Inference and Representation, Fall 2014

Problem Set 2: Undirected graphical models

Due: Friday, September 26, 2014 at 5pm (as a PDF document sent to pg1338@nyu.edu. Please make sure the filename is in the format xyz-ps2.pdf, where xyz is your NetID.)

Important: See problem set policy on the course web site.

1. Exercise 26.1 from Murphy’s book (causal reasoning in the sprinkler network).

2. Recall that an Ising model is given by the distribution
   \[ p(x_1, \cdots, x_n) = \frac{1}{Z} \exp \left( \sum_{(i,j) \in E} w_{i,j} x_i x_j - \sum_{i \in V} u_i x_i \right), \] 
   where the random variables \( X_i \in \{-1, +1\} \). Related to the Ising model is the Boltzmann machine, which is parameterized the same way (i.e., using Eq. 1), but which has variables \( X_i \in \{0, 1\} \). Here we get a non-zero contribution to the energy (i.e. the quantity in the parentheses in Eq. 1) from an edge \((i, j)\) only when \( X_i = X_j = 1 \).
   Show that a Boltzmann machine distribution can be rewritten as an Ising model. More specifically, given parameters \( \vec{w}, \vec{u} \) corresponding to a Boltzmann machine, specify new parameters \( \vec{w}', \vec{u}' \) for an Ising model and prove that they give the same distribution \( p(X) \) (assuming the state space \( \{0, 1\} \) is mapped to \( \{-1, +1\} \)).

3. Give a procedure to convert any Markov network on discrete variables into a pairwise Markov random field. In particular, given a distribution \( p(X) \), specify a new distribution \( p'(X, Y) \) which is a pairwise MRF, such that \( p(x) = \sum_y p'(x, y) \), where \( Y \) are any new variables added.
   Clarification: Assume that the input is specified as full tables specifying the value of the potential for every assignment to the variables for each potential. The new pairwise MRF must have a description which is polynomial in the size of the original MRF.
   Hint: First consider a simple case, such as a MRF on 3 binary variables with a single potential function for the 3 variables, i.e. \( p(X) \propto \psi_{123}(X_1, X_2, X_3) \). Introduce a new variable \( Y \) with \( 2^3 = 16 \) states and let \( p'(X, Y) \propto \psi_Y(Y) \psi_{1Y}(X_1, Y) \psi_{2Y}(X_2, Y) \psi_{3Y}(X_3, Y) \). Figure out how to set the new potential functions \( \psi_Y(Y), \psi_{1Y}(X_1, Y), \psi_{2Y}(X_2, Y) \) and \( \psi_{3Y}(X_3, Y) \) so as to have \( p(x) = \sum_y p'(x, y) \) for all assignments \( x \).

4. Exponential families (see Chap. 9). Probability distributions in the exponential family have the form:
   \[ p(x; \eta) = h(x) \exp\{\eta \cdot f(x) - \ln Z(\eta)\} \]
   for some scalar function \( h(x) \), vector of functions \( f(x) = (f_1(x), \ldots, f_d(x)) \), canonical parameter vector \( \eta \in \mathbb{R}^d \) (often referred to as the natural parameters), and \( Z(\eta) \) a constant (depending on \( \eta \)) chosen so that the distribution normalizes.
   (a) Determine which of the following distributions are in the exponential family, exhibiting the \( f(x) \), \( Z(\eta) \), and \( h(x) \) functions for those that are.
i. \( N(\mu, I) \)—multivariate Gaussian with mean vector \( \mu \) and identity covariance matrix.

ii. \( \text{Dir}(\alpha) \)—Dirichlet with parameter vector \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K) \) (see Sec. 2.5.4).

iii. \( \text{log-Normal distribution} \)—the distribution of \( Y = \exp(X) \), where \( X \sim N(0, \sigma^2) \).

iv. \( \text{Boltzmann distribution} \)—an undirected graphical model \( G = (V, E) \) involving a binary random vector \( \mathbf{X} \) taking values in \( \{0, 1\}^n \).

\[ p(\mathbf{X}) \propto \exp\left\{ \sum_i u_i x_i + \sum_{(i,j) \in E} w_{i,j} x_i x_j \right\}. \]

(b) **Conditional models.** One can also talk about conditional distributions being in the exponential family, being of the form:

\[ p(y \mid x; \eta) = h(x, y) \exp\{\eta \cdot f(x, y) - \ln Z(\eta, x)\}. \]

The partition function \( Z \) now depends on \( x \), the variables that are conditioned on.

Let \( Y \) be a binary variable whose conditional distribution is specified by the logistic function,

\[ p(Y = 1 \mid x; \alpha) = \frac{1}{1 + e^{-\alpha_0 - \sum_{i=1}^n \alpha_i x_i}}. \]

Show that this conditional distribution is in the exponential family.

5. **Conjugacy and Bayesian prediction.**

(a) Let \( \theta \sim \text{Dir}(\alpha) \). Consider discrete random variables \( (X_1, X_2, \ldots, X_N) \), where \( X_i \sim \text{Cat}(\theta) \) for each \( i \) (thus the \( X_i \) are conditionally independent of one another given \( \theta \)).

Show that the posterior \( p(\theta \mid x_1, \ldots, x_N, \alpha) \) is given by \( \text{Dir}(\alpha') \), where

\[ \alpha'_k = \alpha_k + \sum_{i=1}^N 1[x_i = k]. \]

This property, that the posterior distribution \( p(\theta \mid x) \) is in the same family as the prior distribution \( p(\theta) \), is called **conjugacy**. The Dirichlet distribution (see Sec. 2.5.4) is the conjugate prior for the Categorical distribution. Every distribution in the exponential family has a conjugate prior. For example, the conjugate prior for the mean of a Gaussian distribution can be shown to be another Gaussian distribution.

(b) Now consider a random variable \( X_{\text{new}} \sim \text{Cat}(\theta) \) that is assumed conditionally independent of \( (X_1, X_2, \ldots, X_N) \) given \( \theta \). Compute:

\[ p(x_{\text{new}} \mid x_1, x_2, \ldots, x_N, \alpha) \]

by integrating over \( \theta \).

*Hint:* Your result should take the form of a ratio of gamma functions.

This is called **Bayesian prediction** because we put a prior distribution over the parameters \( \theta \) (in this case, a Dirichlet) and are thus able to take into consideration our initial uncertainty over (and prior knowledge of) the parameters together with the evidence we observed (samples \( x_1, \ldots, x_N \)) when giving our predictions for \( x_{\text{new}} \).