CS 539 Machine Learning Homework 3

Conceptual and Theoretical Questions (6 questions, 40 pts)

The joint distribution over three binary variables.

a	b	c	$p(a,b,c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

1. Consider three binary variables a, b, c ∈ {0, 1} having the joint distribution given in the above table. Show by direct evaluation that this distribution has the property that a and b are marginally dependent, so that $p(a, b) \neq p(a)p(b)$, but that they become independent when conditioned on c, so that $p(a, b|c) = p(a|c)p(b|c)$ for both c = 0 and c = 1. (This is question 8.3, from Bishop textbook). **(6 pts)**

2. Evaluate the distributions p(a), p(b|c), and p(c|a) corresponding to the joint distribution given in the above table. Hence show by direct evaluation that $p(a, b, c) = p(a)p(c|a)p(b|c)$. Draw the corresponding directed graph. (This is question 8.4, from Bishop textbook). **(6 pts)**

3. Consider two discrete variables x and y each having three possible states, for example x, y ∈ {0, 1, 2}. Construct a joint distribution p(x, y) over these variables having the property that the value x that maximizes the marginal p(x), along with the value y that maximizes the marginal p(y), together have probability zero under the joint distribution, so that p(x,y) = 0. (This is question 8.27, from Bishop textbook) (**8 pts**)

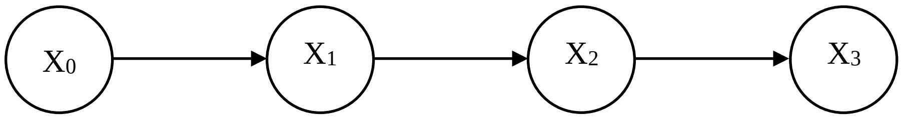
4. Suppose we wish to use the EM algorithm to maximize the posterior distribution over parameters p(θ|X) for a model containing latent variables, where X is the observed data set. Show that the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by $Q(\theta, \theta_{old}) + \ln p(\theta)$ where $Q(\theta, \theta_{old})$ is defined by (9.30). (This is question 9.4, from Bishop textbook) **(4 pts)**

5. Consider a special case of a Gaussian mixture model in which the covariance matrices Σ_k of the components are all constrained to have a common value Σ. Derive the EM equations for maximizing the likelihood function under such a model. (This is question 9.6, from Bishop textbook) (**8 pts**)

6. Consider a Bernoulli mixture model as discussed in Section 9.3.3, together with a prior distribution $p(\mu_k|a_k, b_k)$ over each of the parameter vectors μ_k given by the beta distribution (2.13), and a Dirichlet prior p(π|α) given by (2.38). Derive the EM algorithm for maximizing the posterior probability p(μ,π|X). (This is question 9.18, from Bishop textbook) **(8 pts)**

Application Questions (4 questions, 60 pts)

Graphical Model (15 points) Here, we want to guesstimate what a local weather station will report in 4 consecutive days, starting day 0. The station uses three words, "cold", "hot", and "mild". We use a directed graph to characterize the report over these 4 days, where variables X_0, X_1, X_2, and X_3 represent the station report on these 4 days.

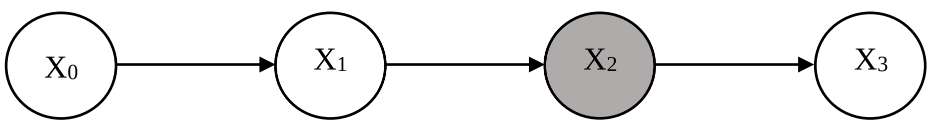


These variables will get three possible states, $X_0, X_1, X_2, X_3 \in$ {'cold', 'hot', 'mild'}. We call 'cold' state 1 (S_1), 'hot' state 2 (S_2), and 'mild' state 3 (S_3). Historically, we have learned that the station has a sort of pattern in its reporting of weather; thus, we might have extra information on what the report is for the next day given we know today's report. The conditional probability of $P(X_k|X_{k-1})$ is defined by the following matrix:

$$P(X_k|X_{k-1}) = \begin{bmatrix} 1/2 & 1/3 & 1/4 \\ 1/4 & 1/3 & 1/4 \\ 1/4 & 1/3 & 1/2 \end{bmatrix}$$

where, the matrix $(i,j)^{th}$ element defines probability $X_k = S_i$ given $X_{k-1} = S_j$. For example, the probability of X_k to be 'cold' given X_{k-1} is 'cold' will be 1/2, and the probability of X_k to be 'mild given X_{k-1} is 'hot' is 1/3. For this problem,

a) Draw 5 samples for the station report over these 4 days. For day zero (X_0), consider $P(X_0$='cold')= $P(X_0$='hot')= $P(X_0$='mild')=1/3.
b) We were not in the town to check the report for day zero (X_0); so, we assume $P(X_0$='cold')= $P(X_0$='hot')= $P(X_0$='mild')=1/3. Derive the marginal probabilities for days 0 to 3. In other words, what is the probability of the station to report 'cold', 'hot', and 'mild' per each day.
c) Someone tells us that they know for sure that the X_2 report will be "hot", what are the conditional probabilities for other days given the X_2 report is 'hot'.

