Project Report
Graphical Models LAB

Maurice Wenig
Friedrich Schiller University Jena
Germany
maurice.wenig@uni-jena.de

1 INTRODUCTION

Learning graphical models is a big aspect of machine learning. But for the parameters of the model to be learned, it is often assumed that the structure is already given, which includes information about dependencies and independencies between features. However in practice, that is not the case. So we must first learn the structure between our features, before we can learn the parameters. This is also the task of our assignment.

For our assignment, we focused on Gaussian Bayesian Networks (GBN). A GBN is a Bayesian Network where the conditional distributions of a feature given their parents is a Gaussian. Another constraint to the GBN is that the mean paramater of the Gaussian, that is the conditional distribution, can only depend linearly on the values of the feature's parents in the Network Graph. So with n as the total number of features, the problem of learning the distribution $p(x_i|x_{\mathrm{pa}(i)})$ for a feature $i \in [n]$, its value $x_i \in \mathbb{R}$, and the value of their parents $x_{\mathrm{pa}(i)} \in \mathbb{R}^{n_i}$ is equivalent to learning $\hat{\beta}_i^* \in \mathbb{R}, \hat{\beta}_i \in \mathbb{R}^{n_i}, \hat{\sigma}_i \in \mathbb{R}$ such that $p(x_i|x_{\mathrm{pa}(i)}) \sim \mathcal{N}(\hat{\beta}_i^* + \hat{\beta}_i^\top x_{\mathrm{pa}(i)}, \hat{\sigma}_i^2)$, This

2 METHODS

2.1 Hill Climbing

2.2 Tabu Walks

2.3 Random Restarts

2.4 Choice of optimized score

For the score function $p(G|D) \propto p(D|G) \cdot p(G)$, i used the approximation

$$
\begin{aligned}
p(D|G) &\approx p(D|G, \hat{\theta}) \\
\hat{\theta} &:= \arg\max_{\theta} p(D|G, \theta)
\end{aligned}
$$

and the prior

$$p(G) \propto \frac{1}{|E|^\lambda}$$

With this, the objective function can be decomposed into the sum of independent node scores and a regularization term.

$$
\begin{aligned}
\arg\max_G p(G|D) &= \arg\max_G \log p(D|G) + \log p(G) \\
&\approx \arg\max_G \log p(D|G, \hat{\theta}) - \lambda\,|E_G| \\
&= \arg\max_G \sum_{i\in[n]} S_i(G) - \lambda\,|E_G|
\end{aligned}
$$

where

$$S_i(G) := -|D| \cdot \log \hat{\sigma}_i - \frac{1}{2} \sum_{x \in D} \left(\frac{x_i - (\hat{\beta}_i^\top x_{\mathrm{pa}(i)} + \hat{\beta}_i^*)}{\hat{\sigma}_i} \right)^2$$

and $\hat{\theta} := \left(\hat{\beta}_i, \hat{\beta}_i^*, \hat{\sigma}_i\right)_{i\in[n]}$ are the respective ML estimates for

$$p(x_i|x_{\mathrm{pa}(i)}) \sim \mathcal{N}(\hat{\beta}_i^\top x_{\mathrm{pa}(i)} + \hat{\beta}_i^*, \sigma_i^2)$$

A derivation of this can be found in subsection A.1.

2.4.1 *Implications for Hill Climbing.*

Elementary changes (addition, substraction, flip of an edge) only influence local distributions. That means if we construct a graph G', where G' was made by applying an elementary change to an edge (u, v) in G, the comparison $p(G'|D) > p(G|D)$ can be evaluated locally:

$$
\begin{aligned}
& \sum_{i\in[n]} S_i(G') - \lambda\,|E_{G'}| > \sum_{i\in[n]} S_i(G) - \lambda\,|E_G| \\
\iff & \sum_{i\in[n]} S_i(G') - S_i(G) > \lambda(|E_{G'}| - |E_G|) \\
\iff & \underbrace{\sum_{i\in\{u,v\}} S_i(G') - S_i(G)}_{:=\Delta_S(G',G)} > \lambda\,\underbrace{(|E_{G'}| - |E_G|)}_{:=\Delta_E(G',G)}
\end{aligned}
$$

Where the second equivalence holds because $S_i(\cdot)$ only depends on node i and its parents. Therefore $S_i(G') = S_i(G)$ for $i \notin \{u, v\}$. Note that $\Delta_E(G', G)$ only depends on the type of change applied to G:

$$\Delta_E(G', G) = \begin{cases} 1 & \text{addition} \\ 0 & \text{flip} \\ -1 & \text{deletion} \end{cases}$$

$\Delta_S(G', G)$ measures the improvement of node scores when the change is applied to G.

