6.047 Problem Set 2 Writeup

Matthew Feng

October 7, 2018

1 Naive Bayes

(a) Naive Bayes Assumption

No, the naive Baye's assumption that the features are independent conditioned on the class does not hold in this case. This is because the complexity of the sequence is likely dependent on the length of the sequence, as well as the GC content. If the GC content of a sequence is high, then the sequence is more likely to be a CpG island, which would mean that the complexity is lower because of the repeated base ordering.

(b) Computing MLEs

Maximum Likelihood Estimates

	X_1	X_2	X_3
$P(\cdot \mid repeat)$	0, short 1, long	0, low content 0, medium content 1, high content	2/3, low complexity 1/3, high complexity
$P(\cdot \mid gene)$	0, short 1, long	2/3, low content 1/3, medium content 0, high content	1/3, low complexity 2/3, high complexity
$P(\cdot \mid motif)$	1, short 0, long	0, low content 1/2, medium content 1/2, high content	1/4, low complexity 3/4, high complexity

Prior probability distribution

$$P(Y) = \begin{cases} 3/10, & Y = repeat \\ 3/10, & Y = gene \\ 4/10, & Y = motif \end{cases}$$

(c) Predicted MAP Class

The maximum a posteriori estimate of class of $(X_1, X_2, X_3) = (long, medium, low)$ is $Y = gene$; we don't need to compute the denominator of Baye's theorem for two reasons: the other classes have probability of 0 in the numerator, and that the MAP estimate of class is proportional only to the numerator; the denominator is just a normalization factor that is the same for all classes. Therefore, if we are only doing classification, we can just compare proportions rather than exact values.

$$\begin{aligned} P(Y|X_1, X_2, X_3) &= \frac{P(X_1, X_2, X_3|Y)P(Y)}{P(X_1, X_2, X_3)} && \text{(Baye's theorem)} \\ &= \frac{P(X_1|Y)P(X_2|Y)P(X_3|Y)P(Y)}{P(X_1, X_2, X_3)} && \text{(Naive Baye's assumption)} \end{aligned}$$

2 Classification of Conserved Regions

(a) Conditional probabilities

Alignment 1 $\log \mathbb{P}(S|N) = -17.098$, $\log \mathbb{P}(S|C) = -24.177$

Alignment 2 $\log \mathbb{P}(S|N) = -17.504$, $\log \mathbb{P}(S|C) = -13.256$

(b) Classification error I

The classification error, or amount of time that $P(S|C) > P(S|N)$ even though S is sampled from N, is 0.128 (12.8%).

(c) Classification error II

The classification error, or amount of time that $P(S|N) > P(S|C)$ even though S is sampled from C, is 0.1412 (14.1%).

(d) Reduce classification error

(i) Good discriminators

Score values 1 and 6 are good discriminators between the two models, because of the difference between the relative frequencies is high.

(ii) Bad discriminators

Score values 2 and 3 are poor discriminators between the two models, because of the difference between the relative frequencies is low.

The rate of classification errors would not decrease if we dismissed alignment scores of 0, because 0 is a good discriminator; alignments of 0 are twice as likely to occur in model N than in model C.

(e)

3 K-means Clustering

(a)

(b)

(c)

(d)