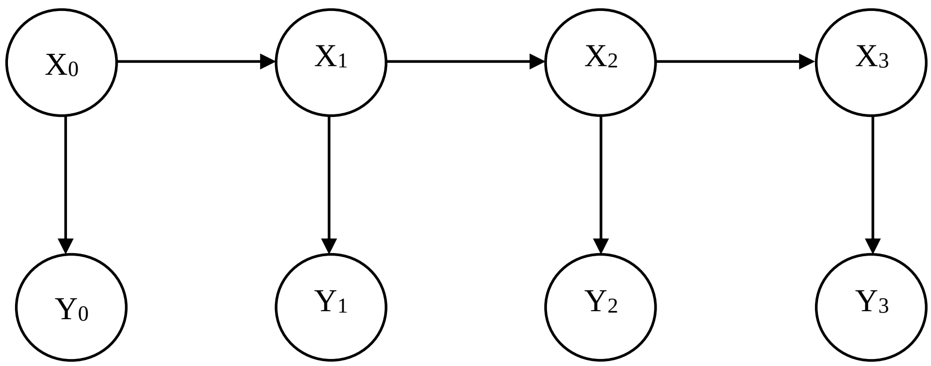


d) What is the most probable report for these four days?

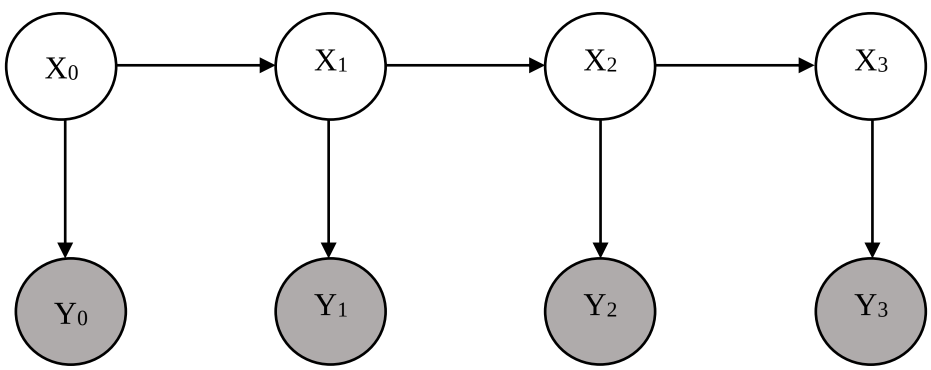
Graphical Model (15 points) The weather station is experimenting with a graphical presentation of the weather report instead of using "cold", "hot", and "mild" words. They have designed a bar display, where the height of the bar is a function of "state". They also considered adding a bit of variability to the bar hoping that the bar display will be more engaging. Their bar height model is defined by

$$p(y_i|s_i) \sim N(m_i, \sigma^2)$$
$$m_i = \begin{cases} -2 & s_i =' cold' \\ 0 & s_i = 'mild' \\ 2 & s_i =' hot' \end{cases}$$
$$\sigma^2 = 1$$

The below figure shows the graphical model for the bar and state.



a) Draw 5 samples for the bar height over these 4 days.
b) Draw 5 samples for the bar display when X_0='hot', X_1='mild', X_2='cold', X_3='cold'.
c) Let's assume, we observed Y_0=0.7, Y_1=1.5, Y_2=-1.8, Y_3=-1. What are marginal distribution of $P(X_k|$ $Y_0=0.7, Y_1=1.5, Y_2=-1.8, Y_3=-1)$ k=0,1,2,3.



Gaussian Mixture Model (15 points) We will use the gmm_data.xlsx dataset for this problem (GMM).

1. Visualize the dataset; and discuss what would you suggest for the number of clusters in the data.
2. Fit a 2-D normal distribution on the data and show your result over the scatter plot of your data points.
3. Fit GMM with K=2 to the data and show your result over the scatter plot of the data points. Repeat this with different initialization and discuss your results.
4. Repeat pat 2 with K=3.
5. For theoretical question 5, we derived the GMM solution with a common covariance matrix for mixtures. Run this solution on the dataset with K=2 and K=3, and show your result.

Poisson Mixture Model (15 points) We will use the poisson_data.xlsx dataset for this problem. Poisson distribution is a discrete distribution with a rate parameter (λ), where the probability of random variable x to be m is defined by

$$p(x = m) = \frac{\lambda^m e^{-\lambda}}{m!}$$

In many datasets, we may observe a mixture of Poisson distributions (I have attached a pdf file showing this form of data in DNA sequencing). We can build a Poisson mixture model similar to what we did for the Gaussian mixture model. The model is defined by

$$p(x = m) = \sum_{k=1}^{K} \pi_k \frac{\lambda_k^m e^{-\lambda_k}}{m!}$$

For the poisson_data.xlsx, we want to fit a Poisson mixture model with K=2.

a) Plot the histogram of the data and discuss your observation. You might compare the histogram with the histogram of Poisson distribution.
b) Fit a Poisson distribution to the data and compare its pmf with the histogram you derived in part (a).
c) Derive the update rule for the mixture model (EM).
d) Apply the EM algorithm to the data with K=2 and plot the model pmf with the histogram of the data.
e) Bonus point: discuss model evidence for the models in parts (b) and (d).