AI & ML Course
MidSem 3(Apr 15, 2024)

Time: 70 minutes

Instructions

- Answer all questions
- All answers must be written in the provided spaces. Answers written outside the boxes will not be not graded.
- Last five pages are for rough work. Will not be graded.

Name: _____ SRNO: _____

Room no: _____ Serial Number: _____

Question:	1	2	3	4	Total
Points:	10	15	10	5	40
Score:					

1. Consider learning a mixture of Bernoulli distribution of the following form.

$$P(X = x|\theta, \alpha) = \sum_{i=1}^{k} \alpha_i \theta_i^x (1-\theta_i)^{(1-x)}$$

$$\alpha_i > 0, \sum_{i=1}^{k} \alpha_i = 1, 0 < \theta_i < 1$$

From Dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ we wish to learn the parameters $\{\theta, \alpha\}$ through the EM algorithm. Let $\bar{\alpha}, \bar{\theta}$ be the current estimate of the parameters and the probabilities $q_i^{(j)} = \frac{\bar{\alpha}_i \bar{\theta}^{x_j)(1-\bar{\theta}_i)^{(1-x_j)}}}{P(X_j=x_j|\bar{\theta},\bar{\alpha})}$

(a) (3 points) The EM algorithm computes the new estimate of the parameters by solving the following problem

$$max_{\alpha,\theta} Q(\alpha, \theta)$$

State $Q(\alpha, \theta)$ in terms of $q_i^{(j)}$, Data and the parameters.

Solution:

$$Q(\alpha, \theta) = \sum_{i=1}^{k} \sum_{j=1}^{N} q_i^j \log \alpha_i + \sum_{j=1}^{N} \left(q_i^{(j)} x_j log\theta_i + (1-x_j) q_i^{(j)} log(1-\theta_i) \right)$$

(b) (3 points) Show that from the above problem the new estimate of $\alpha_i^* = \frac{N_i}{N}$. Your answer should express N_i in terms of $q_i^{(j)}$, Data and the current estimate

Solution: It is to be noted that the optimization problem is separable in α and θ_i. Thus

$$\alpha^* = \text{ argmax }_{0 \le \alpha_i \le 1, \sum_{i=1}^k \alpha_i = 1} \sum_{i=1}^{k} \sum_{j=1}^{N} q_i^j \log \alpha_i = \text{ argmax }_{0 \le \alpha_i \le 1, \sum_{i=1}^k \alpha_i = 1} \sum_{i=1}^{k} N_i \log \alpha_i$$

where $N_i = \sum_{j=1}^{N} q_i^j$. Check that the objective can be written as

$$-NKL(\alpha^*, \alpha) + N \sum_{i=1}^{k} \alpha_i^* \log \alpha_i^*$$

This is maximized at $\alpha = \alpha^*$.

(c) (4 points) Find θ_i^*

Solution: Again noting that the problem is separable in the θ_i we have

$$\theta_i^* = \text{ argmax }_{\theta_i} \sum_{j=1}^{N} \left(q_i^{(j)} x_j log\theta_i + (1 - x_j) q_i^{(j)} log(1 - \theta_i) \right)$$

Using the procedure as in the above question,optimality is obtained at

$$\theta_i^* = \frac{\sum_{j=1}^{n} q_i^{(j)} x_j}{n_i}$$

2. Consider a HMM with three hidden states and two observables.

   (a) (2 points) Find A where $P(Z_{t+1} = s_i | Z_t = s_j) = a_{ij}$. It is given that $a_{ij} = \alpha_i(1 - \delta(s_i, s_j)$. Your answer must be numerical.

Solution: From modelling considerations, for every $i \in \{1, 2, 3\}$,

$$\sum_{j=1}^{3} a_{ij} = 1 \implies 2\alpha_i = 1$$

This is true because $a_{ii} = 0$. Thus $a_{ii} = 0, a_{i \neq j} = \frac{1}{2}$

   (b) (2 points) Find B where $P(X_t = v_l | Z_t = s_i) = b_{il}$. It is given that $b_{i1} - b_{i2} = (i - 1)\beta$ for all i, l. Find β_0, the maximum value of β for which B will be a valid parameter matrix.

