208891 Probabilistic Graphical Models **Spring 2019**

Homework 3: February 12

Due: 02/19/20

Use the EM algorithm `structural.em` provided by `bnlearn` package to learn a PDAG from the Pima Indians Diabetes Dataset (donlapark.cmustat.com/208891/data/pima-indians-diabetes.csv). The variables in this data consist of

1. Number of times pregnant.
2. Plasma glucose concentration at 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

Before doing anything, make sure that `Class`'s type is factor and all other variables' type is numeric. The minimal code to run the algorithm is

```
result = structural.em(data, maximize = "hc", maximize.args = list(restart=40), fit = "mle",
                                          impute="parents", return.all = TRUE, max.iter = 20)
```

which searches for the best structure using hill-climbing method with 40 random restarts. Alternatively, we can try the tabu search:

```
result = structural.em(data, maximize = "tabu", maximize.args = list(tabu=20, max.tabu=10),
                            fit = "mle", impute="parents", return.all = TRUE, max.iter = 20)
```

Here, `result` is a list of three elements,

- `result$dag` is the resulting graph from the search.
- `result$imputed` is the data with missing values filled out. You will need this data to compute BIC scores.
- `result$fitted` indicates the values of all parameters learned from the data.

The documentation about `structural.em` can be found at https://www.bnlearn.com/documentation/man/structural.em.html.

1. Experiment with the codes above and answer the following question.
   (a) Find the searching algorithm, together with the values of `maximize.args` and `max.iter`, that gives the best BIC score. Report your results along with the graph.
   (b) If you run the code with the value of `tabu` less than that of `max.tabu`, then an error will be showing up. Why is this the case?
   (c) Let $X =$ Diastolic blood pressure. What are the parents $\boldsymbol{U}$ of X? What is the distribution of $X|\boldsymbol{U}$ (you have to be explicit about the values of the parameters)?

2. With the model created, we can use it to answer the following type of query:

   Given that $X = x$ and $Z \geq z$, what is the probability that $Y \geq y$?

   which can be answered by running the following code

   ```
   cpquery(result$fitted, (Y>=y), (X==x & Z>=z))
   ```

   Among the group of diabetes patients with $\geq$ 50 BMI, what proportion has more than 80 diastolic blood pressure?

3. Suppose that we want to employ this model to predict each patient's diastolic blood pressure using other variables. We can evaluate the model by first computing the model's predictions (assume that you name the variable `bp`)

   ```
   bp_pred = predict(result($fitted, "bp", result$imputed)
   ```

   and then compute the root-mean-square error using `rmse` from `Metrics` library

   ```
   rmse(bp_pred, result$imputed$bp)
   ```

   What is the value of RMSE?