CSC2506 Assignment 2

Matthew Scicluna

2017-01-13

Graphical model distributions

Directed Graph

1. The statement MUST BE TRUE. We show that x_4 is d-seperated from x_5 conditioned on $x_3 \Rightarrow x_4 \perp x_5 \mid x_1, x_2, x_3$. This is clear though since there are only 2 paths, so we just have to check them both. $x_4 \to x_6 \to x_5$ doesn't work since x_6 explains away x_4 and x_5. $x_4 \to x_3 \to x_5$ doesn't work either since x_3 is the common cause of x_4 and x_5.
2. The statement MUST BE TRUE. Once again we show that x_3 is d-seperated from x_7 conditioned on x_4 and x_5. There are only two paths to check, $x_3 \to x_4 \to x_6 \to x_7$ and $x_3 \to x_5 \to x_6 \to x_7$. x_4 and x_5 block each path respectively as they are being conditioned on and they form chains between x_3 and x_6.
3. The statement MUST BE TRUE. There is only 1 path $x_1 \to x_3 \to x_2$ which is blocked since x_3 and all of its is descendants are not conditioned on, meaning it explains away x_1 and x_2.
4. The statement COULD BE TRUE. We can only determine that x_4 and x_5 are not d-connected, but this is not sufficient to conclude that they are not independant. To show that x_4 and x_5 are not d-connected, take the path $x_4 \to x_6 \to x_5$.

Undirected Graph

1. The statement CANNOT BE TRUE. It is enough to find a path from x_4 to x_5 that doesn't pass through x_1, x_2 or x_3. One such path is $x_4 \to x_7 \to x_5$.
2. The statement MUST BE TRUE. Every path between x_3 and x_7 passes through either x_4 or x_5.
3. The statement CANNOT BE TRUE. x_1 and x_2 share an edge and so cannot be independant!
4. The statement CANNOT BE TRUE. A path that goes from x_4 to x_5 without passing x_3 or x_6 is $x_4 \to x_7 \to x_5$

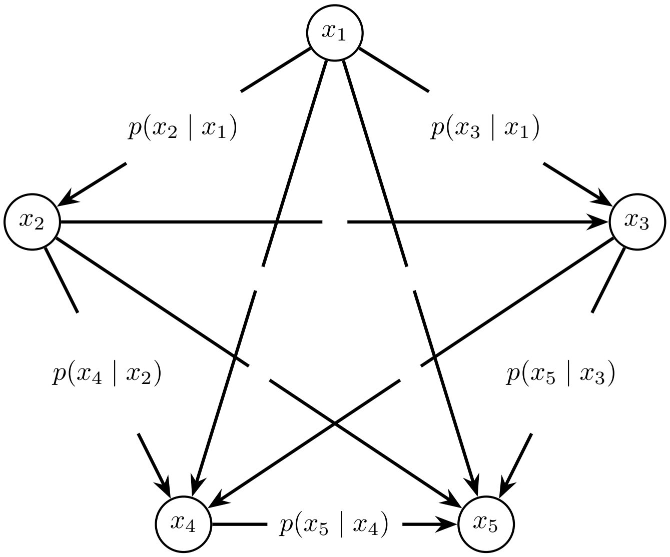
Complete graphs

1. Consider the following 5 variables ordered arbitrarily as x_1, x_2, x_3, x_4, and x_5. We wish to draw a directed graphical model which can capture any joint distribution and is acyclic. This means that we want our graph to capture the following dependancy:

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)P(x_2 \mid x_1)p(x_3 \mid x_2, x_1)p(x_4 \mid x_3, x_2, x_1)p(x_5 \mid x_4, x_3, x_2, x_1)$$

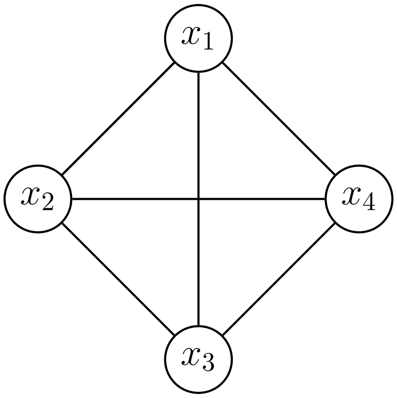
From this we see we must capture all the pairwise dependancies: $P(x_i \mid x_j) \quad \forall i > j \in \{1, 2, 3, 4, 5\}$

We see that the following graph models this dependancy, given our ordering.



