CS 726: Advanced Machine Learning, Spring 2022, Mid-Semester exam

February 23, 2022.
06:00–8:00 pm

Roll: _____

Name: _____

Mode: Credit/Audit/Sit-through _____

Write all your answers in a separate sheet and upload on SAFE. Do not spend time/space giving irrelevant details or details not asked for. Use the marks as a guideline for the amount of time you should spend on a question. You are only allowed to refer your four-page notes, no one else's notes or textbook. You can submit the sheets in which you worked out your answers as a backup. Some background material is also present at the end of the paper.

---

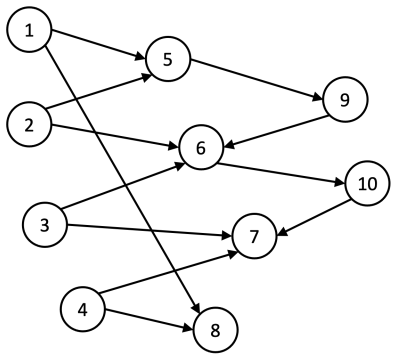
1. Assume there are n people and m virus strains. Each person i has $i-1$,$i+1$ as his friends (except person 1, n who will have one friend each). Let $y_{i,t}$ denote if a person is healthy at time t, if not which of the m strains he is infected with. Thus $y_{i,t}$ takes $m+1$ possible values with $m = 0$ denoting no infection or healthy. Once infected a person stays infected for 15 days. But after that, the person cannot be infected any further into the future, that is an infection provides life-time immunity. An uninfected person at $t-1$ can catch the infection at t with probability p if any person in his social circle is infected at $t-1$. The infection strain will be chosen uniformly from either the strains of his friends at time $t-1$ or another strain within distance 1 of any of his friend's strains. The distance between two strains s and s' is just $|s-s'|$.

   (a) To construct a *minimal and correct* Bayesian network G to represent the dependency among the $y_{i,t}$ variables, specify for each $y_{i,t}$ the set of its parents. ..2

   (b) Compactly write the expression of the probability that person i at time t is not infected ($y_{i,t} = 0$) conditioned on possible states of its parents. ..4

   (c) Write the expression for the probability that person i at time t will be infected by strain 3 given his social circle has strains 4 and 7 (assume $m = 10$) at $t-1$ and i himself is uninfected at $t-1$. ..2

2. For the BN below answer the following questions:



   (a) Provide the smallest subset of variables Z so that $x_5 \perp\!\!\!\perp x_4 | x_7, Z$ ..1

(b) Convert the network to an undirected graphical model that is minimal and correct? ..2

(c) Is the above undirected graphical model triangulated? Justify. ..2

(d) Are there more than one undirected models possible that are both correct and minimal? Justify your answer. ..2

(e) List CIs (if any) that hold in the original BN but not in the graph that you constructed. ..2

(f) Draw the junction tree for calculating $P(x_6|x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 1)$ ..2

(g) Provide the clique potentials of node of the above junction tree that contains x_5, x_9. ..2

3. Consider a parameter learning task for a CRF on n_i variables $\mathbf{y} = [y_1\ y_2\ \ldots\ y_{n_i}]$ where n_i is the length of the input $\mathbf{x}^i$ and each $y_j = +1$ or -1. Let the following 3 features be defined for $f(\mathbf{y}_c, \mathbf{x}, c)$.

$$f_1((y_j, y_{j+1}), \mathbf{x}^i, j) = [\![y_j \neq y_{j+1}]\!], (\forall 1 \leq j < n_i)$$

$$f_2(y_j, \mathbf{x}^i, j) = x^i_j y_j, (\forall 1 \leq j \leq n_i)$$

$$f_3((y_k, y_j), \mathbf{x}^i, (k, j)) = [\![y_k = y_j, x^i_k = x^i_j]\!], (\forall k < j)$$

where $[\![z]\!] = 1$ if z = true and 0 otherwise. That is, $\mathbf{f}(\mathbf{y}, \mathbf{x}) = [f_1\ f_2\ f_3]^T$. Assume the corresponding weight vector to be $\boldsymbol{\theta} = [1, -1, 1]^T$

(a) Draw the underlying graphical models corresponding to the variables for the following two instances: $\mathbf{x}^1 = [0, -1, 1, -1], \mathbf{x}^2 = [2, 0, 1, 2, -1]$ [Note $n_1 = 4, n_2 = 5$] ..3

(b) Draw the junction tree corresponding to any one of the graphs above and assign potentials to each node of your junction tree so that you can run message passing on it to find $Z = \sum_{\mathbf{y}} \mathbf{w}^T \mathbf{f}(\mathbf{x}^i, \mathbf{y})$, that is, define $\psi_c(\mathbf{y}_c, \mathbf{x}^i)$ in terms of the above quantities for each clique node c in the JT. ..3

(c) Suppose you use the junction tree above to compute the marginal probability for each pair of adjacent variables in your graph of part (a). Let $\mu_{kj}(-1, 1|\mathbf{x}^i)$, $\mu_{kj}(1, 1|\mathbf{x}^i)$, $\mu_{kj}(-1, -1|\mathbf{x}^i)$, $\mu_{kj}(1, -1|\mathbf{x}^i)$ denote the marginal probability of variable pairs y_k, y_j taking values (-1,1), (1,1), (-1,-1) and (1,-1) respectively on instance $\mathbf{x}^i$. Express the expected value of the following features in terms of the μ values on both instances.

