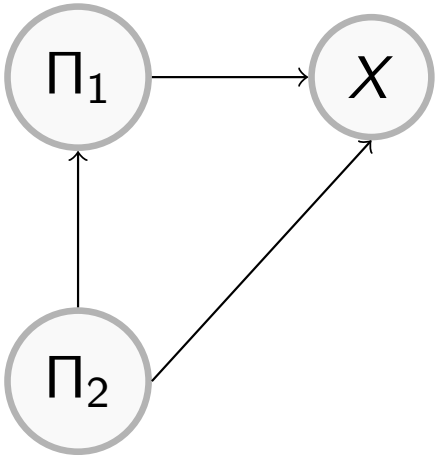
On the Dirichlet Prior and Bayesian Regularization

Бабкин Пётр

2024

Bayesian Network



$$p(x, \pi_1, \pi_2) = p(\pi_2)p(\pi_1|\pi_2)p(x|\pi_1, \pi_2)$$

$$p(D) = \prod_{x_i} p(x_i|\pi_i) \qquad i = 1, \ldots, n$$

$X \sim$ discrete distribution with probability vector $\{\theta_{x_i,\pi_i}\}$
N_{x_i,π_i} =number of realizations
$\{N_{x_i,\pi_i}\} \sim Mult(\{\theta_{x_i,\pi_i}\})$

Dirichlet Prior and Regularization of Parameters

Prior

$$p(\theta_{X_i|\pi_i}) = \frac{1}{B(\alpha_{X_i,\pi_i})} \prod_{x_i} \theta_{x_i|\pi_i}^{\alpha_{x_i,\pi_i}-1}$$

$\sum_{x_i,\pi_i} \alpha_{x_i,\pi_i} = \alpha$ – scale parameter

Likelihood

$$N_{X_i,\pi_i}|\theta_{X_i,\pi_i} \sim Mult(\theta_{X_i,\pi_i})$$

Posterior

$$\theta_{X_i,\pi_i}|N_{X_i,\pi_i} \sim Dir(\alpha_{X_i,\pi_i} + N_{X_i,\pi_i})$$

with $\alpha \to 0, \quad \theta_{X_i,\pi_i} \to MLE$

Regularization of Structure

Prior

$$p(m) > 0 \quad \forall m$$

Posterior

$$p(m|D) = \frac{p(D|m)p(m)}{p(D)}$$

Marginal Likelihood

$$p(D|m) = \int p(D|m,\theta)p(\theta|m)d\theta = \prod_{k=1}^{N}\prod_{i=1}^{n} \frac{N_{x_i^k,\pi_i^k}^{k-1} + \alpha_{x_i^k,\pi_i^k}}{N_{\pi_i^k}^{k-1} + \alpha_{\pi_i^k}}$$

with $\alpha \to 0, \quad \theta_{X_i,\pi_i} \to MLE$

Limit of Vanishing Scale-Parameter

Proposition 1. The marginal likelihood of a Bayesian network structure m vanishes in the limit $\alpha \to 0$ if the data D contain two or more different configurations.
The leading polynomial order is given by

$$p(D|m) \sim \alpha^{d_{EP}^{(m)}} \quad as \quad \alpha \to 0.$$

$$d_{EP}^{(m)} = \sum_i \left(\sum_{x_i,\pi_i} \mathbf{I}(N_{x_i,\pi_i}) - \sum_{\pi_i} \mathbf{I}(N_{\pi_i}) \right)$$

Vanishing Graph

Let edge $A \to B \in m^+$ and $A \to B \notin m^-$

$$d_{EDF} = d_{EP}^{m_+} - d_{EP}^{m_-}$$

Proposition 2. In the limit $\alpha \to 0$:

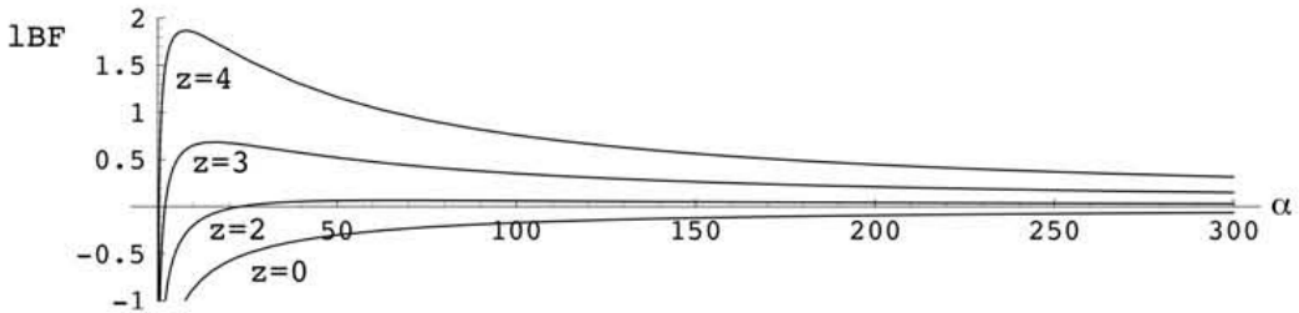
$$\log \frac{p(D|m^+)}{p(D|m^-)} = \begin{cases} -\infty & d_{EDF} > 0 \\ +\infty & d_{EDF} < 0 \end{cases}$$

Proposition is applicable to any edge in the graph.

Empty (complete) graph is assigned the highest relative Bayesian score when EDF are positive (negative).
Positive (negative) EDF relates to large (small) dataset.

Large Scale-Parameter and Example

$$\log \frac{p(D|m^+)}{p(D|m^-)} \to 0 \quad as \quad \alpha \to \infty$$



As $\alpha \to \infty$, edges with weak connections are favored over no connections.

Trade-off

Hence, the scale parameter α: of the Dirichlet prior determines the trade-off between regularizing the parameters vs the structure of the Bayesian network model.