Assignment 3: Graphical Models and Bayesian Treatment of Probabilistic Models

Probabilistic and Unsupervised Learning

Maneesh Sahani and Yee Whye Teh

Due: Mon Nov 19, 2007

Note: all assignments for this course are to be handed in to the Gatsby Unit, **not** to the CS department. Please hand in all assignments at the beginning of lecture on the due date to the lecturer. Late assignments will be penalised. If you are unable to come to class, you can also hand in assignments to Rachel Howes in the Alexandra House 4th floor reception.

Please attempt the first questions before the bonus ones.

1. **[10 marks] Gaussian graphical models.** Consider a multivariate Gaussian variable $\mathbf{x} = (x_1, \ldots, x_n)$ with given mean vector μ and covariance matrix Σ.

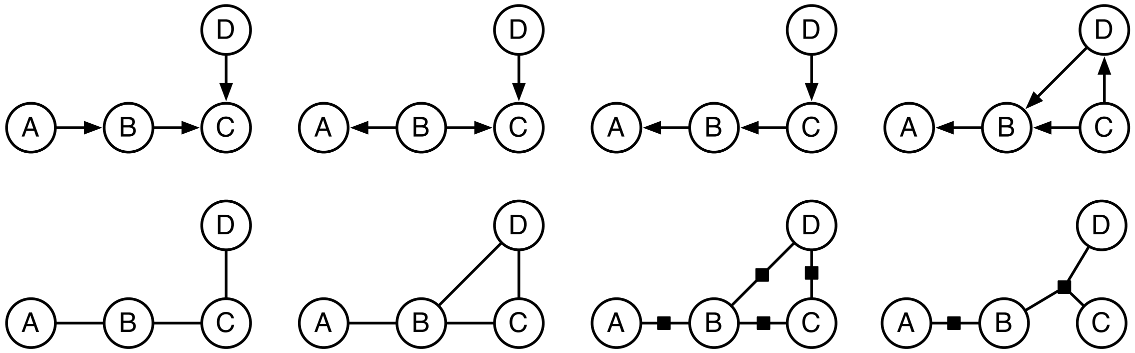
   (a) Write out the probability density function for $\mathbf{x}$. [2 marks]

   (b) Let $n = 4$, $\mu = (0, 1, 1, 0)$ and

   $$\Sigma = \frac{1}{6}\begin{pmatrix} 7 & -2 & -2 & 1 \\ -2 & 7 & 1 & -2 \\ -2 & 1 & 7 & -2 \\ 1 & -2 & -2 & 7 \end{pmatrix},$$

   draw the corresponding undirected graph and define clique potentials consistent with the above Gaussian. [Hint: multiply out the terms that appear in the exponent.] [8 marks]

2. **[25 marks] Conditional independencies and expressiveness of graphical models.** Consider the following graphical models:



   (a) For each graph, write down all the conditional independence relationships for variable C of the form $C \perp\!\!\!\perp \mathbf{X} | \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ can be sets of other variables. [15 marks]

   (b) Two graphs are **equivalent** if they express *all* the same marginal and conditional independence relationships between their variables. A graph G is **subsumed** by graph H if all conditional independence relationships in H are exhibited in G. Divide the above 8 graphs into the smallest number of non-overlapping sets of equivalent graphs, and state which of these sets of equivalent graphs are subsumed by one of other sets. [10 marks]

3. **[30 marks] Constructing directed graphs and junction trees.** You are the doctor on the Star Trek Enterprise and you are attempting to use Bayesian methods to help your diagnosis abilities. You would like to represent your knowledge about the following seven binary random variables describing the state of your patients on any given visit

M = has the disease microsoftus
L = has the disease linuxitis
A = has the disease applosis
V = is a vulcan (V=0 means "is a human")
H = has high temperature
P = likes pizza
B = has blue spots on face

You would like to build a directed graphical model which captures the following background knowledge:

Microsoftus is a rare disease.
Linuxitis and applosis are very rare diseases.
There are about four times as many humans as vulcans on the ship.
Vulcans have higher probability of getting microsoftus than humans.
Most vulcans like pizza, some humans like pizza.
Microsoftus usually causes high temperature and blue spots on the face.
Linuxitis always causes high temperature.
Applosis sometimes causes blue spots on the face.

