

CARNEGIE MELLON UNIVERSITY 07-380

HOMEWORK 4

<https://www.cs.cmu.edu/~07380>

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

You may also digitally annotate the pdf. Illegible handwriting will lead to lost points. However, we suggest that you try to do at least some of your work directly in LaTeX.

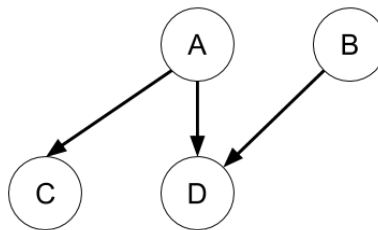
- **How to submit written component:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.
- **Policy:** See the course website for homework policies, including late policy, and academic integrity policies.

Name	
Andrew ID	
Hours to complete all components (nearest hour)	

1 [6 pts] Irrelevant Variables

A variable in a Bayes net is said to be *irrelevant* to a given query if we could remove the variable from the Bayes net without changing the answer to the query. For example, in the Bayes net below, the variable D is irrelevant to the query $P(B)$, because we already have the $P(B)$ table. On the other hand, B is *not* irrelevant to the query $P(D)$, because $P(B)$ is needed to compute $P(D)$.

Consider the Bayes net below:



1. [3 pts] Suppose we are making the query $P(A \mid D = d)$. Prove that C is irrelevant to this query using the following steps.
 - (i) Write the full joint distribution $P(A, B, C, d)$ as a product of the CPTs in the Bayes net.

Your Answer

- (ii) Sum over this product and normalize to get $P(A \mid d)$. Make sure you write out the full equations.

Your Answer

(iii) Explain why this expression does not depend on the variable C .

Your Answer

2. [3 pts] Now suppose we are making the query $P(C \mid D = d)$. Carry out steps (i) and (ii) from part 1 for this query, and then argue why B is *not* irrelevant.

Your Answer

2 [6 pts] Bayes Nets in the Courtroom

Read this article about using a Bayes net to assess new evidence in a criminal appeal case:

Smit, Lagnado, Morgan and Fenton, *Using Bayesian networks to guide the assessment of new evidence in an appeal case*, Crime Science 5:9 (2016).

<https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-016-0057-6>

Figure 3 of the paper presents a diagram of the Bayes net along with the corresponding conditional probability tables used by the authors to analyze the case. (The “Methods” section walks through how each table was filled in.)

1. [2 pts] Read the values off the paper’s Bayes net.

(i) $P(e_2 \mid a_{2p})$	(ii) $P(s \mid a_{1d}, a_{2d})$

2. [2 pts] How were the probabilities above determined? Limit your response to two sentences.

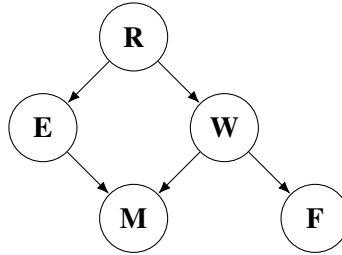
Your Answer

3. [2 pts] The numbers in this Bayes net were not measured from data; they were obtained by converting words such as “very likely” into probabilities. Give one argument *for* and one argument *against* presenting the conclusions of such a Bayes net to a court, and say what you would require before doing so.

Your Answer

3 [10 pts] Sampling

Consider the following Bayes net and corresponding probability tables.



R	$P(R)$
$+r$	0.4
$-r$	0.6

E	R	$P(E R)$
$+e$	$+r$	0.25
$-e$	$+r$	0.75
$+e$	$-r$	0.6
$-e$	$-r$	0.4

W	R	$P(W R)$
$+w$	$+r$	0.9
$-w$	$+r$	0.1
$+w$	$-r$	0.2
$-w$	$-r$	0.8

F	W	$P(F W)$
$+f$	$+w$	0.15
$-f$	$+w$	0.85
$+f$	$-w$	0.75
$-f$	$-w$	0.25

M	E	W	$P(M E, W)$
$+m$	$+e$	$+w$	0.1
$-m$	$+e$	$+w$	0.9
$+m$	$+e$	$-w$	0.45
$-m$	$+e$	$-w$	0.55
$+m$	$-e$	$+w$	0.35
$-m$	$-e$	$+w$	0.65
$+m$	$-e$	$-w$	0.9
$-m$	$-e$	$-w$	0.1

We are sampling to approximate the query $P(R | +f, +m)$. Variables are sampled in the topological order R, E, W, M, F .

