Cody Hubbard 004843389
CSM146, Winter 2018
Problem Set 5

Problem 1

(a) Solution:
When we represent documents using the aforementioned model we lose the actual order of the words and possible meanings of combinations of words in said orders.

(b) Solution:

$$\begin{aligned}
log[Pr(D_i, y_i)] &= log[Pr(D_i|y_i)Pr(y_i)] \\
&= log\theta + log(\frac{n!}{a_i + b_i + c_i}\alpha_1^{ai}\beta_1^{bi}\gamma_1^{ci}) \\
&= log(1-\theta) + \sum_{j=1}^{n} log(j) - \sum_{j=1}^{a_I} log(j) - \sum_{j=1}^{b_i} log(j) - \sum_{j=1}^{c_i} log(j) + a_i log\alpha_0 + b_i log\beta_0 + c_i log\gamma_0
\end{aligned}$$

(c) Solution:
We want to find the$argmax_{(\alpha_0,\beta_0,\gamma_0)}[log(1-\theta) + \sum_{j=1}^{n} log(j) - \sum_{j=1}^{a_I} log(j) - \sum_{j=1}^{b_i} log(j) - \sum_{j=1}^{c_i} log(j) + a_i log\alpha_0 + b_i log\beta_0 + c_i log\gamma_0]$
Which we can reduce to $argmax_{(\alpha_0,\beta_0,\gamma_0)}[a_i log\alpha_0 + b_i log\beta_0 + c_i log\gamma_0]$ due to independence.
We are given that $\alpha_0 + \beta_0 + \gamma_0 = 1$
So

$$\frac{\delta}{\delta\alpha_0}[a_i log\alpha_0 + b_i log\beta_0 + c_i log\gamma_0 - \lambda - \lambda\alpha_0 + \lambda\beta_0 + \lambda\gamma_0] = \frac{\alpha_i}{\alpha_0} - \lambda = 0$$

So its easy to see $\alpha_0 = \frac{\alpha_i}{\lambda}$

Because of our givens we can also infer $\frac{(a_i + b_i + c_i)}{\lambda} = 1$, where $(a_i + b_i + c_i) = n$ so $\frac{n}{\lambda} = 1$ and finally $\lambda = n$
so $\alpha_0 = \frac{a_i}{\lambda} = \frac{a_i}{n}$ and similarly$\alpha_1 = \frac{a_i}{n}$, $\beta_0 = \frac{b_i}{n}$,$\beta_1 = \frac{b_i}{n}$, $\gamma_0 = \frac{c_i}{n}$,$\gamma_1 = \frac{c_i}{n}$.

Problem 2

(a) Solution:

The missing transition probabilities are $q_{21} = P(q_{t+1} = 2|q_t = 1)$ and $q_{22} = P(q_{t+1} = 2|q_t = 2)$ both with value 0.

The missing output probabilities are $e_2(A) = P(O_t = A|q_t = 2)$ and $e_1(B) = P(O_t = B|q_t = 1)$.

$e_1(B) + e_1(A) = 1$ so $e_1(B) = 0.01$ and we know $e_2(A) + e_2(B) = 1$and $e_2(A) = 0.49$

(b) Solution:

$P(O_1 = A) = e_1(A)\pi_1 + e_2(A)\pi_2 = 0.735$

$P(O_1 = B) = e_1(B)\pi_1 + e_2(B)\pi_2 = 0.265$

So the most frequent output symbol to appear in the first position of sequences generated from this HMM is A.

(c) Solution:

Due to the transition probabilities being one and initial probability of A being the largest we can say $P(A|1)$ should max our joint prob for $P(O_1 : 3, q_1 : 3)$.

Lets looks at the probability for AAA

$P(AAA, 111) = (0.99)(0.99)(0.99)(0.49) \approx 0.475$

$P(AAA, 211) = (0.99)(0.99)(0.49)(0.51) \approx 0.245$

So $P(AAA) = 0.475 + 0.245 = .72$

If you were to calculate the other seven sequences probabilities you would fin that they are all less than $P(AAA)$ due to the emission, initial, and transition probabilities which favor A.

Thus AAA is the sequence of three output symbols that has the highest probability of being generated from this HMM model.