i. f_3 ..2

ii. f_2 ..2

4. We introduced the generalized formulation for parameter learning where $P(\mathbf{y}|\mathbf{x}, \theta) = \frac{e^{F_\theta(\mathbf{x}, \mathbf{y})}}{\sum_{\mathbf{y}'} e^{F_\theta(\mathbf{x}, \mathbf{y}')}}$. Assume additionally that F_θ is an arbitrary non-linear function, example obtained from a neural network with parameter θ. Given a training set $D = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), \ldots, (\mathbf{x}^N, \mathbf{y}^N)\}$, the maximum likelihood training objective for finding θ is

$$\max_\theta \sum_i F_\theta(\mathbf{x}^i, \mathbf{y}^i) - \log \sum_{\mathbf{y}} e^{F_\theta(\mathbf{x}^i, \mathbf{y})}$$

(a) When the space of $\mathbf{y}$ is large, computing the training objective could be intractable because of the log-partition function: $\log \sum_{\mathbf{y}} e^{F_\theta(\mathbf{x}^i, \mathbf{y})}$. We wish to make the computation tractable by using the variational method to handle this term. Write the above training objective in variational form with auxillary variational variables. ..3

(b) Next, show how we can use the ideas from VAEs to rewrite the objective jointly in terms of additional 'encoder' parameters ϕ. Your training objective should be expressed jointly over θ, ϕ. ..2

(c) Say the $q_\phi(\mathbf{y}|\mathbf{x}^i)$ is represented as a Bayesian network as follows:

$$q_\phi(\mathbf{y}|\mathbf{x}^i) = q_\phi(y_1, \dots y_n|\mathbf{x}^i) = q_\phi(y_1|\mathbf{x}^i) \prod_{j=2}^{n} q_\phi(y_j|y_{j-1}, \mathbf{x}^i, j)$$

Write the expression for $\sum_{\mathbf{y}} q_\phi(\mathbf{y}|\mathbf{x}^i) \log q_\phi(\mathbf{y}|\mathbf{x}^i)$ only in terms of the potentials in the BN and the simplest possible marginals on the q_ϕ. ..4

Total: 40

Background

- A DAG G is a correct BN for a distribution $P(\mathbf{x})$ if all Local-CIs that are implied in G hold in P.
- A DAG G is minimal if we cannot remove any edge(s) from G and still get a correct BN for P.
- A BN which has no immorality will not require any new edges to be added when converting to MRF. Such networks will have a perfect MRF.
- D-separation test for global-CIS: Three sets of variables X, Y, Z. If Z d-separates X from Y in BN then, $X \perp\!\!\!\perp Y|Z$.
  In a directed graph H, Z d-separates X from Y if all paths P from any X to Y is blocked by Z.
  A path P is blocked by Z when
  1. $x_1 \rightleftharpoons x_2 \ldots \rightarrow x_i \rightarrow \ldots x_{k-1} \rightleftharpoons x_k$ and $x_i \in Z$
  2. $x_1 \rightleftharpoons x_2 \ldots \leftarrow x_i \leftarrow \ldots x_{k-1} \rightleftharpoons x_k$ and $x_i \in Z$
  3. $x_1 \rightleftharpoons x_2 \ldots \leftarrow x_i \rightarrow \ldots x_{k-1} \rightleftharpoons x_k$ and $x_i \in Z$
  4. $x_1 \rightleftharpoons x_2 \ldots \rightarrow x_i \leftarrow \ldots x_{k-1} \rightleftharpoons x_k$ and $x_i \notin Z$ and $Desc(x_i) \notin Z$
- Converting BN to MRFs Efficient: Using the Markov Blanket (MB) (also called the Local-CI) algorithm . The MB of a x_i in a BN can be shown to be:

  $$MB(x_i) = Pa(x_i) \cup Ch(x_i) \cup \text{Spouse}(x_i)$$

  This is essentially obtained by moralizing a BN and removing all directed edges.
- **Chordal or triangulated graphs** A graph is chordal if it has no minimal cycle of length ≥ 4.
- Variational Rewrite

  $$\begin{aligned}
  &\max_{\theta} \sum_{i=1}^{N} \log \sum_{\mathbf{z}: z_1, \ldots, z_m} P(\mathbf{y}^i, \mathbf{z}|\theta, \mathbf{x}^i) \\
  &\equiv \max_{\theta} \sum_{i=1}^{N} \max_{q_{i,\mathbf{z}}: \sum_{\mathbf{z}} q_{i,\mathbf{z}}=1} \sum_{\mathbf{z}} q_{i,\mathbf{z}} \log P(\mathbf{y}^i, \mathbf{z}|\theta, \mathbf{x}^i) - \sum_{\mathbf{z}} q_{i,\mathbf{z}} \log q_{i,\mathbf{z}} \\
  &s.t. \sum_{\mathbf{z}} q_{i,\mathbf{z}} = 1 \text{ and } q_{i,\mathbf{z}} \geq 0
  \end{aligned}$$

- Training CRFs with linear features: Expected value of features: $\sum_i E_{\Pr(\mathbf{y}'|\boldsymbol{\theta}, \mathbf{x}^i)} f_k(\mathbf{x}^i, \mathbf{y}') = \sum_i \sum_c \sum_{\mathbf{y}'_c} f_k(\mathbf{x}^i, \mathbf{y}'_c, c) \Pr(\mathbf{y}'_c|\boldsymbol{\theta}, \mathbf{x}^i) = \sum_i \sum_c \sum_{\mathbf{y}'_c} f_k(\mathbf{x}^i, \mathbf{y}'_c, c) \mu(\mathbf{y}'_c|\boldsymbol{\theta}, \mathbf{x}^i)$