CSED515 Machine Learning - Assignment 4

Dongwoo Kim

Due date: 23:59pm, June 2th, 2023

Remark

Assignment Submission. All students must submit their homework via PLMS. Submit your answer on PLMS in a single PDF file named with `your_student_id.pdf`. You can scan your hand-written answers or write your answer with a tablet (If you are ambitious enough, try to use LaTeX to write your answers).

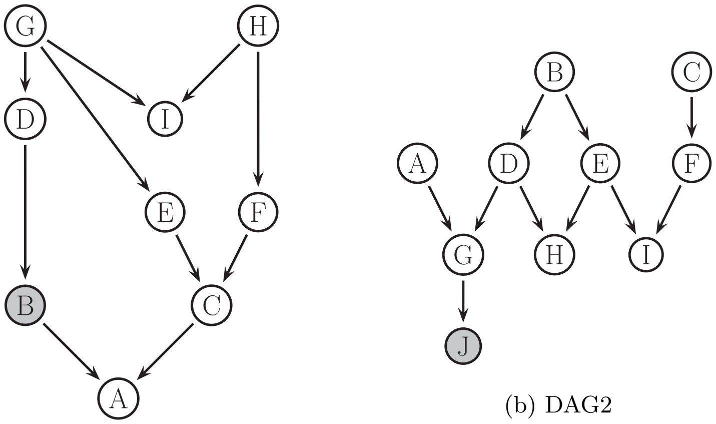
Late Homework Policy. We do not allow late submission. If you have any question regarding this, please send an email to the lecturer.

Honor Code. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Using code or solutions obtained from the web (GitHub/Google/ etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Importance Notice Note that the first assignment aims to identify your background on linear algebra, probability and statistics, and vector calculus. Show your work with your answers. You will get zero score if there is no proper explanation. You can use any external resources to solve the problems, however, please make sure that you have understood the details of what you have written. If you think you are not ready to solve these questions, please consider to take preliminary classes before taking this class. The list of preliminary courses are provided during the second lecture.

1. **Conditional Independence** [20pt]

   Here we compute some global independence statements from some directed graphical models.



(a) DAG1

(b) DAG2

Figure 1: Graphical Models

   (a) Consider the DAG in Figure 1a. List all variables that are independent of A given evidence on B.

   (b) Consider the DAG in Figure 1b. List all variables that are independent of A given evidence on J.

2. **Bayes Net** [20pt]

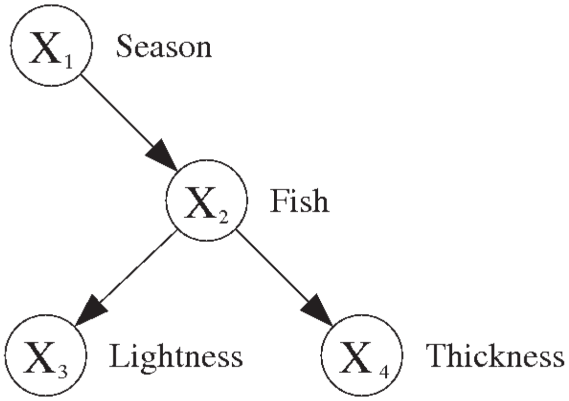


Figure 2: Fishing Net

Consider the Bayes net shown in Figure 2. Here, the nodes represent the following variables $X_1 \in \{$ winter, spring, summer, autumn $\}, X_2 \in \{$ salmon, sea bass $\}$

$$X_3 \in \{ \text{ light, medium, dark } \}, X_4 \in \{ \text{ wide, thin } \}$$

The corresponding conditional probability tables are

$$p(x_1) = \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix}, p(x_2 \mid x_1) = \begin{pmatrix} .9 & .1 \\ .3 & .7 \\ .4 & .6 \\ .8 & .2 \end{pmatrix} \tag{1}$$

$$p(x_3 \mid x_2) = \begin{pmatrix} .33 & .33 & .34 \\ .8 & .1 & .1 \end{pmatrix}, p(x_4 \mid x_2) = \begin{pmatrix} .4 & .6 \\ .95 & .05 \end{pmatrix} \tag{2}$$

Note that in $p(x_4 \mid x_2)$, the rows represent x_2 and the columns x_4 (so each row sums to one and represents the child of the conditional probability). Thus $p(x_4 = \text{thin} \mid x_2 = \text{sea bass}) = 0.05, p(x_4 = \text{thin} \mid x_2 = \text{salmon}) = 0.6$, etc. Answer the following queries.

