Anirudh Jain

Deriving Bayesian Networks from Melanoma Data

Purpose:

Bayesian networks have been previously used to model cancer cell pathways. The Melanoma cell signaling pathway illustrates the genes and their causal relationships that lead to cell proliferation and death which would be extremely useful when attempting to diagnose the causes of Melanoma. The purpose of this project is to learn a Bayesian network and its parameters from the cell dataset. The dataset contains expression values of genes from malignant melanoma cancer cells. Once the network has been derived, inference can be conducted to see the effects of certain genes on others. Thus, this is a useful tool to consider the impact of genes on the cell proliferation of Melanoma and whether are not they are key targets.

Melanoma is one of the most prevalent and aggressive forms of skin cancer. In tumor cells, the cell survival mechanism is hijacked by the mutated genes and exploited to counter medical treatment. Abnormalities in cell cycle control are a characteristic of cancer, and this is accompanied by uncontrolled growth. Typically, most cancer cells deactivate the pathways to apoptosis and simultaneously heighten the effects of the cell proliferation and growth pathways. The overall goal of learning the structure is to investigate the genes instrumental in inducing resistance to cell death in melanoma.

Dataset:

The National Center for Biotechnology Information provides the values of gene expressions from a series of experiments on melanoma cell lines. The dataset contains 50,000 genes, of which 77 have been identified as genes are of interest. The value of the gene expression level indicates how much of the gene has been transcribed or activated. A low value means that the gene has been inhibited or inactive. Since it only matters if a gene is on or off, k-means was used to binarize this data (one centroid for on and one for off). The final output is a 77 × 51 matrix of 0s and 1s indicating which gene was on or off for which experiment (51 experiments).

Learning the Network:

A Bayesian network is used to represent the relationship between the genes. The nodes represent the genes and the edges represent activation or inhibition of children genes based on the parent genes. The model tries to model how the expression of parent genes affects the children.

Check all possible pairs and whether they should be connected. If $P(X) = X_i \perp X_j \mid \underline{U}$ then X_i and X_j cannot be directly connected.

If $X \perp Z \mid \underline{U}_{X,Z}$ then it does not contain the direct edge X-Z.

Check that no child has two or more parents that without dependencies between them. If that is the case for X-Y-Z, then Y cannot be in the network.

Using the Python package, *pgmpy* and the function *testconditional*, the test for conditional independence was conducted using chi-squared with deviance measured by:

$$d_{X,Y}(D) = \sum_{\forall x,y} \frac{(M[x,y] - M\hat{p}(x)\hat{p}(y))^2}{M\hat{p}(x)\hat{p}(y)}$$

Conditional mutual information was also then used as an independence test:

$$I(X;Y|Z) = \sum_{\forall z} p(z) \sum_{\forall x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

Next, steps were taken to avoid cycles and assign directions to any undirected edges. Edges were directed so that no cycles were created, and no child has two or more parents that without dependencies between them. If there were multiple undirected edges, then the direction of how one affected the other was also taken into account.

Final, to estimate the Bayesian parameters, the prior distribution P(theta) was assumed to follow:

$$P(\theta|D) = \prod_i \prod_{Pa(X_i)} P(\theta_{X_i|Pa(X_i)}|D)$$

Where D is the dataset given for X_i. A set of Dirichlet priors was used where each prior had size 2 (since it was binary). If, $P(\theta_{X|U})$ is a Dirichlet prior with hyperparameters $[\alpha_{x^1|u}, \alpha_{x^0|u}]$ where U = Pa(X), then the posterior is a Dirichlet distribution with hyperparameters $[\alpha_{x^1|u} + M[x^1, u], \alpha_{x^0|u}] + M[x^0, u]$

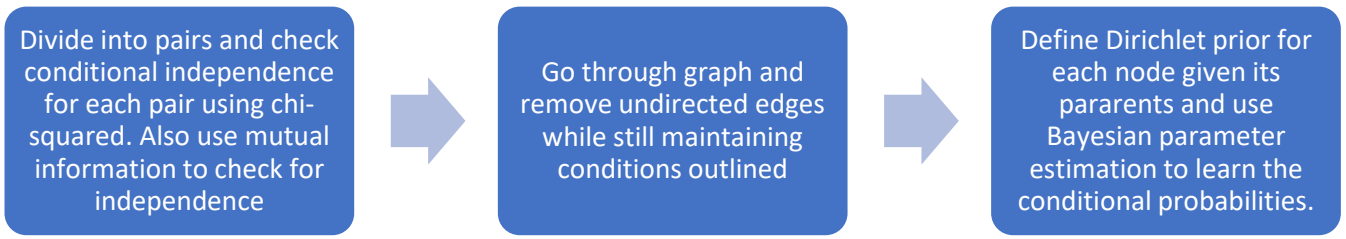
The conditional probabilities are calculated using the definition of the posterior:

$$P(X_i[M+1] = x_i | U[M+1] = u), D = \frac{\alpha_{x_i|u} + M[x_i, u]}{\sum_i \alpha_{x_i|u} + M[x_i, u]}$$