Solution: Since $b_{i1} + b_{i2} = 1, \implies b_{i1} = \frac{1}{2}(1 + (i - 1)\beta)$. Since $b_{i1} \leq 1$, it implies that $\frac{1}{2}(1 + (i - 1)\beta) \leq 1$ and $\beta \leq \frac{1}{i-1}$. Since $i \leq 3$, the maximum value of β is $\beta_0 = \frac{1}{2}$.

   (c) (5 points) Choose A and B as above with $\beta = \frac{1}{2}\beta_0$. Find $\alpha_2(1) = P(X_1 = x_1, X_2 = v_1, Z_2 = s_1)$ for all i. Assume that $\alpha_1(i)$ are all equal.

Solution: We first note that $\beta = \frac{1}{4}$. $b_{11} = b_{12} = \frac{1}{2}, b_{21} = \frac{5}{8}, b_{22} = \frac{3}{8}, b_{31} = \frac{3}{4}, B_{32} = \frac{1}{4}$.
Since $a_{11} = 0$ $\alpha_2(1) = \sum_{i=1}^{N} \alpha_1(i) a_{i2} b_{i1} = \frac{1}{3}\frac{1}{2}(\frac{5}{8} + \frac{3}{4}) = \frac{5}{16}$

(d) (6 points) We wish to compute $P(Z_2 = s_1 | X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T)$. Assuming the knowledge of $\alpha_2(1)$ give an algorithm which is polynomial in T to compute the desired quantity.

Solution: $P(Z_2 = s_1 | X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T) = \frac{P(Z_2 = s_1, X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T)}{P(X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T)}$

$$P(Z_2 = s_1, X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T) = \alpha_2(1)\beta_2(1)$$

$$P(X_1 = x_1, X_2 = v_1, \ldots, X_T = x_T) = \sum_{i=1}^{3} \alpha_2(i)\beta_2(i)$$

Computation of $\beta_2(i)$ can be accomplished through backward algorithm. The computation cost for each i requires $9(T-t)+3$ computations. Thus the total cost of computation is $9(T-t+1)$.

3. Consider the Ising model. $P(S=s)=\frac{1}{Z}e^{-E(s)}$ with $E(s)=-\frac{1}{2}\sum_{ij}w_{ij}s_is_j.s\in\{-1,1\}^d$

   (a) (5 points) Suppose we approximate $P(S)$ by a factorized distribution such that mean of S_i is 0. Derive a lower bound of $log_2 Z$?

Solution: Since

$$m_i = q_i - (1-q_i), q_i = P(S_i=1), \implies q_i = \frac{1}{2}$$

$$KL(q,P) = \sum_s q(s) log_e \frac{P(s)}{q(s)} \leq 0$$

Implies $log_e Z \geq d log_e 2$ and hence the desired answer is d.

   (b) (5 points) In the above question let the mean of each S_i be m_i. In such a case find $f(m)$ such that

$$log_e Z \geq -E(m) + f(m)$$

Solution:

$$f(m) = -\sum_{i=1}^{d} \frac{1}{2}(1+m_i) log_e \frac{1}{2}(1+m_i) + \frac{1}{2}(1-m_i) log_e \frac{1}{2}(1-m_i)$$

4. Mark all correct choices

   (a) (3 points) Match the following

1	Graphical models	a	acyclic graphs
2	Bayesian Networks	b	Inference is intractable
3	HMM	c	Markov blanket
4	Boltzmann Machine	d	Inference is tractable

1 **c**, 2 **a**, 3 **d**, 4 **b**

   (b) (1 point) Mean-field methods for Boltzmann Machines

   A. It yields a non-deterministic system of equations

B. When used in learning it yields a lower bound on the likelihood

C. The mean-field equations are guaranteed to find the true mean of the hidden variables.

D. It approximates the true distributions by assuming that the hidden variables are independent.

(c) (1 point) It is found that for two covariance matrices, $C^{(1)}, C^{(2)}$, of dimension d,

$$r(C^{(1)}) = \frac{1}{d}, r(C^{(2)}) = 1 - \frac{1}{d}$$

where $r(C) = \frac{\lambda_1(C)}{\sum_{i=1}^{d} \lambda_i(C)}$ for any covariance matrix C. When should we apply PCA?

A. To $C^{(1)}$ but not $C^{(2)}$

B. To $C^{(2)}$ but not $C^{(1)}$

C. To both $C^{(1)}$ and $C^{(2)}$

D. None of them

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded