Ve593
Jia Shi
716370990049
Prof. Weng

In this project, I construct a Bayes network to estimate the variable of interest in the wine dataset and protein dataset.

Structure learning:

For Wine dataset
Adjacency matrix

```
[[0 1 1 1 0 1 0 0 0 0 0 1]
 [0 0 1 1 1 1 1 1 1 0 1 1]
 [0 0 0 0 0 0 1 0 0 1 1 0]
 [0 0 0 0 1 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 1 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]]
```

I use BICScore and get the total score of -78794.2796937

For protein dataset
Adjacency matrix

```
[[0 0 1 0 0 1]
 [0 0 1 1 1 0]
 [0 0 0 1 1 1]
 [0 0 0 0 0 0]
 [0 0 0 0 0 0]
 [0 0 0 0 0 0]]
```

I use BICScore and get the total score of -195911.679481

For this part, I implement three scoring function, K2, K2Score and BICScore. For K2Score_log, When the dataset is too big(contain more than 171 data), the log function I implement will show overflow error, this happened even I implement the log function by myself. Since K2Score_log use factorial function, when the result is too big, even the simplest division function (2/1) will overflow as well.
But all the scoring function receive the same result in class when using the little example in the ppt.
Thus, most time I use K2 and BICScore to run the entire dataset.

Parameter learning:
The result is too long and too messy, please refer to the program itself.

General:
For wine dataset, I use K=2 and topological order [1, 10, 0, 3, 7, 4, 6, 5, 9, 2, 8, 11] with training data of 4700, and get the highest accuracy rate. When I using the method of ablation to select the most related variable, I found [1, 10, 0, 3, 7] is the most important variable.
Which is volatileacidity, alcohol, fixedacidity, residualsugar, density in the dataset.

For protein dataset, I use K=2(or 3) and topological order with 19000 data into training

order=[4,1,3,2,5,0] and get a accuracy rate of 93.5%, which is surprisingly high, and I found that that variable 4 is the most important one in the dataset, which is DNA D. When it fall to be placed in the first order, the rate drop tremendously.

With only observe K order, the importance of each variable is list in my topological order numerically.