INFERENCE AND REPRESENTATION REVIEW PROBLEMS

Source: books from class bibliography (Murphy, Koller, MacKay, Bishop), exams and homework from previous instance of the class, and Sam Roweis' machine learning class materials.

Problems are not a comprehensive list. Make sure to study the reading materials related with the classes and recitations. Problems marked with ($\star$) may be harder than midterm problems.

(1) **Bayes theorem, likelihood, priors, posteriors** I have two coins in my pocket, one is fair and one is biased with $P(tails) = 1/3$, $P(heads) = 2/3$. I take a random coin from my pocket and I toss it 4 times obtaining: heads, tails, heads, heads. The goal is to infer what coin I selected. (Write down the computations but you don't need compute the numerical result. Justify your answers.)
- What is is the likelihood function?
- What is the prior?
- What is the posterior probability?
- What is the maximum likelihood estimator? What is the maximum a posteriori estimator? Do you expect them to coincide?

(2) In a game, two coins are tossed. If either of the coins comes up heads, you have won a prize. To claim the prize, you must point to one of your coins that is a head and say 'look, that coin's a head, I've won'. You watch Fred play the game. He tosses the two coins, and he points to a coin and says 'look, that coin's a head, I've won'. What is the probability that the other coin is a head?

(3) **Conditional independence** Are the following statements **true or false** for every random variables X, Y, Z, W? If true prove it, if false give a counter example.
- $X \perp Y|Z$ implies $X \perp Y$.
- $X \perp Y$ and $Y \perp Z$ implies $X \perp Z$.
- $X \perp Y \ |Z, W$ and $X \perp W|Z, Y$ implies $X \perp Y|Z$.
- If a set of random variables is pairwise independent then they are jointly independent.

(4) **Bayesian networks** Why do Bayesian networks need to be acyclic? Show with an example that Bayesian networks with cycles can lead to a contradiction.

(5) From Murphy:

Exercise 2.5 The Monty Hall problem

(Source: Mackay.) On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors. Hint: use Bayes rule.

Exercise 2.6 Conditional independence

(Source: Koller.)

a. Let $H \in \{1, \ldots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), \ldots, P(H = K|e_1, e_2))$$

Which of the following sets of numbers are sufficient for the calculation?

i. $P(e_1, e_2)$, $P(H)$, $P(e_1|H)$, $P(e_2|H)$
ii. $P(e_1, e_2)$, $P(H)$, $P(e_1, e_2|H)$
iii. $P(e_1|H)$, $P(e_2|H)$, $P(H)$

b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficent now?

Show your calculations as well as giving the final result. Hint: use Bayes rule.

Exercise 2.7 Pairwise independence does not imply mutual independence

We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2.125}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2.126}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \quad \forall S \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{2.127}$$

and hence

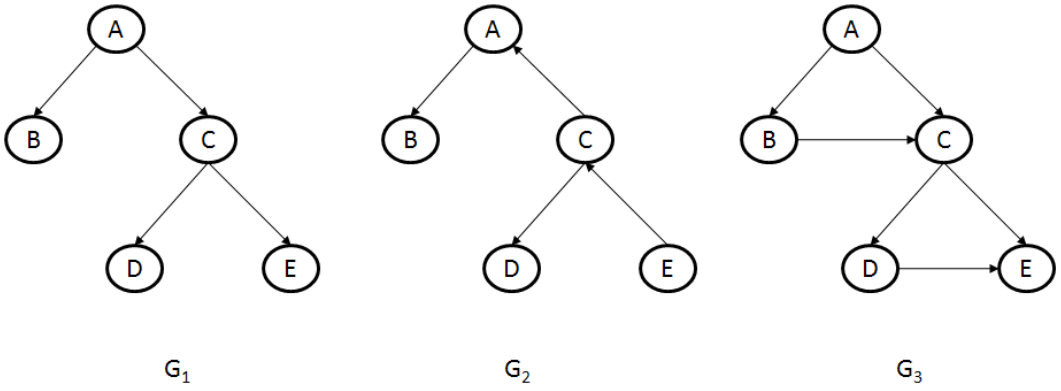
$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{2.128}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

(6) From previous exam:

Consider the Bayesian networks shown in Figure 1. For each of the following questions, answer *true* or *false*, **justifying your answer in 1–3 sentences**:

Figure 1:



(a) G_1 is I-equivalent to G_2.