1. [6 pts] Fill in the following table with the probability of *drawing* each respective sample when using each of the following sampling techniques. In each cell, show the product of CPT entries you used as well as the resulting number. Pay close attention to the domain of samples that can be drawn using each method. *Hint*: $P(+f, +m) = 0.2696$. *Hint 2*: probabilities can be 0.

Method	$\langle +r, +e, -w, +m, +f \rangle$	$\langle +r, -e, +w, -m, +f \rangle$
Prior sampling		
Rejection sampling		
Likelihood weighting		

2. **[2 pts]** What weight does likelihood weighting assign to the sample $\langle +r, +e, -w, +m, +f \rangle$? Show your work.

Your Answer

3. **[2 pts]** If we use rejection sampling for this query, what fraction of the generated samples do we expect to reject? Give a number and a one-sentence justification.

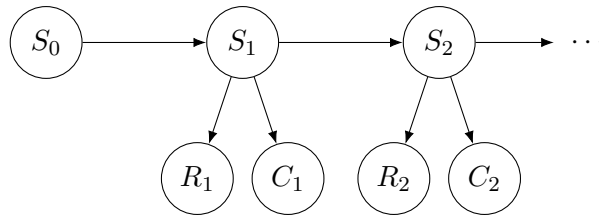
Your Answer

4 [10 pts] Hidden Markov Models: Is Everyone Getting Enough Sleep?

Professor Bayes the Bat wants to know whether the students in 07-380 are getting enough sleep. Each day, the professor observes whether the students sleep in class and whether they have red eyes. Let S_t be the random variable for the students having enough sleep, R_t be the random variable for the students having red eyes, and C_t be the random variable for the students sleeping in class on day t . The professor has the following theory:

- The prior probability of getting enough sleep, with no observations, is 0.6.
- The probability of getting enough sleep on night t is 0.9 given that the students got enough sleep the previous night, and 0.2 if not.
- The probability of having red eyes is 0.1 if the students got enough sleep, and 0.7 if not.
- The probability of sleeping in class is 0.2 if the students got enough sleep, and 0.4 if not.

This theory is the following dynamic Bayes net (an HMM with two evidence variables per time step):



S_0	$P(S_0)$	S_t	S_{t-1}	$P(S_t S_{t-1})$	R_t	S_t	$P(R_t S_t)$	C_t	S_t	$P(C_t S_t)$
$+s$	0.6	$+s_t$	$+s_{t-1}$	0.9	$+r$	$+s$	0.1	$+c$	$+s$	0.2
$-s$	0.4	$-s_t$	$+s_{t-1}$	0.1	$-r$	$+s$	0.9	$-c$	$+s$	0.8
		$+s_t$	$-s_{t-1}$	0.2	$+r$	$-s$	0.7	$+c$	$-s$	0.4
		$-s_t$	$-s_{t-1}$	0.8	$-r$	$-s$	0.3	$-c$	$-s$	0.6

Using this model and the evidence values

- $[-r_1, -c_1]$ = not red eyes, not sleeping in class,
- $[+r_2, -c_2]$ = red eyes, not sleeping in class,

we want to compute $P(S_t | r_{1:t}, c_{1:t})$ for each of $t = 1, 2$ (filtering). To do so, we compute intermediate values that correspond to the **predict** and **update** steps of the forward algorithm, as well as the normalization constant α in each case.

Round your answers to 3 decimal places at the end of each part, and use the rounded values from previous parts in later parts. We can only award partial credit if work is shown. To help you with this problem, you may check your numerical answers in the online HW. This is optional and will not be graded.

S_0	$P(S_0)$
$+s$	0.6
$-s$	0.4

S_t	S_{t-1}	$P(S_t S_{t-1})$
$+s_t$	$+s_{t-1}$	0.9
$-s_t$	$+s_{t-1}$	0.1
$+s_t$	$-s_{t-1}$	0.2
$-s_t$	$-s_{t-1}$	0.8

R_t	S_t	$P(R_t S_t)$
$+r$	$+s$	0.1
$-r$	$+s$	0.9
$+r$	$-s$	0.7
$-r$	$-s$	0.3

C_t	S_t	$P(C_t S_t)$
$+c$	$+s$	0.2
$-c$	$+s$	0.8
$+c$	$-s$	0.4
$-c$	$-s$	0.6

Evidence: $[-r_1, -c_1], [+r_2, -c_2]$. Round to 3 decimal places.