2. We see that no edges Can be added to this graph since it is complete (every pair of nodes is connected by an edge). If any edge is removed from this graph we see that we will be enforcing certain conditional independancies which would limit the class of distributions that this graph can convey. This is since it would only be able to model joint distributions with particular conditional independancies between the variables.

3. We seek to draw an undirected graphical model on four variables which can capture any joint distribution. We also want to list all the maximal cliques. Again we see that we need edges between every pair of vertices or else we will not be able to capture every possible joint distribution (since missing edges imply that we can only model joint distributions with conditional independances). We see the following graph satisfies this since it is complete. Note that this graph is complete meaning that it has exactly one maximal clique - the entire graph itself.



4. Once again, no edges can be added since the graph is complete, and none can be removed or else the graph will not be able to model every possible joint distribution, and only ones with particular conditional independances between the variables.

5. [Bonus] The claim is TRUE. We prove it by showing that every graphical model which can capture any joint distribution must be complete and that a complete graph has $\frac{k^2}{2} - \frac{k}{2}$ edges.

- Firstly, if a graphical model is not complete than it will be missing edges $\Rightarrow$ it can only model joint distributions with conditional independancies imposed between some of it's variables. So clearly the graph must be complete.

- Secondly, a complete graph has edges between every possible pair of nodes. This means if the graph has k nodes it has $\binom{k}{2}$ edges. Note that $\binom{k}{2} = \frac{k!}{(k-2)!2} = \frac{k(k-1)}{2} = \frac{k^2}{2} - \frac{k}{2}$. So the claim holds, as needed.

Learning Undirected Models

We are given the following

$$p(x \mid \theta) = exp\left\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta)\right\}$$

where

$$\Phi(\theta) = log\left\{\sum_{x} exp\left\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right\}\right\}$$

MLE for Fully Disconnected pairwise binary MRF

Note that the maximum likelihood (ML) estimates of the node parameters θ_s when E = $\emptyset$ is as follows:

$$\ell(\theta \mid x) = lnp(x \mid \theta) = ln\left\{exp\left\{\sum_{s \in V} \theta_s x_s - \Phi(\theta)\right\}\right\} = \sum_{s \in V} \theta_s x_s - \Phi(\theta)$$

So for data points $x_s^{(i)}$, $i = 1, 2, \dots N$, each with features $s = 1, 2, \dots, k$.

$$\frac{\partial \ell(\theta_s \mid x)}{\partial \theta_s} = \frac{\partial}{\partial \theta_s} \sum_{i=1}^{N} lnP(x_s^{(i)} \mid \theta) = \sum_{i=1}^{N} \frac{\partial}{\partial \theta_s} \sum_{s \in V} \theta_s x_s^{(i)} - \Phi(\theta) = 0$$