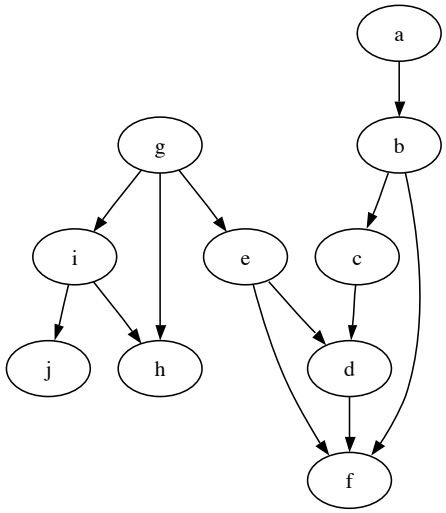
(b) G_1 is I-equivalent to G_3.

(c) If G_1 is a perfect map for $\mathcal{P}$, then G_3 is an I-map for $\mathcal{P}$.

(7) From previous exam

Consider the Bayesian network shown in Figure 2. Which of the following conditional independence statements hold? Answer *true* or *false* – **no justification needed**.

Figure 2:



(a) $a \perp f$

(b) $a \perp g$

(c) $b \perp i \mid f$

(d) $d \perp j \mid g, h$

(e) $i \perp b \mid h$

(f) $j \perp d$

(g) What is the Markov blanket of variable e?

(8) **Hidden Markov Models** Harry lives a simple life. Some days he is Angry and some days he is Happy. But he hides his emotional state, and so all we can observe is whether he smiles, frowns, laughs, or yells. Harry's best friend is utterly confused about whether Harry is actually happy or angry and decides to model his emotional state using a hidden Markov model. Let $X_d \in \{\text{Happy, Angry}\}$ denote Harry's emotional state on day d, and let $Y_d \in \{\text{smile, frown, laugh, yell}\}$ denote the observation made about Harry on day d. Assume that on day 1 Harry is in the Happy state, i.e. X_1 =Happy. Furthermore, assume that Harry transitions between states exactly once per day (staying in the same state is an option) according to the following distribution:

$p(X_{d+1} = \text{Happy}|X_d = \text{Angry}) = 0.1$,
$p(X_{d+1} = \text{Angry}|X_d = \text{Happy}) = 0.1$,
$p(X_{d+1} = \text{Angry}|X_d = \text{Angry}) = 0.9$, and
$p(X_{d+1} = \text{Happy}|X_d = \text{Happy}) = 0.9$.

The observation distribution for Harry's Happy state is given by

$p(Y_d = \text{smile}|Xd = \text{Happy}) = 0.6$,
$p(Y_d = \text{frown}|Xd = \text{Happy}) = 0.1$,
$p(Y_d = \text{laugh}|Xd = \text{Happy}) = 0.2$, and
$p(Y_d = \text{yell}|Xd = \text{Happy}) = 0.1$.

The observation distribution for Harry's Angry state is

$p(Yd = \text{smile}|Xd = \text{Angry}) = 0.1$,
$p(Yd = \text{frown}|Xd = \text{Angry}) = 0.6$,
$p(Yd = laugh|Xd = \text{Angry}) = 0.1$, and
$p(Yd = \text{yell}|Xd = \text{Angry}) = 0.2$.

(a) Model this as an HMM (Draw the corresponding BN).
(b) What is $p(X_2 = \text{Happy})$?
(c) What is $p(Y_2 = \text{frown})$?
(d) What is $p(X_2 = \text{Happy}|Y2 = \text{frown})$?
(e) What is $p(Y_{60} = \text{yell})$? What assumption did you use?
(f) Explain the use of the forward algorithm in the context of this example. What question would it be useful for?
(g) Assume that $Y_1 = Y_2 = Y_3 = Y_4 = Y_5 =$ frown. What is the most likely sequence of the states? That is, compute the MAP assignment

$$\arg \max_{x1,...,x5} p(X_1 = x_1, ..., X_5 = x_5|Y_1 = Y_2 = Y_3 = Y_4 = Y_5 = \text{frown})$$

(9) State the input, output and assumptions of the Baum-Welch algorithm. Give a general explanation of what it does (feel free to use an example).

(10) From Sam Roweis':

In this question, you'll derive for yourself the maximum likelihood estimates for class-conditional Gaussians with independent features (diagonal covariance matrices). Start with the following generative model for a discrete class label $y \in (1, 2, \ldots, K)$ and a real valued vector of D features $\mathbf{x} = (x_1, x_2, \ldots, x_D)$:

$$p(y = k) = \alpha_k$$
$$p(\mathbf{x}|y = k) = \left(\prod_{i=1}^{D} 2\pi\sigma_i^2 \right)^{-1/2} \exp\left\{ -\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2 \right\}$$

where α_k is the prior on class k, σ_i^2 are the shared variances for each feature (in all classes), and μ_{ki} is the mean of the feature i conditioned on class k.

- Use Bayes' rule $p(a|b) = p(b|a)p(a)/p(b)$ to invert the model above and write the expression for $p(y = k|\mathbf{x})$. [Hint: remember that $p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|y = k)\alpha_k$.]