1. [5 pts] State estimation, $t = 1$.

Predict. Use the CPTs to write an expression for $P(+s_1)$, then compute both values.

Expression for $P(+s_1)$ in terms of the CPTs

$P(+s_1) =$

$P(-s_1) =$

Update. Use the CPTs and α to write an expression for $P(+s_1 | -r_1, -c_1)$, then compute both values (do **not** leave α in your numeric answers) and report the value of α .

Expression for $P(+s_1 | -r_1, -c_1)$

$P(+s_1 | -r_1, -c_1) =$

$P(-s_1 | -r_1, -c_1) =$

$\alpha =$

S_0	$P(S_0)$	S_t	S_{t-1}	$P(S_t S_{t-1})$	R_t	S_t	$P(R_t S_t)$	C_t	S_t	$P(C_t S_t)$
$+s$	0.6	$+s_t$	$+s_{t-1}$	0.9	$+r$	$+s$	0.1	$+c$	$+s$	0.2
$-s$	0.4	$-s_t$	$+s_{t-1}$	0.1	$-r$	$+s$	0.9	$-c$	$+s$	0.8
		$+s_t$	$-s_{t-1}$	0.2	$+r$	$-s$	0.7	$+c$	$-s$	0.4
		$-s_t$	$-s_{t-1}$	0.8	$-r$	$-s$	0.3	$-c$	$-s$	0.6

Evidence: $[-r_1, -c_1], [+r_2, -c_2]$. Round to 3 decimal places.

2. [5 pts] State estimation, $t = 2$.

Predict. Compute both values from the transition CPT and your filtered values for $t = 1$, showing the products you used.

$P(+s_2 -r_1, -c_1) =$

$P(-s_2 -r_1, -c_1) =$

Update. Use the CPTs, the probabilities from the predict step, and α to write an expression for $P(+s_2 | r_{1:2}, c_{1:2})$, then compute both values and α .

Expression for $P(+s_2 r_{1:2}, c_{1:2})$

$P(+s_2 r_{1:2}, c_{1:2}) =$

$P(-s_2 r_{1:2}, c_{1:2}) =$

$\alpha =$

5 [4 pts] K-Means: Non-increasing with K

Given a set of N observations $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and cluster centers $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$, assign each observation to its closest center. Each point has a corresponding cluster assignment $z_i \in \{1, \dots, K\}$ for the i -th point.

Consider the optimal K-means objective as a function of the hyperparameter for the number of clusters, K :

$$J(K) = \min_{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, z_1, \dots, z_N} \sum_{i=1}^N \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{z_i}\|_2^2$$

Prove that $J(K)$ is a non-increasing function of $K \in \mathbb{Z}^+$. Write your proof as a series of sentences. To receive full credit, each statement in your proof should have sound reasoning and the proof must be complete, e.g., cover all cases for K .

Hint: Try to consider two cases: 1) $K \geq N$, and 2) $K < N$.

Your Answer

6 [18 pts] Gaussian Mixture Models, EM, and Soft K-Means

This question is about the update rules of Gaussian mixture models (GMMs) fit with the expectation-maximization (EM) algorithm. The walkthrough below is a complete derivation of EM for a general GMM; there is nothing to submit for it. You will then redo the derivation for a simplified GMM called *soft K-means*, run one iteration of it by hand, and compare it with (hard) K-means.

Setup. The dataset is $X = \{\mathbf{x}^{(i)}\}_{i=1}^N$ with $\mathbf{x}^{(i)} \in \mathbb{R}^M$. A GMM with K components has parameters $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^K$: the mixture weights $\pi_k = p(Z_k^{(i)} = 1)$ with $\sum_{k=1}^K \pi_k = 1$, the component means $\mu_k \in \mathbb{R}^M$, and the (invertible) covariance matrices $\Sigma_k \in \mathbb{R}^{M \times M}$. The unobserved cluster assignments are one-hot vectors $Z^{(i)} \in \{0, 1\}^K$, where $Z_k^{(i)} = 1$ denotes the event that the i -th data point was generated by the k -th component. We write $f_{\mathcal{N}}$ for the density of a multivariate Gaussian,

$$f_{\mathcal{N}}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right),$$

so that $p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1, \theta) = f_{\mathcal{N}}(\mathbf{x}^{(i)} \mid \mu_k, \Sigma_k)$.

