CSC412 – Final Test

EXAMPLE QUESTIONS
Time: 60 minutes
Aid Sheet and Calculators Permitted

Name: Student Number:

Use the backs of pages if needed.

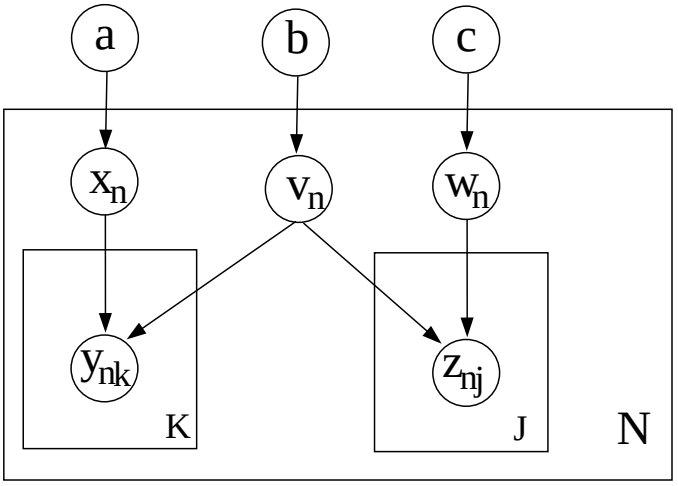
1 Plates and Generative Model Descriptions

Convert the following generative model descriptions into a graphical model diagram using plates when necessary.

- A physics experiment is set up to analyze a set of N radioactive samples. Each sample has a radioactive rate λ_n drawn from $p(\lambda|\alpha)$. The experiment uses J separate detectors, each of which has a sensitivity $\mathbf{y}_j$ drawn from $p(\mathbf{y}|\beta)$. Each sample n is measured using each detector j producing a result $\mathbf{x}_{nj}$ from $p(\mathbf{x}_{nj}|\lambda_n, \mathbf{y}_j)$.

- A global mean μ is drawn from a distribution p$(\mu|\alpha)$. Then, for each of K clusters, a cluster covariance Σ_k is drawn iid from a distribution $p(\Sigma|\beta)$, the number of data points n_k for that cluster is drawn iid from a distribution $p(n|\gamma)$ and then n_k points are drawn iid from the distribution $p(\mathbf{x}|\mu, \Sigma_k)$. All clusters have the same mean.

Describe in words the generative process implied by the graphical model below.



2 Junction Trees

In order to use a junction tree algorithm for inference, we must first "precompile" our graphical model. This involves moralization (if the original GM is directed), triangulation (which is hard), maximal clique finding (which is easy in triangulated graphs), and building the junction tree (which is easy).

- Give a very short recipe for moralizing a directed graphical model.

- Briefly define what it means for an undirected graph to be triangulated.

- Give a short definition of maximal cliques.

- Briefly describe what it means for an undirected graph over cliques to have the "junction-tree property".

- Briefly describe a way to generate a valid junction tree given a set of maximal cliques.

- Not all graphs have junction trees.
  What two other properties are equivalent to the graph property of there existing a valid junction tree over the maximal cliques? Give a definition of each of these properties in terms of the original graph.

- What do the three equivalent properties above all imply about the joint probability distribution represented by the original graphical model?

3 Learning in Undirected Models

- What do "IPF" and "GIS" stand for?

- Give two differences between the IPF updating procedure and the GIS updating procedure.

- For models which have sufficient statistics (i.e. generalized exponential family models), describe what condition is satisfied in terms of the model distribution and the sufficient statistics when the likelihood of a data sample is maximized.

- If, instead of maximizing likelihood, we had assumed the condition satisfied above and maximized some other quantity, we would have found that the exponential family form was the optimal form of the model. Give the name and the mathematical form of the quantity we would have had to maximize subject to the condition above in order to get the exponential form?

4 EM Algorithm for Unobserved Naive Bayes

Consider the following "unobserved naive Bayes" model which has P *observed binary* variables $x_i \in \{0, 1\}$ $(i = 1 \dots P)$, and an *unobserved discrete* latent variable $z \in \{1, 2, \dots, K\}$.

$$p(z = k) = a_k$$
$$p(x_i = 1 | z = k) = b_{ik} \quad \forall i$$

Below you will derive the EM algorithm for maximum likelihood learning in this model.

- Write the complete data log likelihood for a dataset with N observations x_i^n and latent variables z^n, $i = 1 \dots P$, $n = 1 \dots N$.

- Calculate the marginal (incomplete) data log likelihood for some observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$.

- E-step: calculate the posterior $p(z = k | x_1, \dots, x_P)$ of the latent variable given the binary observations.

- Calculate the expected complete data log likelihood for the observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$ under a distribution $p(z^n = k | x_1^n \dots x_P^n) = q_k^n$.

- M-step: For a fixed q_k^n, and fixed observed data x_i^n, find the parameter settings a_k^* and b_{ik}^* which maximize the expected complete log likelihood. Be sure to enforce the normalization constraint $\sum_k a_k = 1$.

- Assume we have some observed data x_i^n, $i = 1 \dots P$, $n = 1 \dots N$ and we want to fit this model using the EM algorithm. Using the results of the previous subquestions, write down the $E - step$ update for q_k^n and the $M - step$ updates for a_k and b_{ik}.
  Make sure the updates you write don't contain unspecified quantities.
  You should be able to turn your updates into code without any further derivations.

5 Bayesian Statistics

5.1 Bayes Error

In a binary classifier we want to decide between two settings $c = 0$ and $c = 1$ of a binary class variable c given some input vector $\mathbf{x}$. The lowest possible average error achievable by a binary classifier on a given problem is the *Bayes error*, obtained by using the optimal Bayesian decision rule.

- If we know the true probability distributions $p(\mathbf{x})$ and $p(c|\mathbf{x})$, what is the optimal Bayesian decision rule for selecting the class given the input?

- What is the Bayes error in terms of $p(\mathbf{x})$ and $p(c|\mathbf{x})$?

- Give simple upper and lower bounds on the Bayes error (but not the trivial bounds 1 and 0).

5.2 Bayesian Evidence

Evidence left at a crime scene identifies the criminal as one who has a rare blood condition, that only one in one thousand people have. The police lab has developed a binary test for this condition that has the following properties: $p(test = 1|condition = 1) = 0.99$ and $p(test = 0|condition = 0) = 0.99$

- What is $p(condition = 1|test = 1)$, i.e. what is the probability that a person for whom test=1 actually has the blood condition?

- What is $p(condition = 1|test = 1)$, if the prior for having the condition is only one in ten thousand?

6 Bayesian Prediction (Poisson)

Consider performing Bayesian learning on a finite data sample $\mathcal{D} = \{x^1, \ldots, x^N\}$ of non-negative integer counts.

- Assume that we want to use Poisson model with rate λ to model the data as iid samples:

$$p(x^n|\lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}$$

If we use a gamma prior on the rate parameter

$$p(\lambda|s, c) = \frac{1}{sG(c)} \left(\frac{\lambda}{s}\right)^{c-1} \exp\left\{-\frac{\lambda}{s}\right\}$$

compute the posterior $p(\lambda|\mathcal{D}, s, c)$ and show it is also gamma.

- Compute the posterior predictive distribution $p(x^{new}|\mathcal{D})$ over a new count.