MACHINE LEARNING COMS 4771, HOMEWORK 5
Assigned April 4, 2013. Due April 18, 2013 before 1:00pm.

Here are the instructions for submitting your homework. Archive/package all of the files you are submitting as a single tarball or zip archive: "UNI-HW5.tar.gz" or "UNI-HW5.zip". For example, a compressed tarball would be "ir2322-HW5.tar.gz". Your homework should contain:

- a writeup (PDF, TXT, or PostScript)
- code (as Matlab M files, shorter code is generally better but include comments)
- any figures/pictures not included in the writeup (PDF or PostScript)
- if you have special instructions, include them as a plain text file called README.txt.

Submit your homework through CourseWorks by doing the following:

1 Log into https://courseworks.columbia.edu/
2 Click "Assignments" on the left side.
3 Choose the appropriate HW Folder to submit to.
4 Use the filename "yourUNI-HW5.tar.gz" or "yourUNI-HW5.zip".
5 Make sure that the "title" is yourUNI-HW5 (example: zz9999-HW5).
6 Add any special instructions in both the description and the README.txt.
7 Click "Submit" at the bottom to upload your file.
8 If you submit multiple times, only the last submission prior to the deadline will count.
9 If something goes wrong, ask the TAs for help.
10 In a dire emergency, if nothing else works, send your homework to the TAs.

Handwritten writeups are not allowed without prior approval.

All your code should be written in Matlab (other languages may be used only with prior permission from an instructor). Please submit all your souce files, each function in a separate file. Clearly denote what each function does, its inputs and outputs, and to which problem it belongs. Do not resubmit code or data provided to you. Do not submit code written by others. Identical submissions will be detected and both parties will get zero credit. Sample code is available on the Tutorials web page. Datasets are available from the Handouts web page. You may include figures directly in your write-up, or separately and refer to them by filename.

Each homework counts equally towards your grade (other than your worst which will be dropped). Points shown here for each problem indicate relative weights for this specific homework. As always, up to 10% bonus points are available for exceptional, relevant work going beyond what is asked.

1 Jensen's inequality (15 points)

Prove the following statements:

a) The arithmetic mean of non-negative numbers is at least their geometric mean.

b) $\sum_{i=1}^m \exp(\theta^\top f_i) \geq \exp\left(\theta^\top \sum_{i=1}^m \alpha_i f_i - \sum_{i=1}^m \alpha_i \log \alpha_i\right)$, where $\alpha_i = \frac{\exp\left(\hat{\theta}^\top f_i\right)}{\sum_{j=1}^m \exp(\hat{\theta}^\top f_j)}$.

HINT: Use Jensen's inequality.

2 EM for mixture of multinomials (20 points)

Consider a random variable x that is categorical with M possible values $1, \dots, M$. Suppose x is represented as a vector in M dimensions s.t. $x(j) = 1$ if x takes the j^{th} value, and $\sum_{j=1}^{M} x(j) = 1$. The distribution of x is described by a mixture of K discrete multinomial distributions such that:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\alpha_k)$$

where

$$p(x|\alpha_k) = \prod_{j=1}^{M} \alpha_k(j)^{x(j)}$$

where π_k denotes the mixing coefficients for the k^{th} component (aka the prior probability that the hidden variable $z = k$), and α_k specifies the parameters of the k^{th} component. Specifically, $\alpha_k(j)$ represents the probability $p(x(j) = 1|z = k)$ (and, therefore, $\sum_j \alpha_k(j) = 1$). Given an observed data set $\{x_i\}$, $i = 1, \cdots, N$, derive the E and M step equations of the EM algorithm for optimizing the mixing coefficients π_k and the component parameters $\alpha_k(j)$ for this distribution. For your reference, here is the generic formula for the E and M steps. Note that θ is used to denote all parameters of the mixture model.
E-step. For each i, calculate $q_i(z_k) = p(z_k|x_i; \theta)$, i.e., the probability of the k^{th} component given observation i and the current parameter settings.
M-step. Set

$$\theta := \arg\max_{\theta} \sum_{i=1}^{N} \sum_{z_k} q_i(z_k) \log \frac{p(x_i, z_k; \theta)}{q_i(z_k)}.$$

3 Conditional Probability (10 points total)

3.1 3 sisters

There are three sisters called Alice, Beatrice and Charlotte. You may assume any ordering of their ages is equally likely.