(a) [10pt] Suppose the fish was caught on December 20 - the end of autumn and the beginning of winter and thus let $p(x_1) = (.5, 0, 0, .5)$ instead of the above prior. (This is called soft evidence, since we do not know the exact value of X_1, but we have a distribution over it.) Suppose the lightness has not been measured but it is known that the fish is thin. Classify the fish as salmon or sea bass.

(b) [10pt] Suppose all we know is that the fish is thin and medium lightness. What season is it now, most likely? Use $p(x_1) = \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix}$

3. **EM for mixtures of Bernoullis** [40pt].

Consider mixtures of Bernoulli distribution where the generative process can be described as
For each data point x_i:

$$\begin{aligned} z_i &\sim \text{Categorical}(\boldsymbol{\pi}) \\ x_i &\sim \text{Bernoulli}(\mu_{z_i}) \end{aligned}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_K]$ and $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_K]$ are model parameters.

(a) [10pt] Write down the complete data log likelihood of the model with dataset $D = \{x_1, ..., x_N\}$.

(b) [10pt] Compute the posterior distribution over z_i given model parameters.

(c) [10pt] Derive the M step for ML estimation of a mixture of Bernoullis.

(d) [10pt] Assume that we place Beta(α, β) prior over μ_k, i.e. $\mu_k \sim \text{Beta}(\alpha, \beta), \forall k$. Show that the M step for MAP estimation of a mixture of Bernoullis with a Beta(α, β) is given by

$$\mu_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{(\sum_i r_{ik}) + \alpha + \beta - 2}$$

4. **PCA** [20pt]

Let $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ be the first k eigenvectors with largest eigenvalues of $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$, i.e., the principal basis vectors. These satisfy

$$\boldsymbol{v}_j^\top \boldsymbol{v}_k = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

We will construct a method for finding the $\boldsymbol{v}_j$ sequentially.

As we showed in class, $\boldsymbol{v}_1$ is the first principal eigenvector of $\mathbf{S}$, and satisfies $\mathbf{S}\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$. Now define $\tilde{\boldsymbol{x}}_i$ as the orthogonal projection of $\boldsymbol{x}_i$ onto the space orthogonal to $\boldsymbol{v}_1$:

$$\tilde{\boldsymbol{x}}_i = \mathbf{P}_{\perp \boldsymbol{v}_1} \boldsymbol{x}_i = \left(\mathbf{I} - \boldsymbol{v}_1 \boldsymbol{v}_1^\top\right) \boldsymbol{x}_i$$

Define $\tilde{\mathbf{X}} = [\tilde{\boldsymbol{x}}_1; \ldots; \tilde{\boldsymbol{x}}_n]$ as the deflated matrix of rank $d-1$, which is obtained by removing from the d dimensional data component that lies in the direction of the first principal direction:

$$\tilde{\mathbf{X}} = \left(\mathbf{I} - \boldsymbol{v}_1 \boldsymbol{v}_1^\top\right) \mathbf{X}$$

(a) [10pt] Using the facts that $\mathbf{X}\mathbf{X}^\top \boldsymbol{v}_1 = n\lambda_1 \boldsymbol{v}_1$ (and hence $\boldsymbol{v}_1^\top \mathbf{X}\mathbf{X}^T = n\lambda_1 \boldsymbol{v}_1^\top$) and $\boldsymbol{v}_1^\top \boldsymbol{v}_1 = 1$, show that the covariance of the deflated matrix is given by

$$\tilde{\mathbf{S}} \triangleq \frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \frac{1}{n}\mathbf{X}\mathbf{X}^\top - \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\top$$

(b) [10pt] Let $\boldsymbol{u}$ be the principal eigenvector of $\tilde{\mathbf{S}}$. Explain why $\boldsymbol{u} = \boldsymbol{v}_2$. (You may assume $\boldsymbol{u}$ is unit norm.)

5. **Belief Propagation** [50pt]

In this problem, we will use sum-product belief propagation (BP) on Conditional Random Field to solve inference problems.

Conditional Random Fields (CRF) are an important special case of Markov Random Fields to model conditional probability distribution. They define a probability distribution p over variables $x_1, x_2, ..., x_n$ and observations $y_1, y_2, ..., y_m$, which in turn elicit an undirected graph G. The distribution has the form:

$$p(x_1, ..., x_n | y_1, ..., y_m) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c, y_c),$$

here C denotes the set of cliques (i.e., fully connected subgraphs) of G, each factor ϕ_c is a non-negative function over the variables in a clique, and Z denotes the normalizing constant that ensures the distribution p sums to 1.