The EM algorithm alternates two steps, starting from some initial parameters $\theta^{(0)}$:

- E-step: (i) compute $E_{Z^{(i)}|\mathbf{x}^{(i)}, \theta^{(t)}}[Z_k^{(i)}] = p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)})$ for all i, k ;
(ii) formulate $E_{Z|X, \theta^{(t)}}[\ell_c(\theta \mid X, Z)]$ as a function of θ
- M-step: $\theta^{(t+1)} = \operatorname{argmax}_{\theta} E_{Z|X, \theta^{(t)}}[\ell_c(\theta \mid X, Z)]$

The first part of the E-step computes the probability of Z given $\mathbf{x}$ under the current parameters, i.e., for every point i and component k , the probability of point i belonging to cluster k . The second part uses these cluster probabilities to write the expectation of the *complete* log likelihood $\ell_c(\theta \mid X, Z) = \log p(X, Z \mid \theta)$ (the log likelihood we would maximize if the cluster assignments were observed) as a function of θ , which the M-step then maximizes.

Walkthrough: EM for a General GMM (nothing to submit)

Step 1: joint probability of a point and its assignment. By the generative story (first pick a component with probability π_k , then draw $\mathbf{x}$ from that Gaussian), the product rule gives

$$\begin{aligned} p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \theta) &= p(Z_k^{(i)} = 1 \mid \theta) p(\mathbf{x}^{(i)} \mid Z_k^{(i)} = 1, \theta) \\ &= \pi_k f_{\mathcal{N}}(\mathbf{x}^{(i)} \mid \mu_k, \Sigma_k) \\ &= \pi_k \frac{1}{\sqrt{(2\pi)^M |\Sigma_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)} - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x}^{(i)} - \mu_k)\right). \end{aligned}$$

Step 2: E-step, part (i): the cluster probabilities for each point. By Bayes' rule, and then marginalizing the denominator over the K possible components,

$$\begin{aligned} p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \theta^{(t)}) &= \frac{p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \theta^{(t)})}{p(\mathbf{x}^{(i)} \mid \theta^{(t)})} \\ &= \frac{p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \theta^{(t)})}{\sum_{j=1}^K p(\mathbf{x}^{(i)}, Z_j^{(i)} = 1 \mid \theta^{(t)})} \\ &= \frac{\pi_k^{(t)} f_{\mathcal{N}}(\mathbf{x}^{(i)} \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^K \pi_j^{(t)} f_{\mathcal{N}}(\mathbf{x}^{(i)} \mid \mu_j^{(t)}, \Sigma_j^{(t)})} \quad (\text{Step 1}) \end{aligned}$$

Step 3: E-step, part (ii): the expected complete log likelihood. From here on we abbreviate $r_{i,k}^{(t)} = p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)})$ and let $N_k^{(t)} = \sum_{i=1}^N r_{i,k}^{(t)}$, the effective number of points assigned to component k (so $\sum_{k=1}^K N_k^{(t)} = N$, since $\sum_{k=1}^K r_{i,k}^{(t)} = 1$ for every i). The expected complete log likelihood decomposes over the data points:

$$\begin{aligned}
 E_{Z|X, \boldsymbol{\theta}^{(t)}} [\ell_c(\boldsymbol{\theta} \mid X, Z)] &= E_{Z|X, \boldsymbol{\theta}^{(t)}} [\log p(X, Z \mid \boldsymbol{\theta})] \\
 &= E_{Z|X, \boldsymbol{\theta}^{(t)}} \left[\log \prod_{i=1}^N p(\mathbf{x}^{(i)}, Z^{(i)} \mid \boldsymbol{\theta}) \right] && \text{(data points are i.i.d.)} \\
 &= E_{Z|X, \boldsymbol{\theta}^{(t)}} \left[\sum_{i=1}^N \log p(\mathbf{x}^{(i)}, Z^{(i)} \mid \boldsymbol{\theta}) \right] \\
 &= \sum_{i=1}^N E_{Z^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}} [\log p(\mathbf{x}^{(i)}, Z^{(i)} \mid \boldsymbol{\theta})] && \text{(linearity of expectation)}
 \end{aligned}$$

