CSE 5523: Problem Set 6

Due via Carmen dropbox at 11:59 PM 11/22.

1. PROGRAMMING:

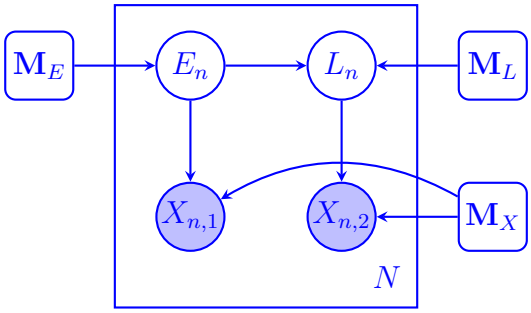
   (In general for programming problems you should hand in the following as separate files:

   - a copy of each program file you write,
   - a representative sample of each input file you use,
   - a representative sample of each output you produce.

   Unless otherwise noted, your programs should be as short as possible, and may be based on linear algebra and data analysis functions used in the lecture notes, but should not use higher-level packages or functions.)

   You are asked to use Gibbs sampling to fit models and latent variables over player types to 'longitudinal' data for subjects' favorite video games at two different time points: early E and late L.

   Here is a plate diagram of the intended model:



   You should assume two different values for latent variables, and should get converged results after about 10 iterations of Gibbs sampling.

   (a) [5 pts.] Using the sample code in the lecture notes on Gibbs sampling as a guide, which pairs of variables will you have to send backward messages between for each training example in this model, and what dimensions will the backward messages have?

   (b) [5 pts.] Using the sample code in the lecture notes on Gibbs sampling as a guide, which variables will you have to resample, and what dimensions will these sampled distributions have?

   (c) [5 pts.] Using the sample code in the lecture notes on Gibbs sampling as a guide, which models will you have to resample, and what dimensions will these models have?

   (d) [20 pts.] Adapt the example sample code from the lecture notes on Gibbs sampling to estimate model parameters and distributions over hidden variable values for the 'games.csv' data on the course web site. Your program should be named 'longi-topic.py' and should take the following as command-line arguments in the following order:

i. a filename of a csv file containing variable dimensions of training data points,

and print out each component probability model.

(e) [5 pts.] Run your inducer on the 'games.csv' data on the course web page and report the resulting models. How would you characterize what happens between time points? Which games are preferred at each time point?

2. PROJECT:

As noted before, your project may be in any area (vision, natural language, etc), may use any programming language you choose, and may use any data, even data used in previous projects or from data repositories or shared task challenges (e.g. https://archive.ics.uci.edu/ml/datasets.php), but should involve new code that you write to implement techniques you did not use previously and should therefore produce new results. You may work in groups, but in this case you should specify what portion of the project code will be produced by you and the other group members.

If you have not changed your project, and your available computer resources are adequate for your project, and you were not asked to modify your project in feedback on the previous problem set, you may skip this sub-question (please report that you are skipping it for this reason). If you ***have*** changed your project, or your compute resources are ***not*** adequate for your project, or you ***were*** asked to modify your project in feedback on the previous problem set, please resubmit the following (as in the previous problem set):

(a) [3 pts.] Describe in just a few sentences what *problem* your project will solve.

(b) [3 pts.] Describe in just a few sentences what *data* your project will use, including how many items (N) the data include, how many variables (K) are involved, and where your data will come from.

(c) [3 pts.] Describe in just a few sentences what *machine learning techniques* your part of your project will use.