Finally, the ultimate learning objective when deriving the network was to maximize the BIC score to get the best model:

$$score_{BIC}(G) = \log L(\hat{\theta_G} : D) - \frac{\log M}{2} Dim[G]$$

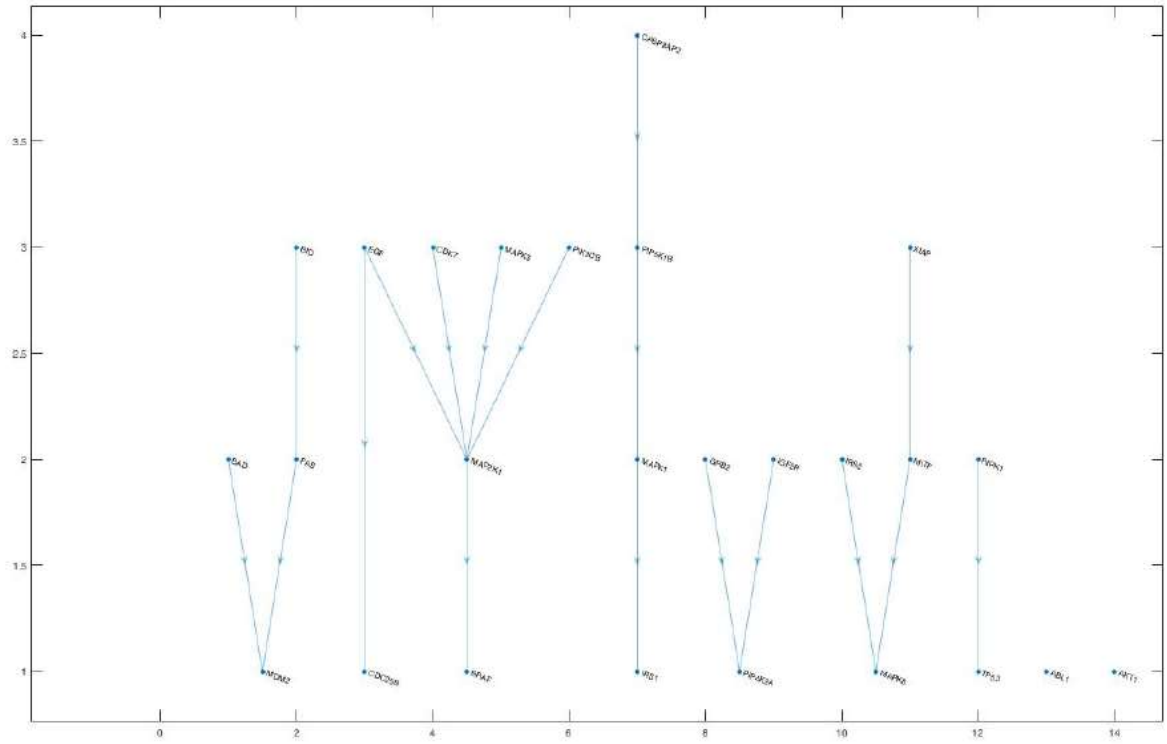
Where Dim[G] is the number of independent parameters in the model (which will depend on the predicted graph). The best BIC score/model is chosen through several trials.