What is the probability that Alice is older than Beatrice?

Now you receive a message informing you that Alice is older than Charlotte. Given this information, what is the probability that Alice is older than Beatrice? (That is, what is the conditional probability $p(A > B|A > C)$?)

3.2 2 children

Joe has 2 children. Assume each child has equal chance of being a boy or a girl.

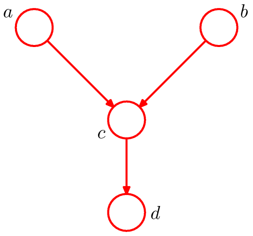
If the elder child is a boy, what's the chance that the younger child is a boy?

If instead, we only know that at least one of Joe's children is a boy, what's the chance that the other is a boy?

4 Bayesian Network Conditional Independence (25 points total)

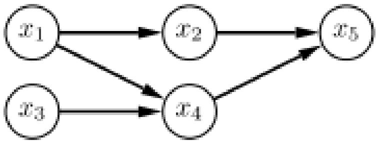
4.1 Explaining away (10 points)

Consider the Bayesian network below, in which none of the variables is observed. Write down the implied factorization of the probability distribution. Use this to show algebraically that a is independent of b. Suppose we now observe the variable d. Show that in general a is not conditionally independent of b given d. Make up a specific example using binary variables to demonstrate this, providing conditional probability tables and interpretations for each variable.



4.2 Bayes ball (15 points)

Consider the Bayesian network below, where binary variables represent the following assertions: x_1 student is intelligent, x_2 student is good at taking tests, x_3 student is hard working, x_4 student understands the material, and x_5 student gets a good grade.



Write out the factorization of the probability distribution $p(x_1, \ldots, x_5)$ implied by this directed graph. Using the Bayes ball algorithm, answer each of the following questions as either True (if true for all possible distributions) or False (if not true for all possible distributions). For each answer you mark False, provide one appropriate path a Bayes ball could take which demonstrates your answer (just list the variables in order). Also, pick *just one* of your solutions marked True and in addition, prove your answer algebraically using the factorization of the probability distribution (note some will be easier than others!).

1. x_2 and x_4 are independent.
2. x_2 and x_4 are conditionally independent given x_1, x_3 and x_5.
3. x_2 and x_4 are conditionally independent given x_1 and x_3.
4. x_5 and x_3 are conditionally independent given x_4.
5. x_5 and x_3 are conditionally independent given x_1, x_2 and x_4.
6. x_1 and x_3 are conditionally independent given x_5.
7. x_1 and x_3 are conditionally independent given x_2.
8. x_2 and x_3 are independent.

9. x_2 and x_3 are conditionally independent given x_5.
10. x_2 and x_3 are conditionally independent given x_5 and x_4.

5 MAP is not Max-Marginals (15 points)

Consider two discrete variables x and y, each with three possible states $x, y \in \{0, 1, 2\}$. Construct a joint probability distribution $p(x, y)$ over these variables such that the value $\hat{x}$ that maximizes the marginal $p(x)$, and the value $\hat{y}$ that maximizes the marginal $p(y)$, together have probability zero under the joint distribution, i.e. $p(\hat{x}, \hat{y}) = 0$. Ensure that $\hat{x}$ and $\hat{y}$ are the unique maximizers of their respective marginals.

Is it possible to construct such a probability distribution if both variables are binary, i.e. $x, y \in \{0, 1\}$? Either provide an example or prove it is not possible.