10708 Probabilistic Graphical Models: Final Exam

Due Dec 15th by 2pm electronically to 10708-instr@cs.cmu.edu or paper version to Monica Hopes,

or by fax to 412-268-3431

Your final must be done individually. You may not discuss the questions with anyone other than Carlos or the TAs (you are free to ask us questions by e-mail or in person if you are having problems with a question). The exam is open book, but not open-Google, *i.e.*, you can use any materials we discussed in class or linked to from the class website. You are not allowed to look at other sources. However, you may use a calculator or Matlab to do numerical computations, if necessary. If you hand in your assignment early, you can get bonus points.

HANDIN	BONUS
Dec 11, 2pm	4 pts
Dec 12, 2pm	3 pts
Dec 13, 2pm	2 pts
Dec 14, 2pm	1 pts

You may *not* use late days on the final.

1 Short answer [9 pts] [Ajit]

1. Using Jensen's inequality show that, for discrete variable X,

   (a) $H_P(X) \geq 0$

   (b) $D(P||Q) \geq 0$

   Theorem 1 (Jensen's Inequality) *Let f be a concave function and P a distribution over a random variable X. Then $E_P[f(X)] \leq f(E_P[X])$.*

   (Hint: intuitively, $H_p(X) = E_P[-log(p(X))]$.)

2. You want to learn the structure of a Gaussian graphical model using a score-based method. If the parameter prior $P(\theta_{\mathcal{G}}|\mathcal{G})$ is Gaussian explain why using the Bayesian score is preferable to BIC.

3. This problem is designed to make you comfortable with the canonical parameterization of Gaussian distributions. (Notation: Throughout this exam, we will refer to Gaussians in standard form as $N(\cdot; \mu, \Sigma)$ and Gaussians in canonical form as $N_c(\cdot; \eta, \Lambda)$. When we ask you to give us a Gaussian distribution, we want you to write it using this notation.) You are given the joint distribution over X and Y in standard form:

$$P(X, Y) = N\big(X, Y; \mu = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.75 \end{bmatrix}\big)$$

   (a) Write down $P(X, Y)$ in canonical form.

   (b) Write down $P(Y)$ in canonical form.

   (c) Using your answers from parts (a) and (b), write down $P(X|Y)$ in canonical form.

   (Hint 1: To multiply two Gaussians in canonical form, you simply add the parameters, and to divide two Gaussians in canonical form, you subtract the parameters, filling in with zeros as necessary. Refer to the Kalman filter slides for more details.)

   (Hint 2: Your answer to part (c) in canonical form will be represented as a multivariate distribution over X and Y, not a univariate distribution over X as would be the case in standard form.)

2 Context-Specific Independence [13 pts] [Khalid]

Let Y be a binary-valued random variable with n binary-valued random variables $X_1, \ldots, X_n$ as its parents. Each X_i in turn has no parents. The variables are all binary and take values in $\{0, 1\}$.
$X_i \sim \text{Bernoulli}(\theta_i), i = 1, \ldots, n$; while Y is a determinisitic OR function of the X_i's.

$$P(Y = \text{OR}(X_1, \ldots, X_n)) = 1$$
$$P(Y = 1 - \text{OR}(X_1, \ldots, X_n)) = 0$$

2.1

What is the time complexity of naively computing the marginal probability $P(Y = 1)$ using the standard tabular variable elimination? (Hint: if you use a tabular representation of $P(Y \mid X_1, \ldots, X_n)$, what is the size of the table?)

2.2

Note that the complexity in the previous question arises because node Y has n parents. Can you introduce some intermediate variables so that no node in the new graph has more than two parents? What is the complexity of computing the marginal probability $P(Y = 1)$ using the standard variable elimination algorithm in this new (equivalent) graph?

2.3

Now, consider a noisy version of the above problem (this is usually called a noisy-OR CPT), where in addition to $X_1, \ldots, X_n, Y$, we have "noisy" variables $X'_1, \ldots, X'_n, Y'$. As before, each X_i has no parent; each X'_i has X_i as a parent; Y has $X'_1, \ldots, X'_n$ as its parents, and Y' has Y as its parent.

$$\begin{aligned} X_i &\sim \text{Bernoulli}(\theta_i) \\ P(X'_i = 0|X_i) &= (1 - \lambda_i)^{X_i} \\ Y &= \text{OR}(X'_1, \ldots, X'_n) \\ P(Y' = 0|Y) &= (1 - \lambda)^{Y} \end{aligned}$$

Thus, the variables X'_i are the noisy versions of X_i; Y is a deterministic OR function of X'_i, and Y' is a noisy version of Y.
Use the intuition of the previous parts and show how to compute the marginal probability of $P(Y' = 1)$ using variable elimination but with a low time complexity.