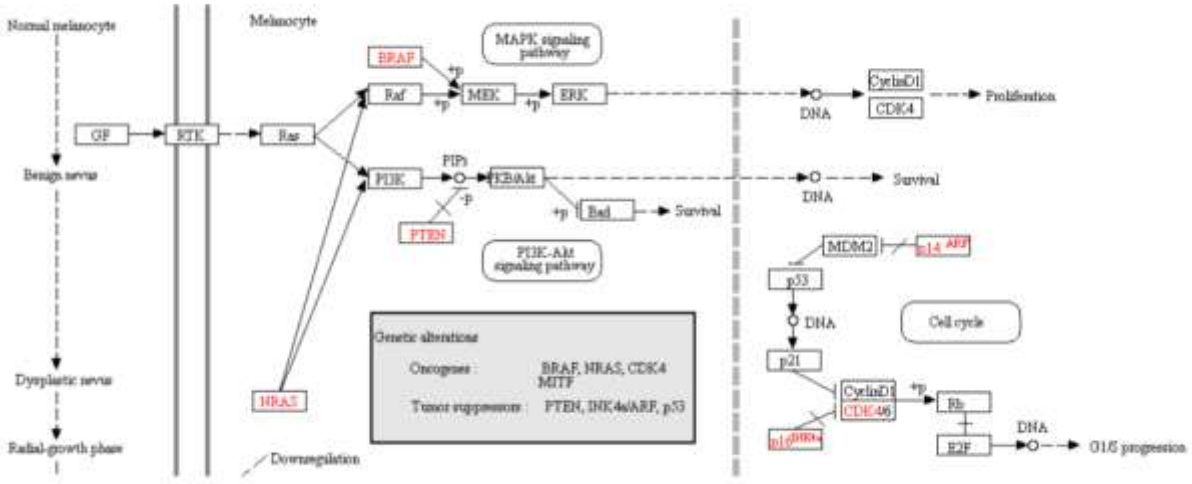
Some equations and process cited from:
http://www.cs.technion.ac.il/~dang/journal_papers/heckerman1995learning.pdf

Results:

The predicted Bayesian Network is shown in the figure below.



The experimentally-verified network (through laboratory experimentation) is shown here:



As we can see, the predicted Bayesian network is much sparser than the original network with fewer genes and less complicated relationships. That being said comparing the two networks it can be seen that the relationships learnt from the data by the model are correct and the order of all parent and children genes are in the correct place and that no two genes are incorrectly placed. While several genes/parts of the pathway were cutout, the computational approach was able to come up with a network that was correct. Moreover, this probabilistic approach is far more efficient than running dozens of long-lasting and time-consuming experiments to experimentally determine the pathway. Thus, this approach showed merit.

As a final step, inference was conducted (i.e. variable elimination) using *pgmpy* to see the effects of different genes conditioned on each other and to verify the network was working. Based on laboratory research, it was known that the growth factors genes matched up to nodes 21, 32, 48, cell death genes corresponded to nodes 73 and 74, and cell growth indicators matched up to nodes 11, 14, and 15. By varying the gene expressions of the growth factor genes/cell death genes, we can observe the changes in the probability of the cell growth indicators (0 or 1):

```
phi_query = inference.query(['11', '14','15'],evidence={'73':0,'74':0,'21':1,'32':1,'48':1})
phi_copy = phi_query.copy()
for X in phi_copy:
    print((X,phi_copy[X].values))
```

```
('11', array([ 0.48199318,  0.51800682]))
('15', array([ 0.30188679,  0.69811321]))
('14', array([ 0.50943396,  0.49056604]))
```

```
phi_query = inference.query(['11', '14','15'],evidence={'73':1,'74':1,'21':0,'32':0,'48':0})
phi_copy = phi_query.copy()
for X in phi_copy:
    print((X,phi_copy[X].values))
```

```
('11', array([ 0.57012259,  0.42987741]))
('15', array([ 0.30188679,  0.69811321]))
('14', array([ 0.50943396,  0.49056604]))
```

```
phi_query = inference.query(['11', '14','15'],evidence={'74':1})
phi_copy = phi_query.copy()
for X in phi_copy:
    print((X,phi_copy[X].values))
```

```
('11', array([ 0.52248333,  0.47751667]))
('15', array([ 0.30188679,  0.69811321]))
('14', array([ 0.50943396,  0.49056604]))
```

As we can see when the growth factor genes are expressed the probability that node 11 (cell growth indicator) will be 1 is much higher than when the cell death genes are expressed. Therefore, this shows that the network is also able to maintain some form of correct causal relationships.

Conclusion:

While this project was not perfect and was not able to completely learn the Bayesian network, it was able to get a reasonable approximation that had appropriate causal relationships. With future advances in the field, this approach may be promising.

Citations:

http://robotics.stanford.edu/~koller/NIPStut01/tut6.pdf

http://www.cs.technion.ac.il/~dang/journal_papers/heckerman1995learning.pdf

NCBI GEO database GSE31534 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31534

Kanehisa M, Sato Y, Kawashima M, Furumichi M, & Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016;44(Database issue) pp. :D457-D462

Wang. L, Hurley. DG, Watkins. W, Araki. H, Tamada. Y, Muthukaruppan. A, Ranjard. L, Derkac. E, Imoto. S, Miyano. S, Crampin. EJ, & Print. CG.(2012) Cell cycle gene networks are associated with melanoma prognosis. PLoS One 7(4), e34247