(a) Draw a directed graphical model representing the relationships between the above variables. If you need to make any additional assumptions to draw your graph, state clearly what they are. [5 marks]

(b) For each variable in your graph, define a conditional probability table for that variable given the settings of its parents. Use the above background knowledge and convert those statements into probability tables which you think reasonably represent them. You will have to make up numbers for what terms like “rare”, “most”, and “usually” mean. [10 marks]

(c) Construct a junction tree for your directed graph, drawing out the intermediate factor graph, undirected graph and chordal graph. Use the minimum deficiency search variable elimination order, and show the clique factors on the resulting junction tree. [10 marks]

(d) Using Shafer-Shenoy propagation on the junction tree, compute the probability

P(patient is a vulcan | patient has blue spots and high temperature)

Show each message computed. Does this probability match your intuitions? [5 marks]

4. **[20 marks] Gaussian processes.** Consider zero-mean Gaussian processes with the following covariance kernels (parameters are all positive, and s and t are real numbers):

$$k_1(s,t) = \theta_1^2 \exp\left(-\frac{2\sin^2(\pi(s-t)/\tau_1)}{\sigma_1^2}\right)\exp\left(-\frac{(s-t)^2}{2\eta_1^2}\right) + \zeta_1^2\delta_{s=t}$$
$$k_2(s,t) = \theta_2^2 \left(\exp\left(-\frac{2\sin^2(\pi(s-t)/\tau_2)}{\sigma_2^2}\right) + \phi_2^2\exp\left(-\frac{(s-t)^2}{2\eta_2^2}\right)\right) + \zeta_2^2\delta_{s=t}$$

   (a) Write a MATLAB function to generate samples drawn from a GP. Specifically, given as input a covariance kernel function and a vector of input points $\mathbf{x}$, return a function $f(\mathbf{x})$ evaluated on the input points $\mathbf{x}$ drawn randomly from a GP with the given covariance kernel and with zero mean. [10 marks]

   (b) For each of the two kernels above, use your MATLAB function to draw samples from GPs with the given covariance kernel for various values of the parameters.

   Produce plots of the drawn functions with input points $\mathbf{x} = -10, -9.9, -9.8, \ldots, 9.9, 10$, and with:

      i. Covariance kernel k_1 with $\theta_1 = 10, \tau_1 = 2, \sigma_1 = 2, \eta_1 = 5$ and $\zeta_1 = .001$.

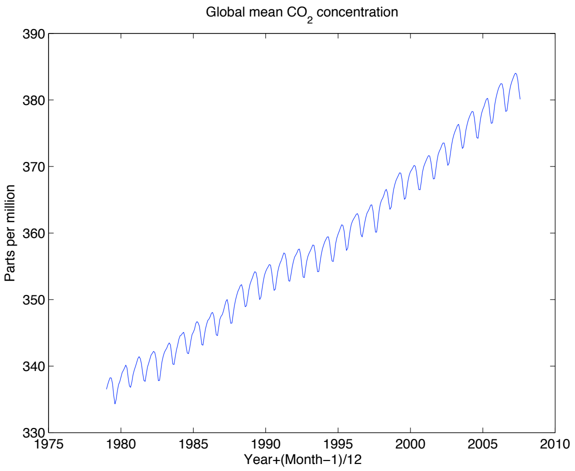
      ii. Covariance kernel k_1 with $\theta_1 = 1, \tau_1 = 1, \sigma_1 = 1, \eta_1 = 1000$ and $\zeta_1 = .001$.

      iii. Covariance kernel k_2 with $\theta_2 = 1, \tau_2 = 2, \sigma_2 = 2, \phi_2 = 2, \eta_2 = 4$ and $\zeta_2 = .001$.

      iv. Covariance kernel k_2 with $\theta_1 = 1, \tau_1 = 2, \sigma_1 = .5, \phi_2 = 4, \eta_2 = 10$ and $\zeta_2 = .001$.

   Describe the characteristics of the drawn functions, and how the characteristics of the functions depend on the parameters. How do the two covariance kernels differ from each other? [10 marks]

5. **[15 marks] Bayesian linear and Gaussian process regression.** The following time series of monthly mean global CO$_2$ concentrations can be obtained from the file co2.txt (original data obtained from http://www.esrl.noaa.gov/gmd/ccgg/trends):



We will apply Bayesian linear and Gaussian process regression to predict the CO$_2$ concentration $f(t)$ as a function of time t, where $t = \text{Year} + (\text{Month} - 1)/12$.