- Write down the expression for the likelihood function $\ell(\theta; \mathcal{D}) = \log p(y^1, x^1, y^2, x^2, \ldots, y^M, x^M|\theta)$ of a particular dataset $\mathcal{D} = \{y^1, x^1, y^2, x^2, \ldots, y^M, x^M\}$ with parameters $\theta = \{\alpha, \mu, \sigma^2\}$. (Assume the data are iid.)

- Take partial derivatives of the likelihood with respect to each of the parameters μ_{ki}, with respect to the shared variances σ_i^2, and with respect to the class priors α_k. Since the variances must be positive you might want to take the derivative with respect to their logarithms.

- Set these partial derivatives to zero and solve for the maximum likelihood parameter values μ_{ki}, σ_i^2 and α_k. When solving for the class priors, remember that α_k, must be between 0 and 1 and sum to unity (across k), so you need to use Lagrange multipliers to enforce this constraint.

(11) **EM** Express a mixtures of Gaussians model as a Bayesian Network. Explain the EM algorithm in this context. How is it different from Lloyd's algorithm for k-means?

(12) **Naive Bayes**

(a) Explain the naive bayes model. Explain the assumptions and write it as a Bayesian Network.

(b) Consider a Naive Bayes model (multivariate Bernoulli version) for spam classification with the vocabulary V="secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza". We have the following example spam messages "million dollar offer", "secret offer today", "secret is secret" and normal messages, "low price for valued customer", "play secret sports today", "sports is healthy", "low price pizza". Show how to compute the MLEs for the following parameters: $\theta_{spam}, \theta_{secret|spam}, \theta_{secret|non-spam}, \theta_{sports|non-spam}, \theta_{dollar|spam}$.

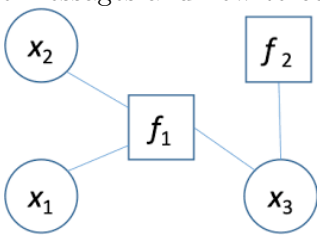
(13) **Markov Chains**

(a) What is a stationary distribution of a Markov Chain?

(b) Does it always exist? Is it always unique?

(c) How is it used in the Google PageRank algorithm? Illustrate with an example.

(14) **Belief propagation** Consider the factor graph below where $X_1 \in \{0, 1\}, X_2 \in \{2, 3\}, X_3 \in \{0, 1, 2\}$. Write the message updates $m_{x_3 \to f_1}$ and $m_{f_1 \to x_2}$. Clearly state the dimension of the messages and how to compute its entries.



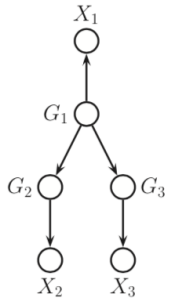
Explain how to compute the marginals.

(15) From Homework 3:

Message passing on a tree. (From Murphy chapter 20)

Consider the DGM below which represents the following fictitious biological model. Each G_i represents the genotype of a person: $G_i = 1$ if they have a healthy gene and $G_i = 2$ if they have an unhealthy gene. G_2 and G_3 may inherit the unhealthy gene from their parent G_1. $X_i \in \mathbb{R}$ is a continuous measure of blood pressure, which is low if you are healthy and high if you are unhealthy. We define the CPDs as follows

$$
\begin{aligned}
p(G_1) &= [0.5, 0.5] \\
p(G_2|G_1) &= \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \\
p(G_3|G_1) &= \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \\
p(X_i|G_i = 1) &= \mathcal{N}(X_i|\mu = 50, \sigma^2 = 10) \\
p(X_i|G_i = 2) &= \mathcal{N}(X_i|\mu = 60, \sigma^2 = 10)
\end{aligned}
$$



(a) Suppose you observe $X_2 = 50$ and X_1 is unobserved. What is the posterior belief on G_1 (i.e. $p(G_1|X_2 = 50)$?)

(b) Now suppose you observe $X_2 = 50$ and $X_3 = 50$. What is $p(G_1|X_2, X_3)$? Explain your answer intuitively.

(c) Now suppose $X_2 = 60$ and $X_3 = 60$. What is $p(G_1|X_2, X_3)$? Explain your answer intuitively.

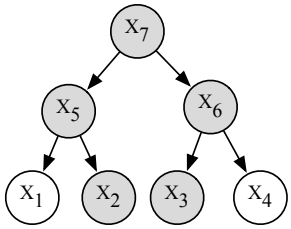
(d) Now suppose $X_2 = 50$ and $X_3 = 60$. What is $p(G_1|X_2, X_3)$? Explain your answer intuitively.