3 Structure Learning in Undirected Models [13 pts] [Khalid]

For this problem, assume that you have i.i.d. data sampled from a distribution $P(\mathcal{X})$. P is represented by a Markov Random Field whose graph structure is unknown. However, you do know that each node has at most d neighbors.

3.1

Show why knowing the Markov blanket of each node is sufficient for determining the graph structure.

3.2

For any node X and its Markov blanket $\text{MB}(X)$, we know that

$$P \models (X \perp \mathcal{X} - \{X\} - \text{MB}(X)|\text{MB}(X)).$$

Briefly, why might you need *a lot* of data to test for this conditional independence directly?

3.3

For disjoint sets of variables $\mathbf{A}$ and $\mathbf{B}$, let conditional entropy be defined as,

$$\mathrm{H}(\mathbf{A}|\mathbf{B}) = -\sum_{\mathbf{a},\mathbf{b}} P(\mathbf{A}=\mathbf{a},\mathbf{B}=\mathbf{b}) \log P(\mathbf{A}=\mathbf{a}|\mathbf{B}=\mathbf{b})$$

Prove that for any node X, $\mathrm{H}(X|\mathrm{MB}(X)) = \mathrm{H}(X|\mathcal{X} - \{X\})$.

3.4

For disjoint sets of variables $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, we have that

$$\mathrm{H}(\mathbf{A}|\mathbf{B},\mathbf{C}) \leq \mathrm{H}(\mathbf{A}|\mathbf{B}).$$

In other words, information never hurts. Prove that $\mathrm{MB}(X) = \mathrm{argmin}_{\mathbf{Y}}\, \mathrm{H}(X|\mathbf{Y})$.

3.5

Using the intuition developed in the previous parts, describe a structure learning algorithm for Markov Random Fields, assuming the constraint that each node has at most d neighbors. Your algorithm should run in $O(n\binom{n}{d}c)$ time, where n is the number of nodes in your model, and c is the complexity of computing the conditional entropy $\mathrm{H}(X|\mathbf{Y})$, when $|\mathbf{Y}| \leq d$.

3.6

If we removed the constraint that each node have at most d neighbors and instead changed our optimization problem to include a penalty term, $\mathrm{MB}(X) = \mathrm{argmin}_{\mathbf{Y}}\{\mathrm{H}(X|\mathbf{Y}) + |Y|\}$, how would the time complexity of the algorithm change?

4 KL Projection in Assumed Density Filtering [13 pts]

[Khalid]

In class, we discussed the Boyen-Koller algorithm, an instance of assumed density filtering, where the belief state is represented by a clique tree. At each time step, the belief state

becomes more complex, and we project it into a simpler clique tree by doing a simple marginalization. This approach may seem like a hack, but, in this question, you will show that this marginalization is a well-defined, KL-minimizing projection.
Consider the clique tree T_1 (corresponding to the complex belief state in Boyen-Koller):

$$ABC - BCD - CDE$$

Let the cliques be calibrated, so that we have all the clique marginal probabilities.
Consider also the clique tree T_2 (corresponding to the simpler belief state in BK):

$$AB - BC - CD - DE$$

A KL projection of a distribution P over a set of distributions S is given by,

$$P_S = \arg\min_{Q \in S} KL(P||Q)$$

Let P denote the distribution given by the calibrated clique tree T_1; and let S denote the set of distributions represented by clique tree T_2 (i.e., for which T_2 is an I-map). Show that the KL Projection of P over S is given by setting the clique probabilities in T_2 to be the marginals of the corresponding clique probabilities in T_1. That is, $P_{T_2}(AB) = \sum_C P_{T_1}(ABC)$, and so on.

5 I-Equivalence [13 pts] [Khalid]

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two graphs over $\mathcal{X}$. For this problem, you will prove that $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton and the same set of immoralities if and only if they are I-equivalent.