Because $Z^{(i)}$ is one-hot, $p(\mathbf{x}^{(i)}, Z^{(i)} \mid \boldsymbol{\theta}) = \prod_{k=1}^K p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \boldsymbol{\theta})^{Z_k^{(i)}}$ (only the factor with $Z_k^{(i)} = 1$ differs from one), so

$$\log p(\mathbf{x}^{(i)}, Z^{(i)} \mid \boldsymbol{\theta}) = \sum_{k=1}^K Z_k^{(i)} \log p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \boldsymbol{\theta}).$$

Taking the expectation replaces each $Z_k^{(i)}$ by $E_{Z^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}} [Z_k^{(i)}] = p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}) = r_{i,k}^{(t)}$, and expanding the log of the joint probability from Step 1 then gives

$$\begin{aligned}
 E_{Z|X, \boldsymbol{\theta}^{(t)}} [\ell_c(\boldsymbol{\theta} \mid X, Z)] &= \sum_{i=1}^N \sum_{k=1}^K r_{i,k}^{(t)} \log p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \boldsymbol{\theta}) \\
 &= \sum_{i=1}^N \sum_{k=1}^K r_{i,k}^{(t)} \left(\log \pi_k - \frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k) \right).
 \end{aligned}$$

The M-step maximizes this over $\boldsymbol{\theta}$ with the $r_{i,k}^{(t)}$ held fixed.

Step 4: M-step for the means. Only the last term of the expected complete log likelihood depends on μ_k , and $\nabla_{\mu} \left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) \right) = \Sigma^{-1}(\mathbf{x} - \mu)$ for symmetric Σ , so

$$\nabla_{\mu_k} E_{Z|X, \theta^{(t)}} [\ell_c(\theta | X, Z)] = \sum_{i=1}^N r_{i,k}^{(t)} \Sigma_k^{-1}(\mathbf{x}^{(i)} - \mu_k).$$

Setting this to zero and solving for μ_k ,

$$\begin{aligned} \sum_{i=1}^N r_{i,k}^{(t)} \Sigma_k^{-1}(\mathbf{x}^{(i)} - \mu_k) &= \mathbf{0} \\ \sum_{i=1}^N r_{i,k}^{(t)}(\mathbf{x}^{(i)} - \mu_k) &= \mathbf{0} \quad (\text{multiply by } \Sigma_k, \text{ which is invertible}) \\ \sum_{i=1}^N r_{i,k}^{(t)} \mathbf{x}^{(i)} &= \left(\sum_{i=1}^N r_{i,k}^{(t)} \right) \mu_k = N_k^{(t)} \mu_k \\ \mu_k^{(t+1)} &= \frac{\sum_{i=1}^N r_{i,k}^{(t)} \mathbf{x}^{(i)}}{N_k^{(t)}}, \end{aligned}$$

a weighted average of the data with weights $r_{i,k}^{(t)}$.

Step 5: M-step for the covariances. The terms of the expected complete log likelihood that involve Σ_k are

$$\sum_{i=1}^N r_{i,k}^{(t)} \left(-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (\mathbf{x}^{(i)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(i)} - \mu_k) \right).$$

We use the matrix derivatives $\frac{d|A|}{dA} = |A|A^{-\top}$ (so $\frac{d \log |A|}{dA} = A^{-\top}$) and $\frac{d\mathbf{v}^\top A^{-1} \mathbf{v}}{dA} = -A^{-\top} \mathbf{v} \mathbf{v}^\top A^{-\top}$, together with $\Sigma_k^{-\top} = \Sigma_k^{-1}$ because Σ_k is symmetric. Setting the gradient with respect to Σ_k to zero,