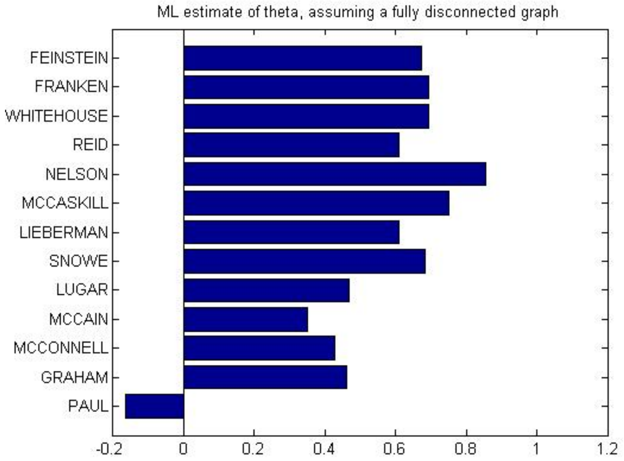


Figure 1: Plot of each Senators theta parameter, assuming a fully disconnected graph

and using the linearity of the differentiation operator and that $\frac{\partial}{\partial \theta_s}\Phi(\theta) = E(x_s \mid \theta)$ since x is in the exponential family we have that, after rearranging terms:

$$\bar{x_s} = \frac{1}{N}\sum_{i=1}^{N} x_s^{(i)} = E(x_s \mid \theta)$$

Finally a plot of the θ_s for each senator can be found in figure 1.

To compute the expectation we see that:

$$\begin{aligned}
E(x_s \mid \theta) &= \sum_{x_s=0}^{1} \frac{exp(\sum_{i=1}^{13} x_i\theta_i)x_s}{\sum_{x_s=0}^{1} exp(\sum_{i=1}^{13} x_i\theta_i)} \\
&= \sum_{x_s=0}^{1} \frac{\prod_{j\neq s} exp(x_j\theta_j)exp(x_s\theta_s)x_s}{\sum_{x_s=0}^{1}\prod_{j\neq s} exp(x_i\theta_i)exp(x_s\theta_s)} \\
&= \sum_{x_s=0}^{1} x_s \frac{e^{x_s\theta_s}}{1+e^{x_s\theta_s}} \\
&= \frac{e^{\theta_s}}{1+e^{\theta_s}}
\end{aligned}$$

And combining this yields:

$$\bar{x_s} = \frac{e^{\theta_s}}{1+e^{\theta_s}} \Rightarrow \theta_s = ln(\frac{\bar{x_s}}{1-\bar{x_s}})$$

Gradient of the Log Likelihood of This Model

First we find the sufficient statistics. Using that $x^{(i)}$ is from the exponential family, we see that:

$$
\begin{aligned}
P(Data) = \prod_{i=1}^{N} P(x^{(i)}) &= \prod_{i=1}^{N} exp\Big\{\sum_{s\in V}\theta_s x_s^{(i)} - \sum_{(s,t)\in E}\theta_{st}x_s^{(i)}x_t^{(i)} - \Phi(\theta)\Big\} \\
&= exp\Big\{\sum_{i=1}^{N}\sum_{s\in V}\theta_s x_s^{(i)} - \sum_{i=1}^{N}\sum_{(s,t)\in E}\theta_{st}x_s^{(i)}x_t^{(i)} - N\Phi(\theta)\Big\} \\
&= exp\Big\{\sum_{s\in V}\theta_s \sum_{i=1}^{N} x_s^{(i)} - \sum_{(s,t)\in E}\theta_{st}\sum_{i=1}^{N}x_s^{(i)}x_t^{(i)} - N\Phi(\theta)\Big\}
\end{aligned}
$$

So the sufficient statistics are $\sum_{i=1}^{N} x_s^{(i)}$ and $\sum_{i=1}^{N} x_s^{(i)}x_t^{(i)}$

Now, given this we derive an expression for the gradient of this log-likelihood with respect to θ_s and θ_{st}

For θ_s we have that:

$$
\begin{aligned}
\frac{\partial\ell(\theta_s \mid x)}{\partial\theta_s} &= \frac{\partial}{\partial\theta_s}\sum_{i=1}^{N}\sum_{s\in V}\theta_s x_s^{(i)} + \sum_{(s,t)\in E}\theta_{st}x_s^{(i)}x_t^{(i)} - \Phi(\theta) \\
&= \sum_{i=1}^{N} x_s^{(i)} - NE(X_s \mid \theta) \\
&= \sum_{i=1}^{N} x_s^{(i)} - NP(X_s = 1 \mid \theta)
\end{aligned}
$$

