STAT 479: Homework 4

Due: 11:59PM Mar 8, 2025 by Canvas

---

1. **MLE in Bayesian Networks** (20 points)

Consider the following Bayesian Network with four binary variables:

$$X_1 \to X_2, \quad X_1 \to X_3, \quad X_2 \to X_4, \quad X_3 \to X_4$$

The dataset consists of:

X_1	X_2	X_3	X_4	**Count**
0	0	0	0	30
0	0	0	1	10
0	0	1	0	20
0	0	1	1	20
0	1	0	0	25
0	1	0	1	15
0	1	1	0	35
0	1	1	1	15
1	0	0	0	20
1	0	0	1	10
1	0	1	0	25
1	0	1	1	15
1	1	0	0	30
1	1	0	1	20
1	1	1	0	40
1	1	1	1	30

(a) Compute the MLE for $P(X_2 = 1|X_1 = 0)$. Round to the nearest tenth.

A. 0.4
B. 0.5
C. 0.6
D. 0.7

(b) Compute the MLE for $P(X_4 = 1|X_2 = 1, X_3 = 0)$. Round to the nearest tenth.

A. 0.3
B. 0.4
C. 0.5
D. 0.6

(c) The variable X_4 is a collider. What happens when we condition on X_4?

A. It blocks the path between X_2 and X_3, making them independent.

B. It introduces a correlation between X_2 and X_3.

C. It has no effect unless we also condition on X_1.

D. It enforces complete independence among all variables.

(d) After fitting our estimate of $P(X_2|X_1)$, someone informs us about an unobserved variable L that influences both X_2 and X_3, but we know we cannot observe L. What should we do to our estimate?

A. Keep the estimate unchanged, since we already conditioned on X_1.

B. Acknowledge potential bias and use sensitivity analysis to estimate the possible impact of L.

C. Widen our confidence intervals to ensure the true value is captured.

D. Ignore L since unobserved variables do not affect MLE estimates.

Answer:

(a)

(b)

(c)

(d)

2. **MRFs** (20 points)

Consider a Markov Random Field (MRF) with the following structure:

- Nodes: X_1, X_2, X_3, X_4
- Edges: $(X_1, X_2), (X_2, X_3), (X_3, X_4), (X_4, X_1), (X_1, X_3)$
- The joint probability distribution is given by:

$$P(X_1, X_2, X_3, X_4) \propto \exp\left(\sum_{(i,j)\in E} \theta_{ij} X_i X_j\right)$$

(a) Which of the following is a correct factorization of this MRF?

A. $P(X_1, X_2, X_3, X_4) = \phi(X_1, X_2)\phi(X_2, X_3)\phi(X_3, X_4)\phi(X_4, X_1)$

B. $P(X_1, X_2, X_3, X_4) = \phi(X_1, X_2, X_3)\phi(X_3, X_4)$

C. $P(X_1, X_2, X_3, X_4) = \phi(X_1, X_2)\phi(X_2, X_3)\phi(X_3, X_4)\phi(X_4, X_1)\phi(X_1, X_3)$

D. $P(X_1, X_2, X_3, X_4) = \phi(X_1, X_2, X_3, X_4)$

(b) Why is computing the partition function Z difficult for large MRFs?

A. The partition function requires summing over an exponential number of terms.

B. The partition function depends on the parameters θ_{ij}, which are unknown.

C. The partition function does not exist for undirected graphical models.

D. The partition function is always equal to 1.

(c) Removing edge (X_1, X_3) changes which independence property?

A. $X_1 \perp X_3 \mid \{X_2, X_4\}$

B. $X_1 \perp X_3 \mid X_2$

C. $X_1 \perp X_4 \mid X_2$

D. $X_1 \perp X_2 \mid X_4$

(d) A researcher is using Graphical Lasso to learn the structure of a Markov Random Field (MRF) from data. However, they observe that small changes in the regularization parameter result in large differences in the learned graph structure, with many edges appearing or disappearing unpredictably.

Which of the following is the most likely reason for this instability?

A. The sample size is too small relative to the number of variables, leading to an unstable covariance estimate.

B. The true MRF is not connected, so regularization causes disjoint components to form.

C. Graphical Lasso is not a consistent estimator and always leads to instability.

D. The data is non-Gaussian, violating the assumptions of Graphical Lasso.

Answer:

(a)

(b)

(c)

(d)

3. **Gaussian Graphical Models – Step-by-Step Proof** (20 points)

We consider an undirected graphical model where a set of random variables $X = \{X_1, ..., X_d\}$ follows a joint Gaussian distribution:

$$X \sim \mathcal{N}(\mu, \Sigma)$$

where Σ is the **covariance matrix** ($\Sigma \in \mathbb{S}^{++}$), and its inverse $\Theta = \Sigma^{-1}$ is the **precision matrix**. Given two nodes X_j, X_k, and the remaining variables $Z = \{X_i \mid i \notin \{j, k\}\}$, we want to show:

$$X_j \perp X_k \mid Z \quad \text{if and only if} \quad \Theta_{jk} = 0.$$

(a) **Step 1: Conditional Independence in Gaussian Graphical Models**
The joint probability density function of a multivariate Gaussian is:

$$p(X) \propto \exp\left(-\frac{1}{2} X^T \Theta X\right)$$

From this, we can see that the entries of the **precision matrix** Θ describe:

A. The marginal variances of X.
B. The correlations between variables.
C. The structure of the conditional dependencies between variables.
D. The eigenvalues of the covariance matrix.

