# EAS 507: STATISTICAL DATA MINING II

Report for Homework #3

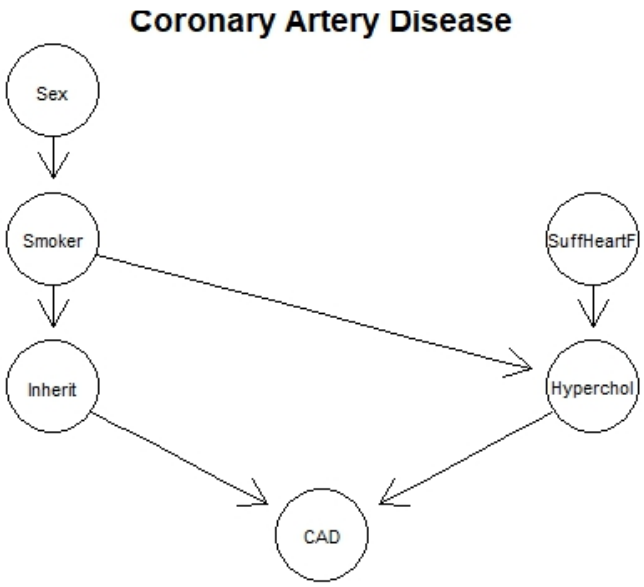
Kishore Ravisankar
UBIT: kravisan

**1.**

In this exercise, the cad1 dataset is given, which contains data observed at a Danish heart clinic. It contains 236 observations across 14 predictors. A structural learning algorithm has been used to construct an optimal network, which contains 6 predictors.

**a)**

The given network is constructed in R, which is shown as follows:




Now we identify some d-separations in the graph. We see that the following nodes are d-separated:

- Sex and SuffHeartF
- Sex and Inherit, given Smoker
- Sex and Hyperchol, given Smoker
- Inherit and SuffHeartF, given Smoker
- Smoker and CAD, given Inherit and Hyperchol

These d-separations are validated using the code attached.

To infer conditional probability tables, we use the compileCPT() function. The tables obtained are as follows.

Sex:

| Female | Male |
|---|---|
| 0.1991525 | 0.8008475 |

Smoker:

| Smoker/Sex | Female | Male |
|---|---|---|
| No | 0.3617021 | 0.1798942 |
| Yes | 0.6382979 | 0.8201058 |

Inherit:

| Inherit/Smoker | No | Yes |
|---|---|---|
| No | 0.8235294 | 0.6486486 |
| Yes | 0.1764706 | 0.3513514 |

CAD:

Hyperchol = No

| CAD/Inherit | No | Yes |
|---|---|---|
| No | 0.8214286 | 0.5 |
| Yes | 0.1785714 | 0.5 |

Hyperchol = Yes

| CAD/Inherit | No | Yes |
|---|---|---|
| No | 0.4487179 | 0.26 |
| Yes | 0.5512821 | 0.74 |

Hyperchol:

SuffHeartF = No

| Hyperchol/Smoker | No | Yes |
|---|---|---|
| No | 0.675 | 0.4645669 |
| Yes | 0.325 | 0.5354331 |

SuffHeartF = Yes

| Hyperchol/Smoker | No | Yes |
|---|---|---|
| No | 0.2727273 | 0.3275862 |
| Yes | 0.7272727 | 0.6724138 |

SuffHeartF:

| No | Yes |
|---|---|
| 0.708 | 0.292 |

**b)**

In this exercise, we have a new observation where the subject is Female, and has high cholesterol. Now, we find the probability of heart failure and CAD, with and without this evidence. The probabilities obtained are as follows:

Conditional probability:

With evidence:

| CAD/SuffHeartF | No | Yes |
|---|---|---|
| No | 0.6137859 | 0.3862141 |
| Yes | 0.6178472 | 0.3821528 |

Without evidence:

| CAD/SuffHeartF | No | Yes |
|---|---|---|
| No | 0.7326698 | 0.2673302 |
| Yes | 0.6782138 | 0.3217862 |

We notice that the conditional probability of having heart failure, irrespective of coronary artery disease (CAD), increases with evidence. In other words, the conditional probability of not having heart failure given the subject is female, and has high cholesterol, decreases with evidence.