Definition 1 (Minimal Active Trail) *Consider an active trail $T = X_1, X_2, \ldots, X_m$. We call this active trail* minimal *if no subset of the nodes in T forms an active trail between X_1 and X_m. In other words, T is minimal if no other active trail between X_1 and X_m "shortcuts" any of the nodes in T.*

Definition 2 (Triangle) *Consider a trail $T = X_1, X_2, \ldots, X_m$. We call any three consecutive nodes in the trail a* triangle *if their undirected skeleton is fully connected (i.e., forms a 3-clique). In other words, X_{i-1}, X_i, X_{i+1} form a triangle if we have $X_{i-1} \rightleftharpoons X_i \rightleftharpoons X_{i+1}$ and $X_{i-1} \rightleftharpoons X_{i+1}$.*

5.1

Give an example of two I-equivalent graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ that have the same skeleton, but different v-structures.

5.2

Prove that the only possible triangle in a minimal active trail is one where $X_{i-1} \leftarrow X_i \rightarrow X_{i+1}$, with an edge between X_{i-1} and X_{i+1}, and where either X_{i-1} or X_{i+1} is the center of a v-structure in the trail. (Hint: prove by cases.)

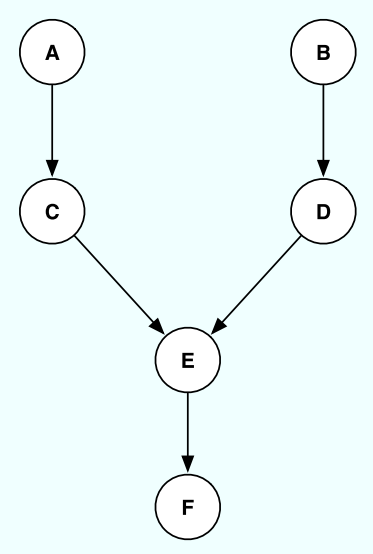
5.3

Consider two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ that have the same skeleton and same immoralities. Prove, using the notion of minimal active trail, that $\mathcal{G}_1$ and $\mathcal{G}_2$ imply precisely the same conditional independence assumptions, i.e., that if X and Y are d-separated given $\mathbf{Z}$ in $\mathcal{G}_1$, then X and Y are also d-separated given $\mathbf{Z}$ in $\mathcal{G}_2$. (Hint: prove by contradiction.)

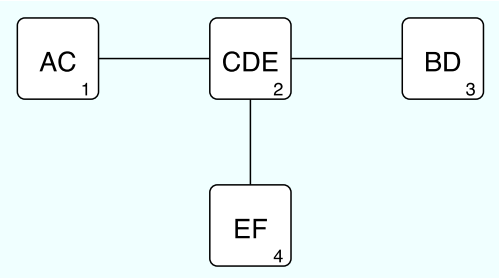
5.4

Finally, prove that two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ that induce the same conditional independence assumptions must have the same skeleton and the same immoralities. (Hint: prove by contradiction.)

6 Gaussian Graphical Models [13 pts] [Khalid]



(a) Gaussian Graphical Model



(b) Junction Tree

In Figure 1(a) we give you a Gaussian graphical model with the following conditional probability distributions (all given in canonical form):

$$
\begin{aligned}
P(A) &= N_c\big(A; \eta = 9, \Lambda = 1\big) \\
P(B) &= N_c\big(B; \eta = 1, \Lambda = 0.6\big) \\
P(C|A) &= N_c\big(C, A; \eta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Lambda = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\big) \\
P(D|B) &= N_c\big(D, B; \eta = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}, \Lambda = \begin{bmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}\big) \\
P(E|C,D) &= N_c\big(E, C, D; \eta = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Lambda = \begin{bmatrix} 1 & -9 & 0.5 \\ -9 & 81 & -4.5 \\ 0.5 & -4.5 & 0.25 \end{bmatrix}\big) \\
P(F|E) &= N_c\big(F, E; \eta = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \Lambda = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\big)
\end{aligned}
$$

In this problem, you will use the Shafer-Shenoy message passing scheme for Gaussians to perform exact inference on this model.

6.1

Give an elimination ordering for the Bayesian network in Figure 1(a) that would result in the junction tree in Figure 1(b).

6.2

Using the Family Preserving Property, assign the given CPDs to appropriate cliques in the junction tree. Then, remembering how to multiply Gaussians in canonical form from question 1.3, compute the initial clique potentials $\Pi_1^{(0)}$, $\Pi_2^{(0)}$, $\Pi_3^{(0)}$, and $\Pi_4^{(0)}$. Your answers should be Gaussian distributions in canonical form.