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^N r_{i,k}^{(t)} \left(-\frac{1}{2} \Sigma_k^{-1} + \frac{1}{2} \Sigma_k^{-1} (\mathbf{x}^{(i)} - \mu_k) (\mathbf{x}^{(i)} - \mu_k)^\top \Sigma_k^{-1} \right) \\ &= \frac{1}{2} \Sigma_k^{-1} \left(\sum_{i=1}^N r_{i,k}^{(t)} \left((\mathbf{x}^{(i)} - \mu_k) (\mathbf{x}^{(i)} - \mu_k)^\top - \Sigma_k \right) \right) \Sigma_k^{-1}. \end{aligned}$$

Multiplying by Σ_k on the left and on the right, and then solving for Σ_k ,

$$\begin{aligned} \sum_{i=1}^N r_{i,k}^{(t)} \left((\mathbf{x}^{(i)} - \mu_k) (\mathbf{x}^{(i)} - \mu_k)^\top - \Sigma_k \right) &= \mathbf{0} \\ \sum_{i=1}^N r_{i,k}^{(t)} (\mathbf{x}^{(i)} - \mu_k) (\mathbf{x}^{(i)} - \mu_k)^\top &= \left(\sum_{i=1}^N r_{i,k}^{(t)} \right) \Sigma_k = N_k^{(t)} \Sigma_k \\ \Sigma_k &= \frac{\sum_{i=1}^N r_{i,k}^{(t)} (\mathbf{x}^{(i)} - \mu_k) (\mathbf{x}^{(i)} - \mu_k)^\top}{N_k^{(t)}}. \end{aligned}$$

Plugging in the updated mean from Step 4 gives the update

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^N r_{i,k}^{(t)} (\mathbf{x}^{(i)} - \mu_k^{(t+1)}) (\mathbf{x}^{(i)} - \mu_k^{(t+1)})^\top}{N_k^{(t)}}.$$

Step 6: M-step for the mixture weights. The weights must satisfy $\sum_{k=1}^K \pi_k = 1$, so we maximize the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = E_{Z|X, \boldsymbol{\theta}^{(t)}} [\ell_c(\boldsymbol{\theta} | X, Z)] - \lambda \left(\sum_{k=1}^K \pi_k - 1 \right).$$

Only the $\log \pi_k$ terms depend on π_k , so setting the derivative to zero and solving for π_k ,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_k} &= \frac{\sum_{i=1}^N r_{i,k}^{(t)}}{\pi_k} - \lambda = 0 \\ \pi_k &= \frac{\sum_{i=1}^N r_{i,k}^{(t)}}{\lambda} = \frac{N_k^{(t)}}{\lambda}. \end{aligned}$$

The multiplier λ follows from the constraint: summing the last line over k ,

$$\begin{aligned} \sum_{k=1}^K \pi_k &= \frac{1}{\lambda} \sum_{k=1}^K N_k^{(t)} \\ 1 &= \frac{N}{\lambda} && \text{(the constraint, and } \sum_k N_k^{(t)} = N) \\ \lambda &= N. \end{aligned}$$

Therefore

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{N}.$$

Summary. One EM iteration computes the cluster probabilities $r_{i,k}^{(t)}$ for each point from the current parameters (Step 2), forms the expected complete log likelihood from them (Step 3), and maximizes it (Steps 4–6); in the resulting updates every point i contributes to every component k in proportion to $r_{i,k}^{(t)}$.

6.1 [11 pts] Soft K-Means: EM for a Simplified GMM

Soft K-means is the special case of a GMM in which the mixture weights are fixed and uniform, $\pi_k = \frac{1}{K}$ for all k , and every covariance matrix is fixed to the identity, $\Sigma_k = I$ (the $M \times M$ identity matrix). These are no longer parameters: the only parameters are the means, $\theta = (\mu_1, \dots, \mu_K)$. Redo the derivation of the walkthrough for this model, simplifying as much as possible and showing your work.

1. **[2 pts]** Derive an expression for the joint probability $p(\mathbf{x}^{(i)}, Z_k^{(i)} = 1 \mid \theta^{(t)})$ for soft K-means (the analogue of Step 1) in terms of $\mu_k^{(t)}$, $\mathbf{x}^{(i)}$, K and M . *Hint:* $|I| = 1$ and $I^{-1} = I$.

Your Answer

2. [3 pts] **E-step, part (i).** Derive an expression for $p\left(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}\right)$ for soft K-means (the analogue of Step 2) in terms of $\boldsymbol{\mu}_j^{(t)}$ and $\mathbf{x}^{(i)}$ only. Show which constants cancel and why.