Similarly, two alterations G_1 and G_2 of G can be compared:

$$
\begin{aligned}
& \sum_{i\in[n]} S_i(G_1) - \lambda\,|E_{G_1}| > \sum_{i\in[n]} S_i(G_2) - \lambda\,|E_{G_2}| \\
\iff & \Delta_S(G_1, G) - \Delta_S(G_2, G) > \lambda\,(\Delta_E(G_1, G) - \Delta_E(G_2, G))
\end{aligned}
$$

A derivation of this can be found in subsection A.2.

The interpretation of this is that a change has to bring an improvement of at least λ per edge in order to improve the whole objective function. But more importantly, this allow a very efficient comparison of two different changes, which we need to efficiently find the best possible change for [Hill Climbing] and [Tabu Walks]

2.5 Parameter Fine-Tuning

3 RESULTS

3.1 Generated Structures

3.2 Performance

Some performance analysis

3.3 Likelihood

Cross validation time.

Table 1: Error Score Comparison

Recommender	RMSE	MAE
user based	0.670	0.301
item based	0.569	0.222
factorization	0.512	0.207
hybrid	0.496	0.193

3.3.1 Submission Scores.

4 CONCLUSION

Some conclusion.

REFERENCES

A CALCULATIONS

A.1 Score Function

Here we derive

$$\max_G \log p(D|G,\hat{\theta}) = \max_G \sum_{i\in[n]} S_i(G)$$

$$\begin{aligned}
\max_G \log p(D|G,\hat{\theta}) &= \max_G \log \left[\prod_{x\in D} p(x|G,\hat{\theta})\right] \\
&= \max_G \log \left[\prod_{x\in D}\prod_{i\in[n]} p(x_i|x_{\mathrm{pa}(i)},\hat{\beta}_i,\hat{\beta}_i^{*},\hat{\sigma}_i)\right] \\
&= \max_G \sum_{x\in D}\sum_{i\in[n]} \log p(x_i|x_{\mathrm{pa}(i)},\hat{\beta}_i,\hat{\beta}_i^{*},\hat{\sigma}_i) \\
&= \max_G \sum_{x\in D}\sum_{i\in[n]} \log\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{1}{2}\left(\frac{x_i-(\hat{\beta}_i^{\top}x_{\mathrm{pa}(i)}+\hat{\beta}_i^{*})}{\sigma_i}\right)^2\right)\right] \\
&= \max_G \sum_{x\in D}\sum_{i\in[n]} \left[-\frac{1}{2}\left(\frac{x_i-(\hat{\beta}_i^{\top}x_{\mathrm{pa}(i)}+\hat{\beta}_i^{*})}{\sigma_i}\right)^2 - \log\hat{\sigma}_i - \frac{1}{2}\log 2\pi\right] \\
&= \max_G \sum_{i\in[n]} \underbrace{\left[-|D|\cdot\log\hat{\sigma}_i - \frac{1}{2}\sum_{x\in D}\left(\frac{x_i-(\hat{\beta}_i^{\top}x_{\mathrm{pa}(i)}+\hat{\beta}_i^{*})}{\hat{\sigma}_i}\right)^2\right]}_{=S_i(G)} \quad \square
\end{aligned}$$

Note that $\hat{\theta}$ $(\hat{\beta}_i,\hat{\sigma}_i)$ depends on G $(\mathrm{pa}(i))$.

A.2 Graph Comparison

For a graph G' that was made by applying one elementary change to a graph G, it holds that

$$\sum_{i\in[n]} S_i(G') = \sum_{i\in[n]} S_i(G) + \Delta_S(G',G) \tag{1}$$

and

$$|E_{G'}| = |E_G| + \Delta_E(G',G) \tag{2}$$

Therefore

$$\begin{aligned}
&\sum_{i\in[n]} S_i(G_1) - \lambda\left|E_{G_1}\right| > \sum_{i\in[n]} S_i(G_2) - \lambda\left|E_{G_2}\right| \\
\overset{(1),(2)}{\iff}\ & \Delta_S(G_1,G) - \lambda\Delta_E(G_1,G) > \Delta_S(G_2,G) - \lambda\Delta_E(G_2,G) \\
\iff\ & \Delta_S(G_1,G) - \Delta_S(G_2,G) > \lambda\left(\Delta_E(G_1,G) - \Delta_E(G_2,G)\right)
\end{aligned}$$