Homework # 5

Reading (optional)

- Section 9.4 discusses the general theory of EM and covers the material I discussed in the lecture.
- Setion 12.1 discusses PCA.

1. In this problem, we will once again revisit the Hope heights problem of HW 3, but this time we will use the formal notation of EM. Recall, we consider the two component Gaussian mixture model,

$$X = \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) & \text{with probability } p_1 \\ \mathcal{N}(\mu_2, \sigma_2^2) & \text{with probability } p_2 \end{cases} \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution and X models the height of a person when gender is unknown. Let $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1, p_2)$. Let $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N$ be the sample heights given in the file.

   (a) Recall that to implement EM we define

$$Q(\theta, \theta') = \sum_{i=1}^{N} E_\theta[\log P(\hat{X}_i, z_i \mid \theta')], \tag{2}$$

   where z_i is either 1 or 2 and determines the mixture $\hat{X}_i$ was sampled from and the θ subscript in the expectation means that we take the expectation with the z_i distributed according to θ. Write down an expression for $Q(\theta, \theta')$ using $r_{i1} = P(z_i = 1 \mid \hat{X}_i, \theta)$, $r_{i2} = P(z_i = 2 \mid \hat{X}_i, \theta)$ and the pdfs of the normals. Then give a formula for r_{i1} and r_{i2} in terms of θ and the $\hat{X}_i$.

   (b) Compute $\text{argmax}_{\theta'} Q(\theta, \theta')$. You should derive an expression for each entry of θ' by solve $\nabla_{\theta'} Q(\theta, \theta') = 0$. Hint: To compute that values of p_1' and p_2' for θ', you can either use a Lagrange multiplier approach, with the constraint $p_1' + p_2' = 1$ or you can simply substitute $p_2' = 1 - p_1'$.

   (c) Compare your updates for the parameters to the heuristic updates of soft EM.

2. See the attached section 9.3.3 from Bishop for a definition and discussion of Bernoulli mixture models. Let X represent 10 bits, i.e. $X = (X_1, X_2, \ldots, X_{10})$ where each coordinate of X is either 0 or 1. Assume the following Bernoulli mixture model for the ith coordinate of X, X_i:

$$X_i = \begin{cases} \text{Bernoulli}(\mu_i^{(1)}) & \text{with probability } p_1 \\ \text{Bernoulli}(\mu_i^{(2)}) & \text{with probability } p_2, \end{cases} \tag{3}$$

where $\mu^{(1)}, \mu^{(2)} \in \mathbb{R}^{10}$ with all coordinates in $[0, 1]$. Assume further that the coordinates of X are always sampled from the same mixture, with probabilities p_1 and p_2 for mixture 1 and 2 respectively, but that the Bernoulli draw of each coordinate is independent.

   (a) Write down an EM iteration for this mixture model.

   (b) Attached is the file `noisy_bits.csv` which contains a 500×10 matrix. Each row of the matrix is a sample of X. If you look at an image of the matrix (in R use **image** on the transposed matrix), you will see that there are two patterns, but with some noise added. Use your EM algorithm to fit the mixture model to the data. Does your fit recover the two underlying patterns?

3. This problem focuses on the computations involved in deriving the PCA. Consider the dataset formed by $X^{(i)} \in \mathbb{R}^n$ for $i = 1, 2, \ldots, N$. Set $\mu = 1/N \sum_{i=1}^n X^{(i)}$.

   (a) Let $a, b \in \mathbb{R}^n$ be column vectors. Show in any way you like - by proof, through example, by intuitive explanation - that $(a \cdot b)^2 = a^T M a$ where M is an $n \times n$ matrix given by $M = bb^T$.

   (b) Let $\hat{\Sigma}$ be the covariance matrix of the data. Then, by definition

$$\hat{\Sigma}_{jk} = \frac{1}{N} \sum_{i=1}^{N} (X_j^{(i)} - \mu_j)(X_k^{(i)} - \mu_k) \tag{4}$$

   Show that $\hat{\Sigma}$ can also be written in the following two forms

- Let $\tilde{X}$ be the $N \times n$ matrix with the $X^{(i)} - \mu$ as the rows

$$\hat{\Sigma} = \frac{1}{N}\tilde{X}^T\tilde{X} \tag{5}$$

- Thinking of the $X^{(i)}$ as column vectors,

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X^{(i)} - \mu)(X^{(i)} - \mu)^T \tag{6}$$

(c) The 1-d PCA involves the parameters μ, $w^{(1)}$ and $c_i \in \mathbb{R}$ for $i = 1, 2, \ldots, N$ that are used to approximate $X^{(i)}$ according to

$$X^{(i)} \approx \mu + c_i w^{(1)}. \tag{7}$$

Derive the values of c_i and $w^{(1)}$ that optimize this approximation. (We did this in class.) Then, compute the mean and variance of the c_i.