Home Work 4:

K-means and K-medoids:

Assume we have a 2D dataset consisting of (0, -6), (4, 4), (0, 0), (-5, 2). We wish to do K-means and K-medoids clustering with K = 2. We initialize the cluster centers with (-5, 2), (0, -6)

For this small dataset, in choosing between two equally valid exemplars for a cluster in K-medoids choose them with priority in the order given above (i.e. all other things being equal, you would choose (0, -6) as a center over (-5, 2)).

For the following scenarios, give the clusters and cluster centers after the algorithm converges. Enter the coordinate of each cluster center as a square-bracketed list (e.g. [0, 0]); enter each cluster's members in a similar format, separated by semicolons (e.g [1, 2]; [3, 4])

Clustering 1:

K-medoid algorithm with I_1 norm:

- First we will (arbitrarily) assign (-5, 2) to cluster 1, and (0, -6) to cluster 2
- Then we update the clusters to be [(4, 4), (-5, 2)] and [(0, -6), (0, 0)]
- At this point we have converged.

Clustering 2:

K-medoid algorithm with I_2 norm:

- First we will asign (-5, 2) to cluster 1, and (0, -6) to cluster 2
- Then, we update the clusters to be [(4, 4), (-5, 2), (0, 0)] and [(0, -6)]
- At this point, we will have converged.

Clustering 3:

K-means algorithm with I_1 norm

Note: For K-means algorithm with I_1 norm, you need to use median instead of mean when calculating the centroid.

- First we will assign (-5, 2) to cluster 1, and (0, -6) to cluster 2
- Then, we update the clusters to be [(4, 4), (-5, 2)] with center (-0.5, 3)
- We update [(0, 6), (0, 0)] with center (0, -3)
- At this point, we will have converged

Maximum likelihood estimation:

Consider a general multinomial distribution with parameters θ. Recall that the likelihood of a Dataset D is given by:

$$p(D;\theta) = \prod_{i=1}^{|N|} \theta_i^{c_i}$$

Where c_i is the occurence count of the i-th event.

The MLE of θ is the setting of θ that maximizes $p(D;\theta)$. In lecture we derived this to be:

$$\theta_i^* = \frac{c_i}{\sum_{j=1}^{|D|} c_j}$$

Unigram Model:

Consider the sequence: A B A B B C A B A A B C A C

A unigram model considers just one character at a time and calculates $p(w)$ for $w \in \{A, B, C\}$

What is the MLE estimate of θ?

We calculate the MLE as $\frac{\text{count}(w)}{N}$ where $N = 14$ and the counts are 6, 5 and 3

$$\theta_A^* = \frac{6}{14} = 0.428$$

$$\theta_B^* = \frac{5}{14} = 0.357$$

$$\theta_C^* = \frac{3}{14} = 0.214$$

Using the MLE estimate of θ on D, which of the following sequences is most likely?

ABC	BBB	ABB	AAC
$6 \times 5 \times 3$	5^3	6×5^2	$6^2 \times 3$

Bigram Model 1:

A bigram model computes the probability $p(D;\theta)$ as:

$$p(D;\theta) = p(w_0) \prod_{w_1, w_2 \in D} p(w_2 | w_1)$$

Where w_0 is the first word, and (w_1, w_2) is a pair of consecutive words in the document.

This is also a multinomial model. Assume the vocab size is N. How many parameters are there?

w_0 is the first word, and (w_1, w_2) is a pair of consecutive words in the document. Denote the set of all N words by V. The set of parameters is:

$$\{p(w_0) : w_0 \in V\} \cup \{p(w_1 | w_2) : w_1 \in V, w_2 \in V\}$$

The only constraints on these parameters are:

$$\sum_{w_0 \in V} p(w_0) = 1$$

$$\sum_{w_1 \in V} p(w_1 | w_2) = 1 \quad \text{for all } w_2 \in V$$

Hence the number of parameters is $(N-1) + (N^2 - N) = N^2 - 1$

Solution to the problem as written:

$$p(D;\theta) = \prod_{w_1 w_2 \in D} p(w_2 | w_1)$$

Without taking into account the likelihood $p(w_0)$ of the first word. In this case, the parameters are

$$\{ p(w_1 | w_2) : w_1 \in V, w_2 \in V \}$$

Where $\sum_{w_1 \in V} p(w_1 | w_2) = 1$ for all $w_2 \in V$. Hence, the number of parameters is $N^2 - N$

MLE for the conditional probability $p(w_2 | w_1)$:

$$p(w_2 | w_1) = \frac{p(w_1, w_2)}{p(w_1)}$$

To compute $p(w_1)$, we marginalize out w_2

$$\frac{\text{count}(w_1, w_2)}{\sum_{w_1, w_2' \in D} \text{count}(w_1, w_2')}$$

