STAT 479: Homework 3

Due: 11:59PM Feb 29, 2025 by Canvas

---

1. **Variable Elimination in a Bayesian Network** (20 points)

   Consider a Bayesian network with the following structure:

   $$A \to B \to C, \quad A \to D \to C, \quad A \to E$$

   with categorical random variables A, B, C, D, E. The joint probability is thus:

   $$P(A,B,C,D,E) = P(A)P(B \mid A)P(D \mid A)P(C \mid B,D)P(E \mid A).$$

   We want to compute the conditional probability:

   $$P(E = e|B = b) = \frac{P(B = b, E = e)}{P(B = b)}$$

   which requires computing $P(B = b, E = e)$ and normalizing over e. Let's walk through Variable Elimination to see the most efficient way to answer this query.

   (a) **Incorporating Evidence**: Variable elimination starts by incorporating the evidence $B = b$ in the joint distribution. What is the correct way to modify the joint distribution?

      A. Keep all factors as is.

      B. Set $B = b$ in all factors where it appears:

      $$P(A)P(B = b|A)P(D|A)P(C|B = b, D)P(E|A)$$

      C. Remove all factors that contain B: $P(A)P(D|A)P(C|D)P(E|A)$

      D. Sum out B immediately: $\sum_B P(A)P(B|A)P(D|A)P(C|B,D)P(E|A)$

   (b) **Starting Elimination**: After incorporating evidence, we have the factors:

   $$f_1(A) = P(A), \quad f_2(A) = P(B = b|A), \quad f_3(A,D) = P(D|A)$$

   $$f_4(D,C) = P(C|B = b, D), \quad f_5(A,E) = P(E|A)$$

   The first variable we eliminate is C, since it only appears in $f_4(D,C)$. What are the remaining factors after summing out C?

      A. $f_1(A), f_2(A), f_3(A,D), f_5(A,E), f_6(D)$

      B. $f_1(A), f_2(A), f_3(A,D), f_5(A,E), f_6(A,D)$

      C. $f_1(A), f_2(A), f_3(A,D), f_5(A,E), f_6(A,C)$

      D. $f_1(A), f_2(A), f_3(A,D), f_5(A,E), f_6(A)$

(c) **Eliminating D first, then A**: Now we need to eliminate D and A. Suppose we eliminate D first. What would be the resulting factorization?

A. $f_1(A), f_2(A), f_7(A), f_5(A,E)$, where $f_7(A) = \sum_D f_3(A,D)f_6(D)$
B. $f_1(A), f_2(A), f_7(D), f_5(A,E)$, where $f_7(D) = \sum_D f_3(A,D)f_6(D)$
C. $f_1(A), f_2(A), f_7(A,E), f_6(D)$, where $f_7(A,E) = \sum_D f_3(A,D)f_5(A,E)$
D. $f_1(A), f_2(A), f_3(A,D), f_6(D), f_7(A,E)$, where $f_7(A,E) = \sum_D f_3(A,D)f_5(A,E)$

(d) **Eliminating A first, then D**: Now, instead, suppose we eliminate A before eliminating D. What would be the resulting factorization?

A. $f_7(D,E), f_6(D)$, where $f_7(D,E) = \sum_A f_1(A)f_2(A)f_3(A,D)f_5(A,E)$
B. $f_1(A), f_2(A), f_7(A,D), f_5(A,E)$, where $f_7(A,D) = \sum_A f_3(A,D)f_6(D)$
C. $f_1(A), f_2(A), f_7(A,E), f_6(D)$, where $f_7(A,E) = \sum_A f_3(A,D)f_5(A,E)$
D. $f_1(A), f_2(A), f_3(A,D), f_7(A,E)$, where $f_7(A,E) = \sum_A f_3(A,D)f_5(A,E)$

(e) **Comparing the Orders**: Based on your calculations above, which order is more efficient in terms of minimizing the largest intermediate factor?

A. Eliminating A before D is always more efficient.
B. Eliminating A before eliminating D is more efficient when $|D| > |A|$.
C. Both orders create the same largest intermediate factor in all cases.

Answer:

(a)

(b)

(c)

(d)

(e)

2. **Newton-Raphson for Poisson GLM** (30 points)

Consider a generalized linear model (GLM) where the response y follows a Poisson distribution in exponential family form $p(y|\eta) = h(y)\exp(\eta T(y) - A(\eta))$, with natural parameter $\eta = \mathbf{x}^T\beta$.