Your Answer

3. **[3 pts] E-step, part (ii).** In the following questions, use $r_{i,k}^{(t)}$ to represent $p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)})$ from the previous part. Express the expected complete log likelihood for soft K-means (the analogue of Step 3), $E_{Z|X, \boldsymbol{\theta}^{(t)}}[\ell_c(\boldsymbol{\theta} \mid X, Z)]$, in terms of $r_{i,k}^{(t)}$, $\boldsymbol{\mu}_k$, $\mathbf{x}^{(i)}$, N , K and M . Collect every term that does not depend on the means into a single constant C , and give the value of C .

Hint: Remember that $\sum_{k=1}^K p(Z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(t)}) = 1$ for every i .

Your Answer

4. [3 pts] **M-step.** Let $N_k^{(t)} = \sum_{i=1}^N r_{i,k}^{(t)}$. Show your work to derive

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^N r_{i,k}^{(t)} \mathbf{x}^{(i)}}{N_k^{(t)}}$$

by setting the gradient of your expression from the previous part with respect to $\boldsymbol{\mu}_k$ to zero. Then explain in one sentence why soft K-means has no M-step update for π_k or Σ_k , and which steps of the walkthrough are therefore not needed.

Your Answer

6.2 [4 pts] Soft K-Means by Hand

Consider soft K-means in one dimension ($M = 1$, so $\Sigma_k = I$ means $\sigma_k^2 = 1$) with $K = 2$ components. The current means are $\mu_1^{(t)} = 1$ and $\mu_2^{(t)} = 4$, and the dataset is

$$x^{(1)} = 0, \quad x^{(2)} = 2, \quad x^{(3)} = 3, \quad x^{(4)} = 5, \quad x^{(5)} = 6.$$

Report numeric answers to 3 decimal places.

1. **[2 pts] E-step.** Use the formula for $p\left(Z_k^{(i)} = 1 \mid x^{(i)}, \theta^{(t)}\right)$ you derived in the previous subsection to compute $r_{2,1}^{(t)}$ and $r_{3,1}^{(t)}$ (the probability of $x^{(2)} = 2$ and of $x^{(3)} = 3$ belonging to cluster 1). Show your work.

Your Answer

2. **[2 pts] M-step.** The table of $r_{i,k}^{(t)}$ values (rounded) is given below for the other three points; the two columns for $x^{(2)}$ and $x^{(3)}$ are the values you computed in the E-step (remember that $r_{i,2}^{(t)} = 1 - r_{i,1}^{(t)}$).

i	1 ($x = 0$)	2 ($x = 2$)	3 ($x = 3$)	4 ($x = 5$)	5 ($x = 6$)
$r_{i,1}^{(t)}$	0.999	(E-step)	(E-step)	0.001	0.000
$r_{i,2}^{(t)}$	0.001	(E-step)	(E-step)	0.999	1.000

Use the M-step update you derived above to compute $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$. Show your work.

$$\mu_1^{(t+1)} =$$

$$\mu_2^{(t+1)} =$$

6.3 [3 pts] Soft vs. Hard K-Means

1. [1 pts] Run one iteration of (hard) K-means on the same five points starting from the centers $\mu_1 = 1$ and $\mu_2 = 4$: assign each point to its closest center, then recompute the centers. Report the cluster assignments and the new centers to 3 decimal places, showing your work.

Your Answer

2. [1 pts] K-means assigned $x^{(2)} = 2$ entirely to cluster 1 and $x^{(3)} = 3$ entirely to cluster 2, whereas soft K-means gave them the cluster probabilities $r_{i,k}^{(t)}$ you computed above. In one or two sentences, explain why the soft K-means means $\mu_1^{(t+1)}$ and $\mu_2^{(t+1)}$ differ from the hard K-means centers, and why this is called *soft* clustering.

Your Answer

3. [1 pts] **True or false:** starting from any initialization, EM (for a GMM or for soft K-means) converges to the global maximum of the (marginal) log likelihood. Justify your answer in one sentence and state what is done in practice to mitigate the issue.

☐ True ☐ False

Your Answer

7 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

Your Answer

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

Your Answer

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.

Your Answer