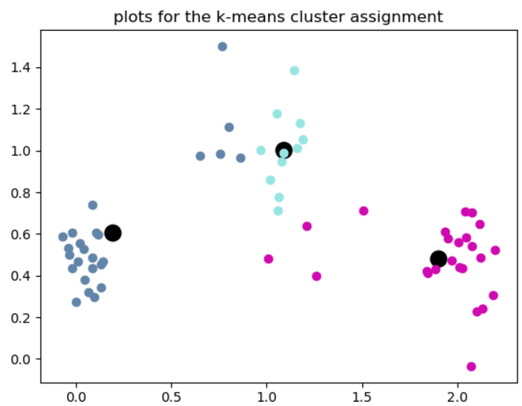
Problem 3

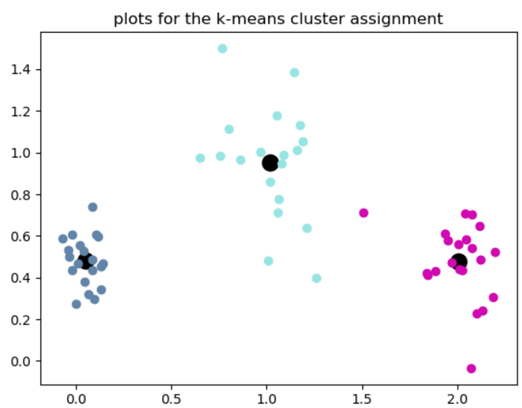
(a) Solution:
The minimum possible value of $J(c, \mu, k)$ is 0. This value is obtained when we have n clusters with $\mu_i = x_i$. Thus including the number of clusters in our minimization is equate a bad idea.

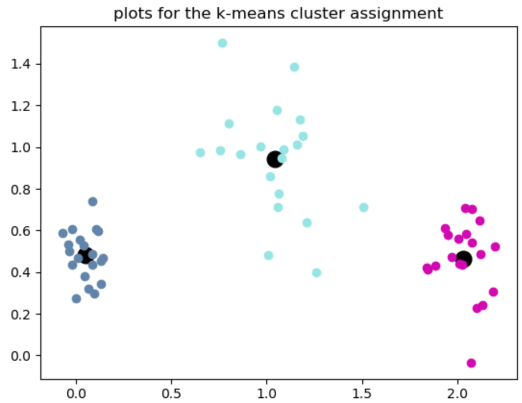
(b) Solution:
I Implemented all of the methods marked TODO in the Cluster and ClusterSet classes

(c) Solution:
I implement random_init(...) and kMeans(...) based on the provided specifications. sensible.

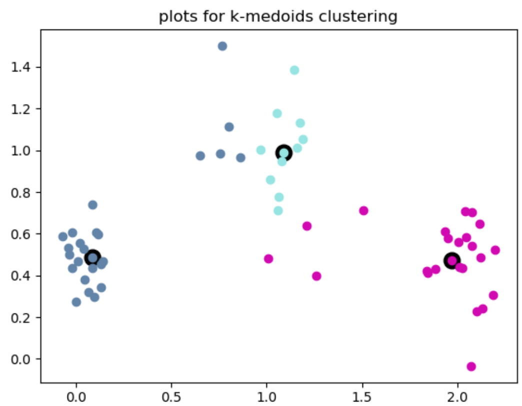
(d) Solution:



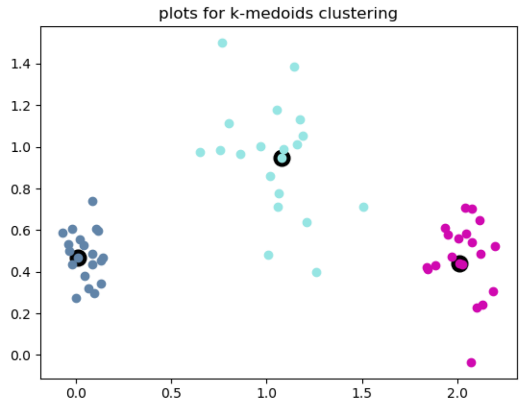



plots for the k-means cluster assignment
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0

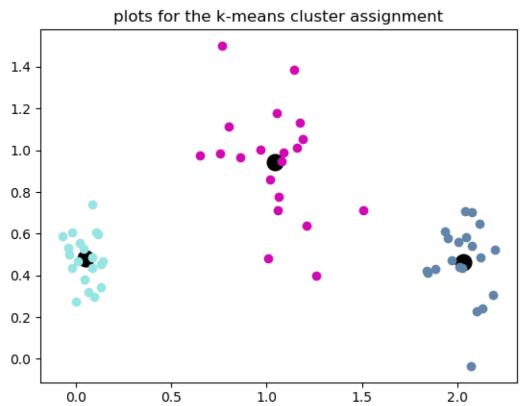
(e) Solution:

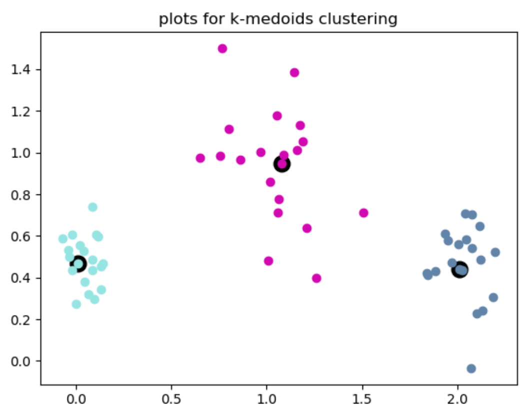

plots for k-medoids clustering
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0


plots for k-medoids clustering
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0


plots for k-medoids clustering

(f) Solution:


plots for the k-means cluster assignment


plots for k-medoids clustering