CS 194-10, Fall 2011
Assignment 7

1. (15) Do Ex. 14.7 from Russell & Norvig.

2. *Exponential Family* (15)
   [[Note: for the purposes of this question, the corresponding Wikipedia articles are off limits.]]
   A probability distribution in the *exponential family* takes the following form:

   $$p(x) = h(x) exp\{\theta^T T(x) - A(\theta)\}$$

   where θ is called the *natural parameter*, $h(x)$ and $T(x)$ are arbitrary functions, and $A(\theta)$ provides the required normalization.

   (a) Show that the following distributions are in the exponential family by defining θ and finding suitable $h(x)$, $T(x)$, and $A(\theta)$.
   (i) Normal$(\mu, 1)$; (ii) Bernoulli(p); (iii) Categorical$(p_1, \ldots, p_K)$.

   (b) A nice fact is that, given N i.i.d. observations from a distribution in the exponential family, the maximum likelihood estimator of θ is consistent, i.e., it converges to the true value in probability as $N \to \infty$. Show that the ML estimate of θ is a function only of $T(x)$, which is referred as a *sufficient statistic* for θ.

   (c) The function $A(\theta)$ has moment generating properties; in particular: $\frac{\partial}{\partial \theta} A(\theta) = E[T(x)]$. Demonstrate this property for the Normal$(\mu, 1)$ distribution.

3. *EM with discrete variables* (10)
   Consider the three-node Bayesian network $X \to Y \to Z$. The following data are and the following data, where missing entries are marked as '?':

x	y	z
0	0	0
0	0	?
0	1	0
1	1	1
1	?	1
1	0	1

   (a) Use the data to estimate initial parameters for this network, using maximum likelihood based only on the observable counts in the data.

   (b) Apply the EM algorithm (by hand) until convergence. In each iteration, compute the expected value for each missing data point given the current parameters and other observed values; then recompute the parameters from the "completed" data. Show your calculations at each step.

   (c) How many iterations does EM take to converge? Will this always be the case? Explain.

4. *Learning with continuous variables* (15)
   Consider a Bayesian network $X \to \mathbf{Y} \to Z$ where X and Z are discrete with values in $\{-1, 1\}$ and $\mathbf{Y}$ is a continuous *vector-valued* variable with values in $\mathbb{R}^D$. Let $P(X) = [0.5, 0.5]$, $P(Y_j \mid X = k) \sim N(\mu_{jk}, \sigma^2)$ where the μ_{jk}s and σ are unknown parameters, and $P(Z = z \mid \mathbf{Y} = \mathbf{y}) = 1/(1 + exp(-z\mathbf{w}^T\mathbf{y}))$, where $\mathbf{w}$ is a fixed set of D unknown weights.

   (a) Given N i.i.d. observations of $(X, \mathbf{Y}, Z)$ write the log-likelihood and its derivative with respect to each unknown parameter.

(b) Consider applying EM under three different conditions: (i) X and $\mathbf{Y}$ are observed and Z is not; (ii) Z and $\mathbf{Y}$ are observed and X is not; (iii) $\mathbf{Y}$ is observed and X and Z are not. Describe qualitatively what learning task EM is solving in each case.

5. *Bayes Net Inference using BayesiaLab* (50)
In this exercise you will use a commercial Bayes net package to learn the **car insurance network** that was discussed in lecture. The observable variables in the **training set** are the ones typically available on a client application form, plus the resulting claim costs incurred by the company in three categories (`PropCost` for property losses, `MedCost` for medical expenses, and, `ILiCost` for intangible liabilities). After training, the network can be used to predict those expenses for new applications in the **test set**. The same training and test sets can be used with a traditional algorithm such as decision tree learning.

   (a) Download and install BayesiaLab following the instructions posted on piazza (and sent to you by email),

   (b) Open the insurance network file by BayesiaLab→Network→Open. To view the network, click on BayesiaLab→View→Automatic Layout→Force Directed Layout.

   (c) Learn conditional probabilities by clicking on BayesiaLab→Data→Associate Data Source→Text File and select the `training.csv` file. Then select Structural EM algorithm for missing data.

   (d) Now we have a complete generative model of the data. To do inference, click on Validation Mode button located at the bottom left corner of the network window. Double click on some nodes to see their probability distributions on the right. Then right click on the probabilities at the right to assign values to some nodes (evidence), and observe how the conditional probabilities change in other nodes. You should be able to fairly quickly find the setting of the observable variables that maximizes the probability of the `Million` outcome for the cost variables; record this setting and the corresponding probabilities.

   (e) Now predict the probabilities for the cost variables for the test cases. For each of the three cost nodes, do the following:

   - Set the node as the target node by right-clicking on it.
   - Supply the test data using BayesiaLab→Inference→Batch Inference and save the predictions to another file.

   (f) The true loss measure for the insurance problem is quite complicated, not only because the company can vary the premium according to the perceived risk but also because a higher premium is more likely to cause the applicant to go elsewhere for insurance. We will instead measure the quality of a probabilistic prediction using the predicted log probability that the model assigns to the correct answer. Report this value (summed over all cases and the three predictions in each case) using twofold cross-validation on the training data, both for Bayes net training and for decision tree learning using the code you wrote in A3. (The code may need to be modified to return probabilities at the leaves and you may need to smooth the probabilities to avoid zeroes.)

Turn in the results from part (d), the Bayes net prediction file for part (e), and the code for calculating prediction quality as well as the cross-validation estimates in (f). Supply documentation and explanations where appropriate. Submit your files collected together as a7.tar.gz using submit a7 as described **here**.