   (a) First we model the function using linear regression, that is, using the functional form

$$f(t) = at + b + \epsilon(t),$$

with i.i.d. noise residual $\epsilon(t) \sim \mathcal{N}(0,1)$ and prior $a \sim \mathcal{N}(0,10^2)$, $b \sim \mathcal{N}(360,100^2)$. Compute (using MATLAB) the posterior mean and covariance over a and b given the CO_2 data. [5 marks]

(b) Let $a_{\text{MAP}}, b_{\text{MAP}}$ be the MAP estimate in the question above. The residual is the difference between the observed function values and the predicted mean function values

$$g_{\text{obs}}(t) = f_{\text{obs}}(t) - (a_{\text{MAP}}t + b_{\text{MAP}}),$$

where $f_{\text{obs}}(t)$ is the observed value of the CO_2 concentration at time t. Plot $g_{\text{obs}}(t)$. Do you think these residuals conform to our prior over $\epsilon(t)$? State, with justifications, which characteristics of the residual you think do or do not conform to our prior belief. Suppose we were to consider modelling the residual function $g(t)$ using a zero mean GP with covariance kernel k_2 as given in question 4. Based on the plot of $g(t)$ what do you think will be suitable values for the parameters of k_2? [10 marks]

(c) [Bonus] Extrapolate the CO_2 concentration levels to 2020 using the GP with covariance kernel k_2 and your chosen parameter values. Specifically, compute the predictive mean and variance of the residual $g(t)$ for every month between September 2007 and December 2020 given the observed residuals $g_{\text{obs}}(t)$. Plot the means and one standard deviation error bars of the extrapolated CO_2 concentration levels

$$f(t) = a_{\text{MAP}}t + b_{\text{MAP}} + g(t)$$

along with the observed CO_2 levels. Does the behaviour of the extrapolation conform to your expectations? [5 bonus marks]

(d) [Bonus] Why is the above procedure not Bayesian? How would we go about modelling $f(t)$ in a Bayesian framework? [5 bonus marks]

6. **[Bonus: 25 marks]** Consider the following two HMMs:

$$P_1(\mathbf{x}_{1:T}, \mathbf{s}_{1:T}) = P(\mathbf{s}_1)P(\mathbf{x}_1|\mathbf{s}_1)\prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{s}_t)P(\mathbf{s}_t|\mathbf{s}_{t-1})$$

and

$$P_2(\mathbf{x}_{1:T}, \mathbf{r}_{1:T}) = P(\mathbf{r}_1)P(\mathbf{x}_1|\mathbf{r}_1)\prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{r}_t)P(\mathbf{r}_t|\mathbf{r}_{t-1})$$

where $\mathbf{x}_t$ is the observation at time t, $\mathbf{s}_t$ and $\mathbf{r}_t$ are the hidden state variables for each HMM, respectively, and the notation $\mathbf{x}_{1:T}$, for example, denotes the sequence $\mathbf{x}_1 \ldots \mathbf{x}_T$, Now form a new model for the data by multiplying these two models and renormalizing:

$$P_3(\mathbf{x}_{1:T}, \mathbf{s}_{1:T}, \mathbf{r}_{1:T}) = \frac{1}{Z}P_1(\mathbf{x}_{1:T}, \mathbf{s}_{1:T})P_2(\mathbf{x}_{1:T}, \mathbf{r}_{1:T})$$

(a) Draw a factor graph—with a node for each variable $\mathbf{x}_t$, $\mathbf{s}_t$, and $\mathbf{r}_t$—representing the conditional independence relationships in this new model, P_3. [2 marks]

(b) Given a sequence $\mathbf{x}_{1:T}$, describe how you would compute $P(\mathbf{s}_t, \mathbf{r}_t|\mathbf{x}_{1:T})$. What is the time complexity of your algorithm? [8 marks]

(c) If $\mathbf{s}_t$ and $\mathbf{r}_t$ are both discrete, taking on at most K states, is this model equivalent to an HMM with K^2 states? Why or why not? [5 marks]

(d) Assume you want to learn the parameters of this model from data. Let's re-write the model more explicitly to make it clear:

$$P_3(\mathbf{x}_{1:T}, \mathbf{s}_{1:T}, \mathbf{r}_{1:T}|\theta, \phi) = \frac{1}{Z(\theta, \phi)} P_1(\mathbf{x}_{1:T}, \mathbf{s}_{1:T}|\theta) P_2(\mathbf{x}_{1:T}, \mathbf{r}_{1:T}|\phi)$$

where θ and ϕ are the usual transition, emission, and initial state HMM parameters for HMM 1 and 2, respectively, and Z is the normalization term, which depends on these parameters. What is the derivative of the log likelihood of P_3 with respect to the transition parameter, $\theta_{ij} \equiv P(\mathbf{s}_{t+1} = j|\mathbf{s}_t = i)$? Calculate the derivative of Z explicitly. [5 marks]

(e) Assume there are L possible symbols: $\mathbf{x}_t \in \{1, \dots, L\}$ and each HMM has K states. What is the maximum mutual information between $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ in this model, maximizing over all parameters? [5 marks]