6.3

Compute $P(C, D, E)$ using Shafer-Shenoy message passing. Write down the three messages that were needed to compute this probability. Both your final answer and your messages should be represented as Gaussian distributions in canonical form.

(Hint: The Shafer-Shenoy algorithm for Gaussian graphical models is analogous to the algorithm presented in class for discrete models. However, you will need to take into account how to multiply potentials together and how to marginalize out variables in the canonical

Gaussian setting. See the Kalman filter slides, as well as problem 1.3 of this exam, for more details.)

6.4

Given the messages computed for 6.3, what additional message would you need if you wanted to compute $P(A|C)$? Write down this message as a Gaussian distribution in canonical form. (Note: you do not need to compute $P(A|C)$.)

6.5

Given that a minimal junction tree for a Bayesian network with n nodes can have at most n cliques, what is the time complexity of Shafer-Shenoy on a Gaussian graphical model with n nodes and induced tree width w? Briefly justify your answer in one or two sentences. (Hint: Time complexity of matrix inversion for a $k \times k$ matrix is $O(k^3)$.) What is the running time of a discrete model over the same BN structure, where each variable takes on at most c values?

6.6 [Extra Credit] [3 pts]

If we wanted to compute $P(C, D, E)$, an alternative to message passing would have been to multiply together all the CPDs, form a single matrix for the distribution $P(A, B, C, D, E, F)$ and directly marginalize out all the other variables. What is the time complexity of this operation? Briefly justify your answer in one or two sentences.

7 Tree-augmented Naïve Bayes [13 pts] [Ajit]

You are given R complete records over discrete variables (features) $X_1, \ldots, X_n$ and over a discrete class C. One approach to modeling the class-conditional distribution $P(C|X_1, \ldots, X_n)$ is tree-augmented naïve Bayes (TAN), which finds the optimal spanning tree on the features and induces a directed tree by picking an arbitrary root

1. Given infinite data $R \to \infty$ why does the choice of directed tree not matter ?
2. Fix the spanning tree on features. Let θ^* be the parameters learned using infinite data. Let θ be the parameters learned using finite data. We want to minimize the variance of our parameters θ. Explain why the choice of directed tree matters. (Hint: What part of the dataset is used to learn each parameter of a CPT? The more data you have, the lower you expect the variance over the parameter to be.)

3. Given the spanning tree on features, provide pseudocode for an algorithm that picks the directed tree that maximizes a decomposable score in $O(n^2c)$ time, where c is the cost of computing the score of a node given its parents.

4. *(Extra Credit [5 pts])* Given the spanning tree on features, provide pseudocode for an dynamic programming algorithm that picks the directed *poly-tree* that maximizes a decomposable score. A poly-tree is a directed acyclic graph where each node can have more than one parent, and where it's skeleton has no undirected cycles. What is the complexity of your algorithm in terms of n ?

8 Variational Free Energy [13 pts] [Ajit]

Once upon a time there were three bears, a mother bear, a father bear, and a baby bear. The mother was a frequentist; the father a pragmatic Bayesian. However, the baby bear thought that his parents were terribly silly – his mother prone to knitting sweaters that were far too snug, that overfit; his father far too stern in his insistence that he choose a single sweater to wear each day. The baby bear preferred to think of himself as wearing a distribution over sweaters, making him the only proper Bayesian bear in the whole forest.

(interlude) Raindrops on roses, whiskers on kittens, bright copper kettles, warm woolen mittens, brown paper packages tied up with string, along with parameter estimation, these are a few of bears' favorite things *(end interlude).*

Knowing the bears' fondness for parameter estimation, a little blonde girl comes along to steal their work. Finding no one home, she looks at the desks of each bear and finds,

Definition 3 (Maximum Likelihood) *If y are the observed variables and θ the model parameters then the maximum likelihood criterion is*

$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} \log p(y|\theta)$$

Definition 4 (Maximum a Posteriori) *If y are the observed variables and θ the model parameters then the maximum a posteriori criterion is*

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_{\theta} \log p(\theta|y)$$

Definition 5 (Fully Bayesian) *If y are the observed variables and θ the model parameters then the fully Bayesian criterion demands the full posterior*

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

8.1 EM-ML

It is well known to any frequentist bear that, denoting the latent variables z and introducing any distribution over latent variables, $q(z)$,

$$\begin{aligned}\log p(y|\theta) &\geq \sum_z q(z) \log \frac{p(y,z|\theta)}{q(z)} \\ &= E_{q(z)}[\log p(y,z|\theta)] + H[q(z)] \equiv F(q,\theta).\end{aligned}$$