(a) **Exponential Family Form**: Derive the forms of $T(y)$, $A(\eta)$, and $E[y]$ for the Poisson distribution $P(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$ when $\eta = \mathbf{x}^T\beta$. Select the correct expression:

A. $T(y) = y$, $A(\eta) = e^\eta$, $E[y] = e^{\mathbf{x}^T\beta}$

B. $T(y) = e^y$, $A(\eta) = \eta$, $E[y] = \mathbf{x}^T\beta$

C. $T(y) = y$, $A(\eta) = \eta^2$, $E[y] = 2\mathbf{x}^T\beta$

D. $T(y) = \ln y$, $A(\eta) = e^{-\eta}$, $E[y] = e^{-\mathbf{x}^T\beta}$

(b) **Log-Likelihood**: For a dataset $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$ with $y_j \sim \text{Poisson}(e^{\mathbf{x}_j^T\beta})$, write the log-likelihood $\ell(\beta)$. Select the correct expression:

A. $\ell(\beta) = \sum_{j=1}^n [y_j\mathbf{x}_j^T\beta - e^{\mathbf{x}_j^T\beta} - \ln(y_j!)]$

B. $\ell(\beta) = \sum_{j=1}^n [y_j e^{\mathbf{x}_j^T\beta} - \mathbf{x}_j^T\beta - y_j^2]$

C. $\ell(\beta) = \sum_{j=1}^n [\ln(y_j) - e^{\mathbf{x}_j^T\beta} + \mathbf{x}_j^T\beta]$

D. $\ell(\beta) = \sum_{j=1}^n [y_j - \ln(e^{\mathbf{x}_j^T\beta}) + \ln(y_j!)]$

(c) **Gradient**: Compute the gradient $\nabla\ell(\beta)$ of the log-likelihood. Select the correct expression:

A. $\nabla\ell(\beta) = \sum_{j=1}^n (y_j - e^{\mathbf{x}_j^T\beta})\mathbf{x}_j$

B. $\nabla\ell(\beta) = \sum_{j=1}^n (e^{\mathbf{x}_j^T\beta} - y_j)\mathbf{x}_j$

C. $\nabla\ell(\beta) = \sum_{j=1}^n y_j\mathbf{x}_j e^{\mathbf{x}_j^T\beta}$

D. $\nabla\ell(\beta) = \sum_{j=1}^n \mathbf{x}_j / e^{\mathbf{x}_j^T\beta}$

(d) **Newton-Raphson Update**: Given the Hessian $\mathbf{H} = \sum_{j=1}^n e^{\mathbf{x}_j^T\beta}\mathbf{x}_j\mathbf{x}_j^T$, derive the Newton-Raphson update rule for β. Select the correct expression:

A. $\beta^{(t+1)} = \beta^{(t)} - \left(\sum_{j=1}^n e^{\mathbf{x}_j^T\beta^{(t)}}\mathbf{x}_j\mathbf{x}_j^T\right)^{-1} \sum_{j=1}^n (y_j - e^{\mathbf{x}_j^T\beta^{(t)}})\mathbf{x}_j$

B. $\beta^{(t+1)} = \beta^{(t)} + \left(\sum_{j=1}^n y_j\mathbf{x}_j\mathbf{x}_j^T\right)^{-1} \sum_{j=1}^n e^{\mathbf{x}_j^T\beta^{(t)}}\mathbf{x}_j$

C. $\beta^{(t+1)} = \beta^{(t)} - \left(\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^T\right)^{-1} \sum_{j=1}^n y_j\mathbf{x}_j$

D. $\beta^{(t+1)} = \beta^{(t)} + e^{\mathbf{x}_j^T\beta^{(t)}} \sum_{j=1}^n (y_j - \mathbf{x}_j^T\beta^{(t)})\mathbf{x}_j$

Answer:

(a)

(b)

(c)

(d)

3. **MLE for Bayesian Network** (20 points)

Consider a simple Bayesian network $A \to B$ with binary variables $A, B \in \{0, 1\}$ and joint distribution $P(A, B) = P(A)P(B|A)$. You have a complete dataset of n observations, where $n_{a,b}$ denotes the number of times $(A = a, B = b)$ occurs.

Let $\theta_1 = P(A = 1), \theta_2 = P(B = 1 \mid A = 0), \theta_3 = P(B = 1 \mid A = 1)$.

(a) **Log-Likelihood**: Write the log-likelihood $\ell(\theta_1, \theta_2, \theta_3)$. Select the correct expression:

A. $\ell(\theta_1, \theta_2, \theta_3) = n_{0,0} \ln(1 - \theta_1) + n_{0,1} \ln(1 - \theta_1)\theta_2 + n_{1,0} \ln \theta_1(1 - \theta_3) + n_{1,1} \ln \theta_1\theta_3$

B. $\ell(\theta_1, \theta_2, \theta_3) = n_{0,0} \ln[(1-\theta_1)(1-\theta_2)] + n_{0,1} \ln[(1-\theta_1)\theta_2] + n_{1,0} \ln[\theta_1(1-\theta_3)] + n_{1,1} \ln[\theta_1\theta_3]$