(b) **Step 2: Identifying the Relevant Conditional Distribution**
To determine whether X_j and X_k are conditionally independent given Z, we need to examine:

A. The marginal covariance matrix of X_j and X_k.
B. The conditional precision matrix of X_j and X_k given Z.
C. The determinant of Θ.
D. The sum of all precision matrix entries.

(c) **Step 3: Expressing the Conditional Distribution**
When conditioning on Z, the precision matrix for $X_j, X_k \mid Z$ is given by:

$$\Theta_{(j,k)|Z} = \Theta_{\{j,k\},\{j,k\}} - \Theta_{\{j,k\},Z} \Theta_{Z,Z}^{-1} \Theta_{Z,\{j,k\}}$$

Given this, which of the following must be true for X_j and X_k to be **conditionally independent given** Z?

A. The determinant of $\Theta_{(j,k)|Z}$ is zero.
B. The off-diagonal entry of $\Theta_{(j,k)|Z}$ (i.e., $(\Theta_{(j,k)|Z})_{jk}$) is zero.
C. The sum of all precision matrix entries is zero.

D. The covariance matrix entry Σ_{jk} must be zero.

(d) **Step 4: Relating This to the Original Precision Matrix**

From our result, we see that the conditional independence condition is met when:

A. $\Theta_{jk} = 0$.
B. $\Sigma_{jk} = 0$.
C. $\Theta_{jk}^{-1} = 0$.
D. $\Theta_{jj} = 0$.

(e) **Step 5: Interpreting the Graph Structure**

In an undirected Gaussian graphical model, an edge exists between two nodes X_j and X_k if and only if:

A. $\Theta_{jk} \neq 0$.
B. $\Sigma_{jk} \neq 0$.
C. $\Theta_{jk}^{-1} \neq 0$.
D. $\Theta_{jj} = 0$.

Answer:

(a)

(b)

(c)

(d)

(e)

4. **Introduction to Conditional Random Fields** (20 points)

A **Conditional Random Field (CRF)** is an undirected graphical model that models a conditional probability distribution $P(Y \mid X)$ instead of the joint distribution $P(X, Y)$. This is particularly useful for structured prediction problems where labels $Y = (Y_1, ..., Y_n)$ depend on input features $X = (X_1, ..., X_n)$.

Consider a linear-chain CRF with sequence labels Y_1, Y_2, Y_3 conditioned on observations X_1, X_2, X_3. The probability of a labeling is given by:

$$P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{3} \psi(Y_i, Y_{i+1}, X)$$

where:

- $\psi(Y_i, Y_{i+1}, X) = \exp(\theta_{Y_i,Y_{i+1}} + \sum_k w_k f_k(Y_i, X))$
- $Z(X)$ is the partition function ensuring the probability distribution normalizes properly.

(a) **Understanding Factorization** What is the key difference between a CRF and a Markov Random Field (MRF)?

A. CRFs model $P(Y \mid X)$, whereas MRFs model $P(X, Y)$.

B. CRFs use directed edges, while MRFs use undirected edges.

C. CRFs require a fully connected graph, whereas MRFs do not.

D. MRFs only allow discrete variables, while CRFs allow continuous ones.

(b) **Computing Conditional Probabilities** Suppose we are given parameter values $\theta_{Y_i,Y_{i+1}}$ and feature weights w_k. Which of the following is true about the conditional probability $P(Y \mid X)$?

A. It is computed by normalizing the product of potential functions over all possible label sequences.

B. It can be computed directly without using the partition function.

C. It is always independent of the input features X.

D. It is equal to the sum of all local potential functions.

(c) **Parameter Estimation** What is the most common method for learning CRF parameters θ and w from labeled data?

