Data driven Business analytics – MGT-302
Homework 2

Due time: Friday May 19th, 8:00PM

Please submit your answers on Moodle by the due time.

Problem 1. Decision trees [25]
Fill in the jupyter notebook `Decision_Trees.ipynb`.

Problem 2. Neural networks [25]
We are interested in a classification problem for diabetes test. We have a small dataset "diabetes.csv" which you can find in the attachments, containing samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ with binary labels $y_i \in \{0, 1\}$.

Requirements
You are asked to:

- Implement a MLP neural network with two hidden layers to do classification. In your implementation, set $h_1 = 8$ and $h_2 = 8$, where h_i is the number of neurons in i-th hidden layer and use ReLU as your activation function. Use the logistic regression log-likelihood as your loss function:

$$\ell_\theta(\mathbf{x}, y) := \sum_{i=1}^{m} \left(y_i \log f_\theta(\mathbf{x}_i) + (1 - y_i) \log(1 - f_\theta(\mathbf{x}_i)) \right).$$

- Use the first 500 instances as your training data and the remaining as your test data.
- Train your neural network and 'show' (graphically) the convergence of your algorithm.
- Compute your test error. A discussion on your choice of the learning rate, the optimizer, and other parameters will be appreciated.

Some pieces of advice:

- You are free to use `pytorch`.
- Please note that the data is unbalanced, so you might want to use downsampling or upsampling to get better performance on the small class.

Dataset description

- Number of instances: 768
- Number of attributes: 8 plus class
- Attributes description: (all numeric-valued)
  - Number of times pregnant
  - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Diastolic blood pressure (mm Hg)
  - Triceps skin fold thickness (mm)
  - 2-Hour serum insulin (mu U/ml)
  - Body mass index (weight in kg/(height in m)ˆ2)
  - Diabetes pedigree function
  - Age (years)
  - Class variable (0 or 1)
- Missing Attribute Values: Yes
- Class Distribution: (class value 1 is interpreted "as tested positive for diabetes")

Class Value	Number of instances
0	500
1	268

Problem 3. Causal discovery [50]
Fill in the jupyter notebook `Causal_Discovery.ipynb`.

Problem 4. Causal inference [15]
Consider the directed acyclic graph (DAG) graph in Figure 1.

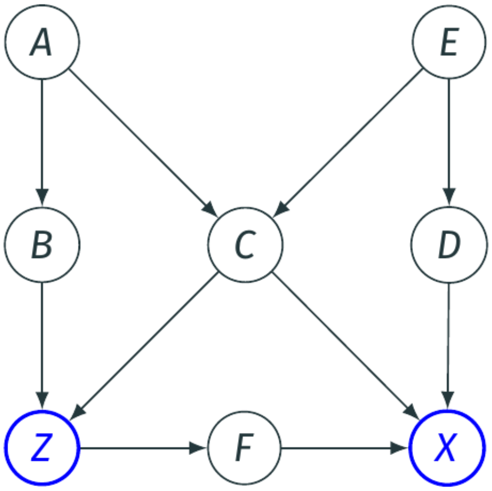

Figure 1: A directed acyclic graph.

1. Find a subset of variables that d-separates $\{Z\}$ from $\{X\}$.
2. Find two different valid adjustment sets for the ordered pair (Z, X), i.e., for computing $\mathbb{P}_{do(Z=z)}(X)$.
3. What are all the edges that one can modify in the graph so that the modified graph stays in the same Markov equivalence class?