(16) **Sampling** Write two algorithms for sampling from the distribution with density $p(y) = \lambda \exp(-\lambda y)$.

(17) By making use of the sampler from the previous exercise for sampling from a single exponential distribution, devise an algorithm for sampling from the piecewise exponential distribution defined by

$$q(z) = k_i \lambda_i \exp(-\lambda_i (z - z_{i-1})) \quad z_{i-1} < z \leq z_i.$$

(18) ($\star$) Consider the Bayesian network shown below whose parameters we would like to learn from data $\mathcal{D}$. However, for each data point $\mathbf{x} \in \mathcal{D}$, some of the bottom layer variables are unobserved (X_5, X_6 and X_7 are always observed). For example, the figure shows an observation $(?, x_2, x_3, ?, x_5, x_6, x_7)$ where X_1 and X_4 are unobserved. Give a closed-form formula for the maximum likelihood estimate of the parameters, and explain why it is correct. You may assume $X_i \in \{0, 1\}$ for all i.



(19) ($\star$) Suppose that you had a black-box inference engine that could take as input a pairwise Markov random field and give as output the single-node marginal probabilities. Could you use this for gradient-ascent based maximum likelihood learning of the parameters? If not, describe how to modify the black-box's output so that it would be helpful for learning. Be sure to state how the black-box is used within learning and why.

(20) Consider the problem of maximum likelihood learning of a pairwise Markov random field on three binary variables, $\Pr(x_1, x_2, x_3) = \frac{1}{Z} \exp\Big(\theta_{12}(x_1, x_2) + \theta_{23}(x_2, x_3) + \theta_{13}(x_1, x_3)\Big)$, from the following data:

X_1	X_2	X_3
0	0	1
1	0	1
1	0	1
0	0	0

Let θ^{ML} be the solution to the maximum likelihood optimization problem. What are the marginals $\Pr(x_1; \theta^{\mathrm{ML}})$, $\Pr(x_2; \theta^{\mathrm{ML}})$, $\Pr(x_3; \theta^{\mathrm{ML}})$? Explain your reasoning.

(21) From previous homework. Last two are ($\star$).

3. **Tree factorization.** Let T denote the edges of a tree-structured pairwise Markov random field with vertices V. For the special case of trees, prove that *any* distribution $p_T(\mathbf{x})$ corresponding to a Markov random field over T admits a factorization of the form:

$$p_T(\mathbf{x}) = \prod_{(i,j)\in T} \frac{p_T(x_i, x_j)}{p_T(x_i)p_T(x_j)} \prod_{j\in V} p_T(x_j), \tag{2}$$

where $p_T(x_i, x_j)$ and $p_T(x_i)$ denote pairwise and singleton marginals of the distribution p_T, respectively.

Hint: consider the Bayesian network where you choose an arbitrary node to be a root and direct all edges away from the root. Show that this is equivalent to the MRF. Then, looking at the BN's factorization, reshape it into the required form.

4. *Hammersley-Clifford and Gaussian models:* Consider a zero-mean Gaussian random vector $(X_1, \ldots, X_N)$ with a strictly positive definite $N \times N$ covariance matrix $\Sigma \succ 0$. For a given undirected graph $G = (V, E)$ with N vertices, suppose that $(X_1, \ldots, X_N)$ obeys all the basic conditional independence properties of the graph G (i.e., one for each vertex cut set).

   (a) Show the sparsity pattern of the inverse covariance $\Theta = (\Sigma)^{-1}$ must respect the graph structure (i.e., $\Theta_{ij} = 0$ for all indices i, j such that $(i, j) \notin E$.)

   (b) Interpret this sparsity relation in terms of cut sets and conditional independence.

5. *Undirected trees and marginals:* Let $G = (V, E)$ be an undirected graph. For each vertex $i \in V$, let μ_i be a strictly positive function such that $\sum_{x_i} \mu_i(x_i) = 1$. For each edge, let μ_{ij} be a strictly positive function such that $\sum_{x_i} \mu_{ij}(x_i, x_j) = \mu_j(x_j)$ for all x_j, and $\sum_{x_j} \mu_{ij}(x_i, x_j) = \mu_i(x_i)$ for all x_i. Suppose moreover that $\mu_{ij}(x_i, x_j) \neq \mu_i(x_i)\mu_j(x_j)$ for at least one configuration (x_i, x_j). Given integers $m_1, \ldots, m_n$, consider the function

$$r(x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \left[\mu_i(x_i)\right]^{m_i} \prod_{(i,j)\in E} \mu_{ij}(x_i, x_j).$$

Supposing that G is a tree, can you give choices of integers $m_1, \ldots, m_n$ for which r is a valid probability distribution? If so, prove the validity. (*Hint:* It may be easiest to first think about a Markov chain.)