For the example (Figure 3) below, the filled-in circles represent the observed nodes y_i and the empty circles represent the variables x_i. A clique in this graph is just a pair of two connected nodes, and then we have $p(x_1, ..., x_n | y_1, y_2, ..., y_n) = \frac{1}{Z} \prod_{i=2}^{n} \phi_{i-1,i}(x_{i-1}, x_i) \prod_{j=1}^{n} \phi_j(x_j, y_j)$, where $\phi_j(x_j, y_j)$ describes some statistical dependency between x_j and y_j at each position j, and $\phi_{i-1,i}(x_{i-1}, x_i)$ describes compatibility between the nearby variables x_{i-1} and x_i.

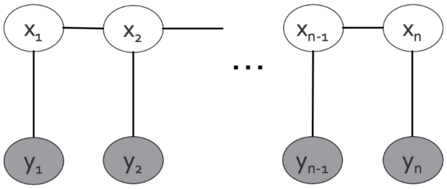


Figure 3: Example of Conditional Random Fields.

We will use $\phi_j(x_j) = \phi_j(x_j, y_j)$ since we have already observed nodes y_j, and use $\psi_{i,k}(x_i, x_k) = \phi_{i,k}(x_i, x_k)$ where k is an adjacent node of node i. Then the joint distribution over x's can be written as $p(x_1, ..., x_n | y_1, ..., y_n) = \frac{1}{Z} \prod_{i}^{n-1} \psi_{i,i+1}(x_i, x_{i+1}) \prod_{j=1}^{n} \phi_j(x_j)$.

Let's work with a graph with cycles as shown in Figure 4. Assume x and y only have two states (0 and 1) and the graphical model has 5 hidden variables, and $\underline{\text{two variables observed with } y_1 = 0,\ y_3 = 1}$. The potential functions are given in the arrays below. In $\psi_{1,2}(x_1, x_2)$, row indices are states 0, 1 for x_1, and column indices are states 0, 1 for x_2. For example, $\psi_{1,2}(x_1 = 1, x_2 = 0) = 0.4$ and $\phi_1(x_1 = 0, y_1 = 1) = 0.2$. Remember that in this case there won't be $\phi_2(x_2), \phi_4(x_4)$ since there is no observation for them.

$$\psi_{1,2}(x_1, x_2) = \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}$$

$$\phi_1(x_1, y_1) = \phi_3(x_3, y_3) = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$$

X_2	X_3	X_4	$\psi_{2,3,4}(x_2, x_3, x_4)$
0	0	0	0.1
0	0	1	0.1
0	1	0	0.2
0	1	1	0.2
1	0	0	0.1
1	0	1	0.1
1	1	0	0.1
1	1	1	0.1

We need to compute the marginal probability of x_2 for some reason. For this task, your job is to

(a) [10pt] draw the factor graph of the undirected graph (in your factor graph, each observed variable can be represented as a factor),

(b) [10pt] identify the messages that need to be computed to compute the marginal of x_2 (use the notations μ, ν to denote the different type of messages as we have seen during the lecture),

(c) [10pt] schedule the message passing algorithm (indicate in which order you will compute the messages), and

(d) [20pt] compute the marginal of x_2 by performing sum-product BP.

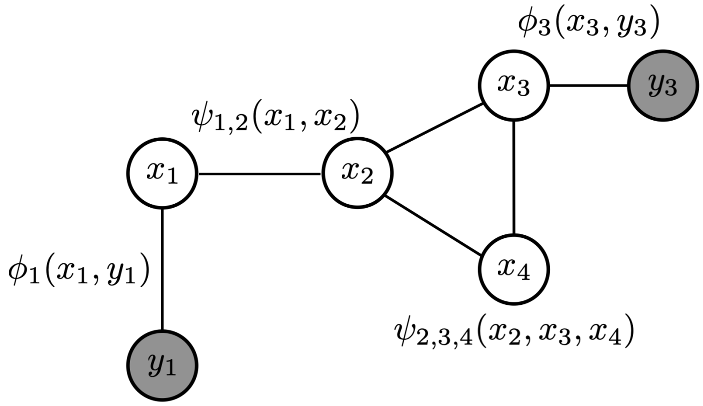


Figure 4: Undirected graph between variables.