The EM algorithm consists of maximizing a lower bound on $p(y|\theta)$ by iterating the following steps over time t:

$$\begin{aligned}\textbf{E-step:} \quad & q^{(t+1)} = \operatorname*{argmax}_q F(q, \theta^{(t)}) \\ \textbf{M-step:} \quad & \theta^{(t+1)} = \operatorname*{argmax}_\theta F(q^{(t+1)}, \theta)\end{aligned}$$

1. In the E-step, prove that $q^{(t+1)} = p(z|y, \theta^{(t)})$.

2. In class it was explained that the M-step updated parameters using expected counts. Show that in the E-step expected counts are computed using

$$E_{q(z)}[\text{Count}(\mathbf{A}_O = \mathbf{a}_O, \mathbf{A}_H = \mathbf{a}_H)] = \sum_{j=1}^{R} \mathbf{1}(\mathbf{A}_O^{(j)} = \mathbf{a}_O) P(\mathbf{A}_H = \mathbf{a}_H | O^{(j)}, \theta^{(t)})$$

   where $O^{(j)}$ is the j^{th} record in the data set, $\mathbf{A}_O$ are the observed variables, $\mathbf{A}_H$ the unobserved (latent) variables, and $\theta^{(t)}$ the estimate of the Bayesian network parameters at step t of the EM algorithm. $\mathbf{1}(\mathbf{A}_O^{(j)} = \mathbf{a}_O)$ is an indicator function for whether variables $\mathbf{A}_O$ take on value $\mathbf{a}_O$ is record j.

3. You are given two inference routines, one for variable elimination and another for junction trees. Which routine is more appropriate for the E-step in part 2 ? Briefly explain you answer.

8.2 EM-MAP

The father bear has scrawled down the following equation:

$$\log p(\theta|y) \geq E_{q(z)}[\log p(y,z|\theta)] + H[q(z)] + \log p(\theta) \equiv F(q,\theta) \tag{1}$$

where $q(z)$ is some distribution over unobserved variables z, which is typically called the variational free distribution, and $p(\theta)$ is a parameter prior. The EM algorithm consists of maximizing a lower bound on $p(\theta|y)$.

$$\begin{aligned}\textbf{E-step:} \quad & q^{(t+1)} = \operatorname*{argmax}_q F(q, \theta^{(t)}) \\ \textbf{M-step:} \quad & \theta^{(t+1)} = \operatorname*{argmax}_\theta F(q^{(t+1)}, \theta)\end{aligned}$$

1. Prove equation 1.

8.3 Fully Bayesian

The bears come home to find the little blonde girl rummaging through their desks. Regular bears would simply maul her. However, since these are not regular bears they chain the girl up and give her choice: answer the following questions or be eaten alive.

1. Briefly explain why computing $p(y)$ exactly is difficult ?
2. Assuming some free distribution that factors over latent variables and parameters, $q(z, \theta) = q(z)q(\theta)$, prove that

$$\log p(y) \geq \int q(\theta) \ln \frac{p(\theta)}{q(\theta)} \, d\theta + \int q(\theta) \sum_z q(z) \ln \frac{p(y, z|\theta)}{q(z)} \, d\theta \equiv F(q(z), q(\theta))$$

3. Prove that maximizing $F(q(z), q(\theta))$ corresponds to minimizing the KL-divergence $D(q(z, \theta)||p(z, \theta|y))$.

We can estimate the marginal likelihood $p(y)$ by maximizing the lower bound $F(q(z), q(\theta))$ with the following EM-style algorithm (which you do not need to prove):

$$\textbf{E-step:} \quad q^{(t+1)}(z) \propto \exp \int q^{(t)}(\theta) \ln p(y, z|\theta) \, d\theta$$

$$\textbf{M-step:} \quad q^{(t+1)}(\theta) \propto p(\theta) \exp \int \ln p(y, z|\theta) q^{(t+1)}(z) \, dz$$

When asked to derive this, the little blonde girl decided that she'd rather be eaten alive. The end.

9 Feedback [0 pts]

The following are questions that we use to calibrate the exam in future years. Your answers are appreciated.

1. How many hours did it take to complete the exam ?
2. Which two questions did you find hardest ?
3. Which two questions did you find easiest ?