Joint probability:

With evidence: Without evidence:

| CAD/SuffHeartF | No | Yes | CAD/SuffHeartF | No | Yes |
|---|---|---|---|---|---|
| No | 0.2408676 | 0.1515618 | No | 0.3957368 | 0.1443930 |
| Yes | 0.3753858 | 0.2321848 | Yes | 0.3118903 | 0.1479799 |

We see that the joint probability of not having both CAD and heart failure, decreases with evidence. In all other cases, the joint probability increases with evidence.

Marginal probability:

CAD:

With evidence: Without evidence:

| No | Yes | No | Yes |
|---|---|---|---|
| 0.3924294 | 0.6075706 | 0.5401298 | 0.4598702 |

SuffHeartF:

With evidence: Without evidence:

| No | Yes | No | Yes |
|---|---|---|---|
| 0.6162534 | 0.3837466 | 0.7076271 | 0.2923729 |

It can be observed that the marginal probability of having CAD increases with evidence. In addition, we can observe that the probability of having a heart failure also increases with evidence.

**c)**

In this exercise, we simulate 5 observations based on the network constructed with evidence. Using this data, we predict the probability of being a smoker, and having coronary artery disease.

On executing the code, we see that the probability of being a smoker and having CAD is 100%. In other words, the subject is a smoker, and has CAD in all the five cases.

The probability of finding evidence associated with these predictions are as follows:

```
$pEvidence
[1] 0.02882428 0.02882428 0.02882428 0.04488406 0.04488406
```

**d)**

The same exercise is repeated, but by simulating 500 observations this time. The simulated data is saved in simulated_obs.RData.

We make predictions on this dataset, and calculate the misclassification rate.

We obtain the following results.

The misclassification rate of the subject being a smoker is 30%.

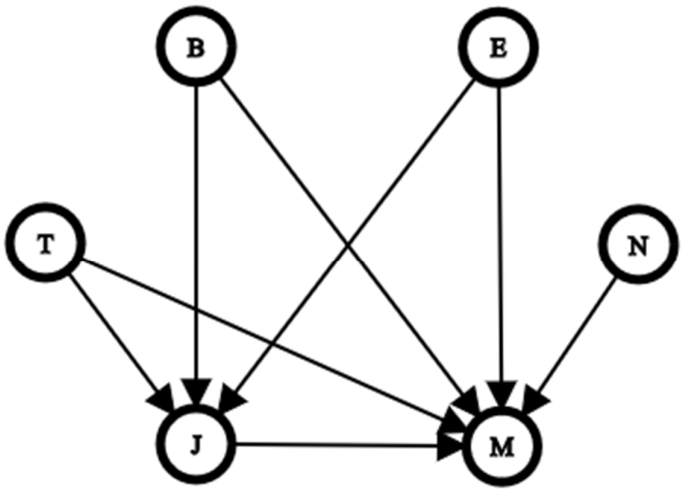
The misclassification rate of the subject having CAD is 39.6%.

Since the given network is based only on subjects who are female, and have high cholesterol levels, the accuracy and prediction performance of the network is questionable.

Therefore, to improve the performance of the network, the data should be balanced. To achieve that, the data can be sampled. Another aspect which can contribute to improving network performance would be selecting optimal nodes using cross-validation.

**2.**

For the given network B, we construct a Bayesian network B', which does not contain the node 'Alarm'. While constructing B', we make sure that all the original dependencies are preserved. The network B' is presented as follows:




B: Burglary

E: Earthquake

J: John Call

M: Mary Call

N: Nap

T: TV

**3.**

**a)** False

**b)** True

**c)** True

**d)** False

**e)** True