C. $\ell(\theta_1, \theta_2, \theta_3) = n \ln \theta_1 + n_{0,1} \ln \theta_2 + n_{1,1} \ln \theta_3$

D. $\ell(\theta_1, \theta_2, \theta_3) = n_{0,0} \ln \theta_1 + n_{0,1} \ln \theta_2 + n_{1,0} \ln \theta_3$

(b) **MLE for** $P(B = 1|A = 0)$: Derive the MLE for $P(B = 1|A = 0)$ by maximizing the log-likelihood. Select the correct estimator:

A. $\hat{\theta}_{2,MLE} = \frac{n_{0,1}}{n_{0,0}+n_{0,1}}$

B. $\hat{\theta}_{2,MLE} = \frac{n_{0,1}}{n}$

C. $\hat{\theta}_{2,MLE} = \frac{n_{0,0}+n_{0,1}}{n}$

D. $\hat{\theta}_{2,MLE} = \frac{n_{1,1}}{n_{1,0}+n_{1,1}}$

Answer:

(a)

(b)

4. **Sharding a Bayesian Network** (20 points)

Consider a Bayesian network with structure $X_1 \to X_2 \to X_3$, $X_1 \to X_4$, $X_2 \to X_4$, and joint distribution $P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_1, X_2)$.

(a) **Sharding Variable**: Determine which set of variables, when conditioned on, shards the BN into two conditionally independent subgraphs, one containing X_1 and the other X_3. Select the correct minimal set:

A. $\{X_2\}$
B. $\{X_1, X_2\}$
C. $\{X_2, X_4\}$
D. $\{X_4\}$

Hint: Use d-separation to find a minimal set that blocks all paths between X_1 and X_3.

(b) **Using Sharding to Simplify Computation**: Using the minimal sharding set from part (a), how can we compute $P(X_4)$ in a way that takes advantage of conditional independence?

A. Compute $P(X_4)$ in two independent steps: first sum over X_1, then over X_2.
B. Compute $P(X_4)$ in one step by marginalizing over X_1 and X_2 together.
C. Compute $P(X_4)$ by conditioning on X_3, then summing over X_1, X_2.
D. Compute $P(X_4)$ by first marginalizing X_3, then summing over X_1, X_2.

(c) **Efficiency of Distributed Computation**: Suppose we distribute inference across separate computing units, where one unit handles (X_1, X_4) and another handles (X_3). Which of the following best describes how sharding reduces computational complexity?

A. It allows us to compute $P(X_4)$ and $P(X_3)$ independently, reducing redundant summations.
B. It changes the factorization structure to remove dependencies between all variables.
C. It eliminates the need for marginalization when computing any query.
D. It makes all variables independent, allowing direct computation of individual probabilities.

Answer:

(a)

(b)

(c)

5. **Mid-Semester Feedback** (10 points)

   We're one-third of the way through Stat 479! Your feedback will help improve the course.

   (a) **Course Pacing**: How do you feel about the pace of the course so far?
      A. The pace is much too fast, and I struggle to keep up with the content.
      B. The pace is slightly too fast, but I can manage with extra effort.
      C. The pace is about right, balancing challenge and understanding.
      D. The pace is too slow, and I'd prefer more challenging material sooner.

   (b) **Lecture Clarity**: How clear are the lectures' explanations and examples?
      A. Lectures are very unclear, and I often leave confused.
      B. Lectures are somewhat unclear, needing more examples or simpler explanations.
      C. Lectures are mostly clear, with minor areas for improvement.
      D. Lectures are very clear, and I grasp the material well from them.

   (c) **Assignment Difficulty**: How do you find the difficulty of the homework assignments?
      A. Assignments are far too difficult, requiring excessive time or external help.
      B. Assignments are challenging but manageable with effort and course resources.
      C. Assignments are appropriately difficult, aligning well with lecture content.
      D. Assignments are too easy, and I'd like more complex problems.

   (d) **Resource Usefulness**: How useful are the course resources (e.g., slides, notes, textbooks, office hours) in supporting your learning?
      A. Resources are not useful, and I rarely rely on them.
      B. Resources are somewhat useful, but I need more or better options.
      C. Resources are generally useful, meeting most of my needs.
      D. Resources are highly useful, significantly aiding my understanding.

   (e) **Time on Homework**: On average, how much time do you spend per week on homework assignments for this course?
      A. 0-5 hours
      B. 6-10 hours
      C. 11-15 hours
      D. 16+ hours

   (f) **Time on Readings**: On average, how much time do you spend per week on assigned readings or supplemental materials for this course?

A. 0-1 hours
B. 2-3 hours
C. 4-5 hours
D. 6+ hours

(g) **Time on Lecture Review**: On average, how much time do you spend per week reviewing lecture notes or recordings outside of class?

A. 0-1 hours
B. 2-3 hours
C. 4-5 hours
D. 6+ hours

(h) **Open Feedback**: In 3-5 sentences, provide any additional feedback about your experience in the course so far. What's working well, and what could be improved for the second half of the semester?

Answer:

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)