} &= argmax_\theta P(S|\theta) \\ &= argmax_\theta \prod_i P(S_i|\theta) \\ &= argmax_\theta \sum_i log(P(S_i|\theta)) \\ &= argmax_\theta(2log(P(S_A|\theta)) + log(P(S_C|\theta)) + log(P(S_G|\theta)) + 6log(P(S_T|\theta))) \\ &= argmax_\theta(2log\theta_A + log\theta_C + log\theta_G + 6log\theta_T)\end{aligned}$$

$$\begin{aligned}\frac{d\theta_{MLE}}{d\theta_A} &= \frac{2}{\theta_A} - \frac{6}{1-\theta_A-\theta_C-\theta_G} = 0 \\ \frac{d\theta_{MLE}}{d\theta_C} &= \frac{1}{\theta_C} - \frac{6}{1-\theta_A-\theta_C-\theta_G} = 0 \\ \frac{d\theta_{MLE}}{d\theta_G} &= \frac{1}{\theta_G} - \frac{6}{1-\theta_A-\theta_C-\theta_G} = 0\end{aligned}$$

Therefore, $\theta_A = 0.2$, $\theta_C = \theta_G = 0.1$, $\theta_T = 0.6$

(b).

$$\begin{aligned}P(S_1|\theta) &= \theta_A^5\theta_T^4\theta_G = 0.2^5 * 0.6^4 * 0.1 \\ P(S_2|\theta) &= \theta_C^3\theta_G^5\theta_A\theta_T = 0.1^8 * 0.2 * 0.6 \\ P(S_1|\theta) &> P(S_2|\theta)\end{aligned}$$

Therefore, S_1 is more likely to be generated by this model.

Questions 1)

D: Disease, M: Mutation, $P(M) = 0.02, P(D|M) = 0.9, P(D|\bar{M}) = 0.05$

(a). $P(M|D) = \frac{P(M,D)}{P(D)} = \frac{P(M)P(D|M)}{P(M)P(D|M)+P(\bar{M})P(D|\bar{M})} = \frac{0.02*0.9}{0.02*0.9+0.98*0.05} = 0.269$

(b). $P(M|\bar{D}) = \frac{P(M,\bar{D})}{P(\bar{D})} = \frac{P(M)P(\bar{D}|M)}{P(M)P(\bar{D}|M)+P(\bar{M})P(\bar{D}|\bar{M})} = \frac{0.02*0.1}{0.02*0.1+0.98*0.95} = 0.0021$ sum of all samples in cluster C_j. The denominator and nominator combined represents the weighted centroid of a cluster.