A. Maximum Likelihood Estimation (MLE) via gradient descent.

B. Expectation-Maximization (EM) since CRFs have latent variables.

C. Bayesian inference using Gibbs Sampling.

D. k-Nearest Neighbors since CRFs are nearest-neighbor models.

Answer:

(a)

(b)

(c)

5. **Introduction to Importance Sampling** (20 points)

Importance sampling is a method for estimating expectations of a function $f(x)$ under a distribution $p(x)$, when direct sampling from $p(x)$ is difficult. Instead, we sample from an easier proposal distribution $q(x)$ and use importance weights to correct for the difference. We define:

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx$$

and approximate this expectation using L samples $x^{(1)}, \ldots, x^{(L)} \sim q(x)$:

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{\sum_{i=1}^{L} u_i} \sum_{i=1}^{L} f(x^{(i)})u_i,$$

where the unnormalized importance weights are:

$$u_i = \frac{p(x^{(i)})}{q(x^{(i)})}.$$

(a) **Why Use Importance Sampling?** Which of the following best describes why importance sampling is useful?

A. It allows us to estimate expectations when direct sampling from $p(x)$ is difficult.

B. It generates exact samples from $p(x)$ without error.

C. It reduces variance compared to sampling directly from $p(x)$.

D. It removes the need to compute the normalizing constant of $p(x)$.

(b) **Understanding Importance Weights** What can we say about the expected value of the importance weight u_i under the proposal distribution $q(x)$?

A. $\mathbb{E}_{q(x)}[u_i] = \mathbb{E}_{p(x)}[u_i]$, since importance weights correct for sampling bias.

B. $\mathbb{E}_{q(x)}[u_i] = 0$ if $p(x) \neq q(x)$.

C. $\mathbb{E}_{q(x)}[u_i] = 1$, ensuring the estimator remains unbiased.

D. $\mathbb{E}_{q(x)}[u_i]$ grows as the difference between $p(x)$ and $q(x)$ increases.

(c) **Variance of Importance Weights** What is the closed-form expression for the variance of the importance weights $u_i = \frac{p(x)}{q(x)}$ under the proposal distribution $q(x)$?

A. $\text{Var}_{q(x)}[u_i] = \mathbb{E}_{q(x)}[u_i^2] - \mathbb{E}_{q(x)}[u_i]^2$.

B. $\text{Var}_{q(x)}[u_i] = \mathbb{E}_{p(x)}[u_i] - 1$.

C. $\text{Var}_{q(x)}[u_i] = \mathbb{E}_{q(x)}[p(x)] - \mathbb{E}_{q(x)}[q(x)]$.

D. $\text{Var}_{q(x)}[u_i] = \frac{\mathbb{E}_{p(x)}[u_i^2]}{\mathbb{E}_{q(x)}[u_i]}$.

(d) **Variance of Importance Weights and High-Dimensional Spaces** Suppose that $p(x)$ and $q(x)$ are both factored distributions over d independent dimensions:

$$p(x) = \prod_{i=1}^{d} p_i(x_i), \quad q(x) = \prod_{i=1}^{d} q_i(x_i).$$

How does the variance of importance weights scale with d, assuming that $p_i(x_i)/q_i(x_i)$ has variance v for each individual dimension?

A. The variance remains constant as d increases.
B. The variance scales linearly as $\mathcal{O}(d)$.
C. The variance scales exponentially as $\mathcal{O}(v^d)$, where v is the per-dimension variance of importance weights.
D. The variance decreases as d increases due to averaging effects.

Answer:

(a)

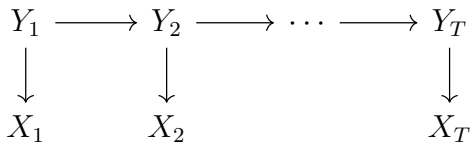
(b)

(c)

(d)

6. **Parameter Learning in HMMs** (0 points (OPTIONAL))

For those of you who want to start flexing your programming muscles, please enjoy implementing this E-M algorithm. This question is OPTIONAL and will not contribute to your grade.



Consider an HMM with $Y_t \in [M]$, $X_t \in \mathbb{R}^K$ $(M, K \in \mathbb{N})$. Let $(\pi, A, \{\mu_i, \sigma_i^2\}_{i=1}^M)$ be its parameters, where $\pi \in \mathbb{R}^M$ is the initial state distribution, $A \in \mathbb{R}^{M \times M}$ is the transition matrix, $\mu_i \in \mathbb{R}^K$ and $\sigma_i^2 > 0$ are parameters of the emission distribution, which is defined to be an isotropic Gaussian. In other words,

$$P(Y_1 = i) = \pi_i \tag{1}$$
$$P(Y_{t+1} = j | Y_t = i) = A_{ij} \tag{2}$$
$$P(X_t | Y_t = i) = \mathcal{N}(X_t; \mu_i, \sigma_i^2 I). \tag{3}$$

In the attached `baum_welch.py` file, implement the Baum-Welch (EM) algorithm that estimates parameters from data $\boldsymbol{X} \in \mathbb{R}^{N \times T \times K}$, which is a collection of N observed sequences of length T. Please find unimplemented TODO blocks in the template for you to implement. The template has its own toy problem to verify the implementation.