EM Algorithm:

Consider the following mixture of two Gaussians:

$$p(x;\theta) = \pi_1 N(x;\mu_1,\sigma_1^2) + \pi_2 N(x;\mu_2,\sigma_2^2)$$

This mixture has parameters $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$ They correspond to the mixing proportions, means, and variances of each Gaussian. We initialize θ as $\theta_0 = \{0.5, 0.5, 6, 7, 1, 4\}$

We have a dataset D with the following samples of x: $x^{(0)} = -1$, $x^{(1)} = 0$, $x^{(2)} = 4$, $x^{(3)} = 5$, $x^{(4)} = 6$

We want to set our parameters θ such that the data log-likelihood $\ell(D;\theta)$ is maximized:

$$\underset{\theta}{\arg\max} \sum_{i=0}^{4} \log p(x^{(i)};\theta)$$

Recall that we can do this with EM algorithm. The algorithm optimizes a lower bound on the log-likelihood, thus iteratively pushing the data likelihood upwards. The iterative algorithm is specified by two steps applied successively:

1. E Step: Infer component assignments from current $\theta_0 = \theta$ (complete the data)

$$p(y=k \mid x^{(i)}) := p(y=k \mid x^{(i)};\theta_0) \text{ for } k = 1,2 \text{ and } i = 0,\dots,4$$

2. M Step: maximize the expected log-likelihood

$$\hat{\ell}(D;\theta) = \sum_i \sum_k p(y=k \mid x^{(i)}) \log \frac{P(x^{(i)}, y=k;\theta)}{p(y=k \mid x^{(i)})}$$ with respect to θ while keeping fixed

To see why this optimizes a lower bound, consider the following inequality:

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_y p(x,y;\theta) \\
&= \log \sum_y q(y|x) \frac{p(x,y;\theta)}{q(y|x)} \\
&= \log E_{y\sim q(y|x)}\left[\frac{p(x,y;\theta)}{q(y|x)}\right] \\
&\geq E_{y\sim q(y|x)}\left[\log \frac{p(x,y;\theta)}{q(y|x)}\right] \\
&= \sum_y q(y|x) \log \frac{p(x,y;\theta)}{q(y|x)}
\end{aligned}$$

where the inequality comes from Jensen's inequality. EM makes this bound tight for the current setting $q(y|x)$ to be $p(y|x;\theta_0)$.

Likelihood Function:

What is the log-likelihood of the Data $\ell(D;\theta)$ given the initial setting of θ?

The likelihood can be written as:

$$p(D;\theta) = \prod_{i=0}^{4} p(x;\theta)$$

$$= \prod_{i=0}^{4} \pi_1 N(x^{(i)}; \mu_1, \sigma_1^2) + \pi_2 N(x^{(i)}; \mu_2, \sigma_2^2)$$

Taking the log gives:

$$\ell(D;\theta) = \sum_{i=0}^{4} \log\left(\pi_1 N(x^{(i)}; \mu_1, \sigma_1^2) + \pi_2 N(x^{(i)}; \mu_2, \sigma_2^2)\right)$$

We then evaluate each Gaussian using the standard formulation:

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Answer: $-24,5$

E-Step:

What is the formula for $p(y=k \mid x,\theta)$? Write in terms of π_k, π_1, π_2, N_k, N_1 and N_2

(Where $N_k = N(x \mid \mu_k, \sigma_k^2)$)

Following Bayes Rule we have:

$$p(y \mid x) = \frac{p(y)\,p(x \mid y)}{\sum_{y'} p(y')\,p(x \mid y')}$$

For this problem, this equates to:

$$p(y=k \mid x;\theta) = \frac{\pi_k N(x;\mu_y,\sigma_y^2)}{\sum_{i=1}^{2} \pi_i N(x;\mu_i,\sigma_i^2)}$$

$$= \frac{\pi_k N_k}{\pi_1 N_1 + \pi_2 N_2}$$

E-Step Weights:

For each of the given data points say which Gaussian (1 or 2) they are given more weight towards in the first E-step using the given setting of θ_0. This is answer 2 if $p(y=2|x,\theta_0) > p(y=1|x,\theta_0)$ and 1 otherwise.

Note that x will more likely be assigned to Gaussian 2 ($y=2$) instead of Gaussian 1 ($y=1$) when the following is true:

$$\frac{P(y=2|x^{(i)};\theta_0)}{P(y=1|x^{(i)};\theta_0)} > 1$$

$$\frac{P(x^{(i)}|y=2)\,P(y=2)}{P(x^{(i)}|y=1)\,P(y=1)} > 1$$

$$\frac{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\{-\frac{1}{2}(x-\mu_2)^2/\sigma_2^2\}}{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\{-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2\}} > 1$$

$$\frac{\frac{1}{\sqrt{2\pi\cdot 4}} \exp\{-\frac{1}{2}(x-7)^2/4\}}{\frac{1}{\sqrt{2\pi\cdot 1}} \exp\{-\frac{1}{2}(x-6)^2\}} > 1$$

$$\frac{1}{2}\exp\{-\frac{1}{2}((x-7)^2/4-(x-6)^2)\} > 1$$

$$\frac{1}{2}\exp\{\frac{1}{8}(x-5)(3x-19)\} > 1$$

$$\log(\tfrac{1}{2}) + \frac{1}{8}(x-5)(3x-19) > 0$$

The x-intercept of this parabola are $x_1 \approx 4.1525$, $x_2 \approx 7.1809$, Thus we can see that all points $x \in [4.15, 7.18]$ have higher probability under class $y=1$, and all other points have higher probability under $y=2$. Thus $x^{(0)}$, $x^{(1)}$ and $x^{(2)}$ are more likely (but not entirely) assigned to Gaussian 2, and The rest of the points ($x^{(3)}, x^{(4)}$) are more likely (but not entirely) assigned to Gaussian 1.

M-Step:

Fixing $p(y=k|x,\theta_0)$, we want to update θ such that our lower bound is maximized.

What is the optimal $\hat{\mu}_k$? For simplicity, assume we only have two data points $x^{(1)}$ and $x^{(2)}$ for this particular question. Answer in terms of $x^{(1)}, x^{(2)}$ and $\gamma_{k_1}, \gamma_{k_2}$ which are defined to be

$$\gamma_{ki} = p(y=k|x^{(i)};\theta_0)$$

The function we are optimizing is now:

$$\sum_i \sum_k \gamma_{ki} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right)$$

Taking $\frac{\partial}{\partial \mu_k}$ and setting to 0 gives:

$$\frac{\partial}{\partial \mu_k} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right) = \sum_i \gamma_{ki} \frac{\partial}{\partial \mu_k} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right)$$

$$= \sum_i \gamma_{ki} \frac{\partial}{\partial \mu_k} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} - \frac{(x^{(i)}-\mu_k)^2}{2\sigma_k^2} \right)\right)$$

$$= \sum_i \gamma_{ki} \frac{x^{(i)}-\mu_k}{\sigma_k^2} = 0$$

Separating out μ_k gives:

$$\mu_k = \frac{\sum_i \gamma_{ki} x^{(i)}}{\sum_i \gamma_{ki}}$$

We can interpret this as a weighted average of the data points, normalized by the "total mass" assigned to Gaussian k. The weight is the probability that point $x^{(i)}$ "belongs" to Gaussian k.

What is the optimal $\hat{\sigma}_k^2$?

Taking $\frac{\partial}{\partial \sigma_k^2}$ and setting to 0 gives:

$$\frac{\partial}{\partial \sigma_k^2} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right) = \sum_i \gamma_{ki} \frac{\partial}{\partial \sigma_k^2} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right)$$

$$= \sum_i \gamma_{ki} \frac{\partial}{\partial \sigma_k^2} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \right) - \frac{(x^{(i)}-\mu_k)^2}{2\sigma_k^2} \right)$$

$$= \sum_i \gamma_{ki} \left(-\frac{1}{2\sigma_k^2} + \frac{(x^{(i)}-\mu_k)^2}{2\sigma_k^4} \right) = 0$$

Separating out σ_k^2 gives:

$$\sigma_k^2 = \frac{\sum_i \gamma_{ki} (x^{(i)}-\mu_k)^2}{\sum_i \gamma_{ki}}$$

What is the optimal $\hat{\pi}_k$?

Finally we solve for π_k while including a lagrange multiplier for the constraint: $\sum_k \pi_k = 1$

$$\frac{\partial}{\partial \pi_k} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right) + \lambda\left(\sum_k \pi_k - 1\right) = \sum_i \gamma_{ki} \frac{\partial}{\partial \pi_k} \log(\pi_k) + \frac{\partial}{\partial \pi_k} \lambda\left(\sum_k \pi_k - 1\right)$$

$$= \frac{\sum_i \gamma_{ki}}{\pi_k} + \lambda = 0 \implies \pi_k = -\frac{\sum_i \gamma_{ki}}{\lambda}$$

Solving for λ gives:

$$\frac{\partial}{\partial \lambda} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k N(x^{(i)}; \mu_k, \sigma_k^2)\right) + \lambda\left(\sum_k \pi_k - 1\right) = \sum_k \pi_k - 1 = 0$$

Combining the two gives:

$$\lambda = -\sum_i \sum_k \gamma_{ki}$$ which we recognize as N. Thus $$\hat{\pi}_k = \frac{\sum \gamma_{ki}}{N}$$