Since $x_s \in \{0, 1\}$

For θ_{st} we have that:

$$
\begin{aligned}
\frac{\partial\ell(\theta_{st} \mid x)}{\partial\theta_{st}} &= \frac{\partial}{\partial\theta_{st}}\sum_{i=1}^{N}\sum_{s\in V}\theta_s x_s^{(i)} + \sum_{(s,t)\in E}\theta_{st}x_s^{(i)}x_t^{(i)} - \Phi(\theta) \\
&= \sum_{i=1}^{N} x_s^{(i)}x_t^{(i)} - NE(X_sX_t \mid \theta) \\
&= \sum_{i=1}^{N} x_s^{(i)}x_t^{(i)} - NP(X_s = 1 \cap X_t = 1 \mid \theta)
\end{aligned}
$$

Computing ML Estimates of Model Parameters

Using the above gradient objective formulas and the MATLAB optimization package `L1General`, we computed the ML estimates of a fully connected pairwise graphical model. We used this on the full Senate voting record. Figure 2 contains a plot the log-likelihood of the model after each iteration.

Representational Power of this Model

The fully connected, pairwise graphical model estimated cannot represent any joint distribution on N binary variables? This is because it can only capture pairwise dependancies, but not higher order ones. An example of one such higher order dependancy is as follows: suppose Senator Paul always voted with Senator Franken, provided Senator Reid did too. Our pairwise model would be unable to capture this three-way interaction.

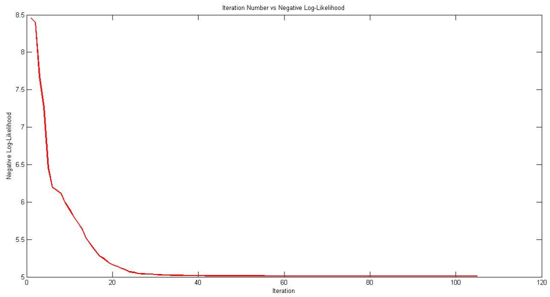


Figure 2: Plot negative log-likelihood of fully connected model

Binary Entropy Comparison

We compared the factorized model and the fully connected model by computing the binary entropy of each of the corresponding joint distributions. We found that the binary entropy was 12.2 and 7.23 respectively. Since the difference is not that large, this suggests that the voting patterns of Senators are largely unaffected by the voting patterns of other Senators.

Using Factorized Laplacian Priors

We explore what happens when we put a Laplacian prior on each θ_s and $\theta_{s,t}$. The prior has the following form:

$$P(\theta \mid \lambda) = \prod_{s \in V} \frac{\lambda_s}{2} e^{(-\lambda_s |\theta_s|)} \prod_{(s,t) \in E} \frac{\lambda_{st}}{2} e^{(-\lambda_{st} |\theta_{st}|)}$$

We see that:

$$\begin{aligned}
\ell(\theta \mid x, \lambda) &= \sum_{i=1}^{N} lnP(x^{(i)} \mid \theta, \lambda) \\
&= \sum_{i=1}^{N} lnP(x^{(i)} \mid \theta) + lnP(\theta \mid \lambda) \\
&= \sum_{i=1}^{N} lnP(x^{(i)} \mid \theta) + \sum_{s \in V} log\left(\frac{\lambda_s}{2}\right) - \lambda_s \mid \theta_s \mid + \sum_{(s,t) \in E} log\left(\frac{\lambda_{st}}{2}\right) - \lambda_{st} \mid \theta_{st} \mid \\
&\propto \sum_{i=1}^{N} lnP(x^{(i)} \mid \theta) - \bar{\lambda} \sum_{(s,t) \in E} \mid \theta_{st} \mid
\end{aligned}$$

Note that the last step follows if we impose the contstraints that $\lambda_s \to 0$ and $\lambda_{st} = \bar{\lambda}$.

