Homework 4 - Introduction to Probabilistic Graphical Models

kipngeno koech - bkoech

April 23, 2025

1 Structure Learning

1.1 Tree-Selection and the Chow-Liu Algorithm

Use the Chow-Liu Algorithm to learn the model (tree and parameters) that generated the following data.

Data

1	0	0	1
0	1	0	0
1	1	0	1
1	1	1	1
0	1	0	1
1	0	0	1
1	1	0	1
0	0	1	0
0	1	0	1

Figure 1: The tree structure of the data.

1.2 Scoring Function

Using the BIC scoring metrics below, compute the model (graph and potentials) that generated the data below.

BIC:

$$S(G, \theta; D) = LL(\theta; D) - \phi(|D|)\|G\| \quad \text{where} \quad \phi(t) = \frac{\log(t)}{2}$$

Maximizing the score:

$$S_{\max}(G, D) = \max_{\theta} \big(S(G, \theta; D)\big)$$

Data

1	0	0
0	1	0
1	1	0
1	1	1
0	1	0
1	0	0
1	1	0
0	0	1
0	1	0

Figure 2: The data used to compute the BIC score.

2 Variational Inference

2.1 Mean-Field Approximation for Multivariate Gaussians

In this question, we'll explore how accurate a Mean-Field approximation can be for an underlying multivariate Gaussian distribution. Assume we have observed data $X \in \mathbb{R}^{2\times n}$ where each column $X_{\cdot,i} \triangleq x^{(i)} \in \mathbb{R}^2$ is a sample that was drawn from a 2-dimensional Gaussian distribution $x^{(i)} \sim p(\cdot; \mu, \Lambda^{-1})$.

$$p(x; \mu, \Lambda) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}^{-1}\right) \tag{1}$$

Note here that we're using the precision matrix $\Lambda = \Sigma^{-1}$. An additional property of the precision matrix is that it is symmetric, so $\Lambda_{12} = \Lambda_{21}$. (This is a convenient simplifying assumption.) We will approximate this 2-dimensional Gaussian with a mean field approximation, $q(x) = q(x_1)q(x_2)$, the product of two 1-dimensional distributions $q(x_1)$ and $q(x_2)$. For now, we won't assume any form for these distributions.

1. **Short Answer:** Write down the equation for $\log p(X)$. (For this question, you can leave all of the parameters in terms of vectors and matrices, not their subcomponents.)

2. **Short Answer:** Group together everything that involves X_1 and remove anything involving X_2. We claim that there exists some distribution $q^*(X) = q^*(X_1)q^*(X_2)$ that minimizes the KL divergence $q^* = \arg\min_q \mathrm{KL}(q\|p)$. Furthermore, said distribution will have a component $q^*(X_1)$ that will be proportional to the quantity you find below. Write that term that is proportional to $q^*(X_1)$.

   It can be shown that this implies that $q(X_1)$ (and therefore $q(X_2)$) is a Gaussian distribution:

   $$q(x_1) = \mathcal{N}\left(x_1; m_1, \Lambda_{11}^{-1}\right)$$

   where

   $$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\big(E[x_2] - \mu_2\big)$$

   Using these facts, we'd like to explore how well our approximation can model the underlying distribution.

3. Suppose the parameters of the true distribution are

   $$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}.$$

   (a) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(X_1)$?

   (b) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(X_1)$?

   (c) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(X_2)$?

   (d) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(X_2)$?

   (e) (5 points) **Plot:** Provide a computer-generated contour plot to show the result of our approximation $q^*(X)$ and the true underlying Gaussian $p(X; \mu, \Lambda)$ for the parameters given above.

4. Suppose the parameters of the true distribution are

   $$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

   (a) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(X_1)$?

   (b) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(X_1)$?

   (c) (2 points) **Numerical Answer:** What is the value of the mean of the Gaussian for $q^*(X_2)$?

   (d) (2 points) **Numerical Answer:** What is the value of the variance of the Gaussian for $q^*(X_2)$?

   (e) (5 points) **Plot:** Provide a computer-generated contour plot to show the result of our approximation $q^*(X)$ and the true underlying Gaussian $p(X; \mu, \Lambda)$ for the parameters given above.

5. (2 points) Describe in words how the plots you generated provide insight into the behavior of minimization of KL(q——p) with regards to the low probability and high probability regions of the true vs. approximate distributions.

   The contour plots illustrate the differences between the true distribution $p(X; \mu, \Lambda)$ and the approximate distribution $q^*(X)$. In high-probability regions, where the true distribution has significant density, the approximation $q^*(X)$ closely aligns with $p(X)$. This is because minimizing $\mathrm{KL}(q\|p)$ prioritizes matching the high-probability regions of $p(X)$, as discrepancies in these regions contribute more to the KL divergence. Conversely, in low-probability regions, the approximation $q^*(X)$ may deviate more significantly from $p(X)$, as these regions have less influence on the overall KL divergence. This behavior reflects the trade-off inherent in the mean-field approximation, where the focus is on capturing the most probable regions of the true distribution while potentially sacrificing accuracy in less probable areas.

2.2 Variational Inference vs. Monte Carlo Methods

Let's end with a brief comparison between variational methods and MCMC methods. We have seen that both classes of methods can be used for learning in scenarios involving latent variables, but both have their own sets of advantages and disadvantages. For each of the following statements, specify whether they apply more suitably to VI or MCMC methods:

1. (2 points) Transforms inference into optimization problems.
   ◯ Variational Inference ✓
   ◯ MCMC

2. (2 points) Is easier to integrate with back-propagation.
   ◯ Variational Inference ✓
   ◯ MCMC

3. (2 points) Involves more stochasticity.
   ◯ Variational Inference
   ◯ MCMC ✓

4. (2 points) Converges to the true distribution.
   ◯ Variational Inference
   ◯ MCMC ✓

5. (2 points) Is higher variance under limited computational resources.
   ◯ Variational Inference
   ◯ MCMC ✓