Assignment 2
Miodrag Bolic

Due: February 25, 2021
Total number of points is 10. Please solve 2 out of 3 problems.

Instructions:
Upload your answers in a ipynb notebook to UOttawa Bright Space.

Your individual submissions should use the following filenames:
ELG_5218_YOURNAME_HW2.ipynb

Your code should be in code cells as part of your notebook. Do not use any different format.

*Do not just send your code. The homework solutions should be in a report style. Be sure to add comments to your code as well as markdown cells where you describe your approach and discuss your results. *

Please submit your notebook in an executed status, so that we can see all the results you computed. However, we will still run your code and all cells should reproduce the output when executed.

If you have multiple files (e.g. you've added code files or images) create a tarball for all files in a single file and name it: ELG_5218_YOURNAME_HW2.tar.gz or ELG_5218_YOURNAME_HW2.zip

Problem 1: Mixture model and variational inference

Mixture model is given in https://turing.ml/dev/tutorials/1-gaussianmixturemodel/.

a) Compare inference results obtained using Monte Carlo and variational inference.
b) Implement stochastic variational inference algorithm in Turing by yourself (from scratch) for GMM model. You can use as an example Algorithm 2 from this report https://www.in.tum.de/fileadmin/w00bws/daml/seminars_ss17/inference/student_submissions/paper_final_shugurov.pdf .
c) Compare results obtained using your code and Turing implementation.

Problem 2: Mixture model and Bayesian performance metrics

The question is about modeling traffic congestion in some part of the road using real sensor data. The dataset is collected in the highway of Los Angeles County in real time by loop detectors. A total of 207 sensors along with their traffic speed from Mar. 1 to Mar. 7, 2012 were selected. Dataset is attached with the assignment – its source is at https://github.com/lehaifeng/T-GCN. The first row represents the station where the data is collected and the other rows are the speed of traffic collected every 12 min approximately for about 7 days.

a) Draw data from the first column (rows 2-2017). Draw data as a histogram. Implement the model as a single Gaussian model as we studied in lectures 1 and 2. Criticize the model based on model checking criteria we studied in the last class.

b) Using Gaussian mixture model, fit this data to the mixture with 2 and 3 clusters. Plot the empirical distribution of means and variances of the clusters as well as of the probability that data belong to a particular cluster. You can use Turing example for mixture models starting with http://www.emmanuelkidando.com/Blog/GMM.html .

c) Evaluate performance of the models using WAIC and LOO metrics. You can obtain these metrics using ArviZ tool. Or, you can code Waic by yourself. Comment on your results.

d) Draw data from multiple colums. Do they have congestion (low speed) at about the same time?

e) (Advanced) Now, instead of looking at the data from a single column, use data from first 10 columns to train the model. Assume that there are 2 clusters. Develop hierarchical model for this data analysis. How do hierarchical models help here? One related work (however not the same) can be seen at https://link.springer.com/article/10.1007/s40534-019-00199-2 .

Problem 3: Bayesian neural networks and measurement error

Consider regression problem shown in this example:

1. Implement it in Julia using either Lengevin Monte Carlo or Variational Inference the following model: https://colab.research.google.com/github/papercup-open-source/tutorials/blob/master/intro_bnn/Bayesian_neural_networks_Part_2.ipynb#scrollTo=d-f2rHjfTvg9 where the datapoints are:
   {x,y}={ (-3, -3.5), (2, 6), (3, 4), (-3.5, -3), (4, 4.5)}.
   Replicate results from the web site for Langevin Monte Carlo and for Variational inference and present them with 95 % confidence intervals.
2. Add measurement error to x variable so that x~N(0,0.5). Observe what happens with your regression results if you do not include the noise into the model. What are 95% confidence intervals now? Then, modify the model to include measurement error. Observe results in this case.
3. Add measurement error to y variable so that y~N(0,0.5). Observe what happens with your regression results if you do not include the noise into the model. What are 95% confidence intervals now? Then, modify the model to include measurement error. Observe results in this case.