Comparing Different Values of Our L1 Penalization Term

From the above we saw that, given $\lambda_s \to 0$ and $\lambda_{st} = \bar{\lambda}$, adding the Laplacian prior is equivalent to adding an identical L1 regularization penalty term on each of the λ_{st} terms. We used the votes for the first 400 bills

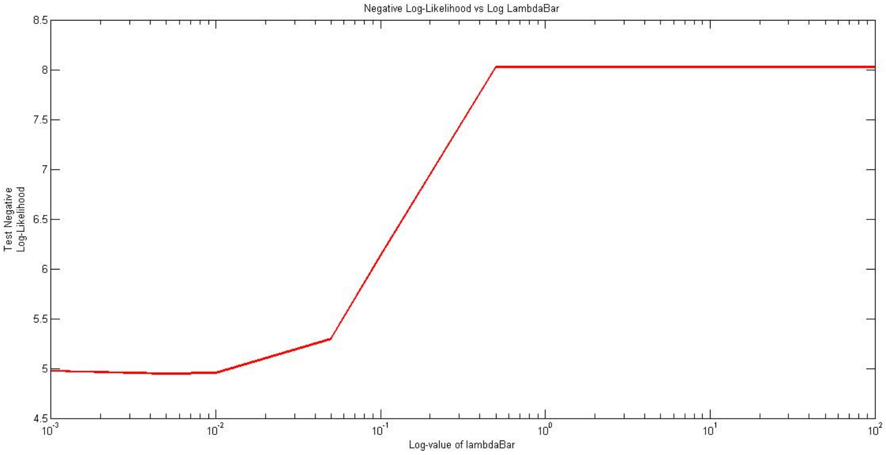


Figure 3: Plot of negative log-likelihood vs Log Value of Theta Bar Hyperparameter

as a training set, and the remaining 86 bills as a validation set to train seperate ising models, varying the value of $\bar{\lambda}$. We evaluated the log-probability of the validation data, and plotted this in figure 3.

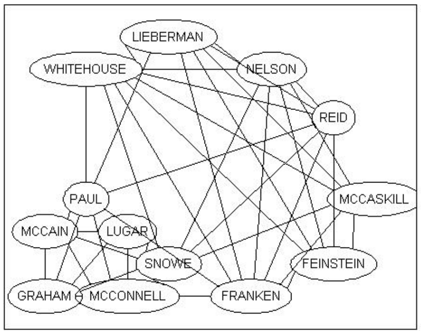


Figure 4: Graph which produced the highest validation log-likelihood

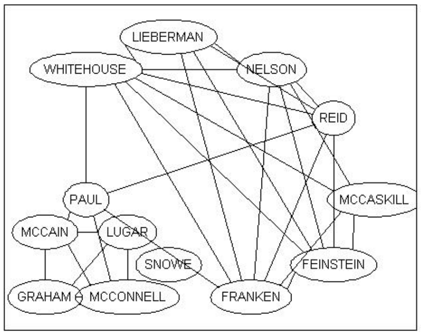


Figure 5: Graph with the smallest number of edges that was nevertheless connected

Associating Graphical Structures

We notice that setting $\bar{\lambda} = 0.5$ gave the model which produced the highest validation log-likelihood, and that $\bar{\lambda} = 0.1$ gave the graph with the smallest number of edges that was nevertheless connected. One way we could associate graphical structures to the models learned is by using the `drawlayout` function. This function creates a graph with distances of nodes based on the value of the θ parameters between the senators. We plot the aformentioned models in figures 4 and 5 respectively.

It is interesting to note that upon ℓ^1 regularization, edges are maintained disproportionately along party lines. This means that senators tend to vote along party lines, with the noticable exception of Senator Paul. Anyone who follows politics will notice that this isn't surprising, since this is something he is known for.