KRR - Learning the parameters of a BN

Tudor Berariu, Alexandru Sorici

December 2018

We assume there is a real distribution $P_r(\mathbf{X})$ we do not have access to. Instead we either have a collection of samples from that distribution, or we are able to sample from it. In what follows we are concerned with learning some parametric model $P_\theta(\mathbf{X})$ that models as good as possible the real distribution $P_r(\mathbf{X})$.

We also assume the structure of a Bayesian Network that represents $P_\theta(\mathbf{X})$ to be known. We are left to learn the CPDs for each variable $X \in \mathbf{X}$. In what follows we will use $\mathbf{Y} \overset{not.}{=} Par(X)$ to denote the parents of X.

One way to ensure that we have good representations of the probabilities is learning a set of parameters $\theta_{X|\mathbf{Y}=\mathbf{y}}$ such that $P(X \mid \mathbf{Y}=\mathbf{y}) = \sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right)$ where $\sigma(x) = \left(1+e^{-x}\right)^{-1}$ is the *sigmoid* function. The sigmoid function is continous and differentiable in $\mathbb{R}$ and has a nice derivative $\sigma'(x) = \sigma(x)(1-\sigma(x))$.

Learning the model of a distribution. A common metric between distributions is the KL divergence. We therefore might use it to perform stochastic optimization of the parameters θ in order to increase the cross-entropy between the two distributions.

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, KL\left(P_r \,||\, P_\theta\right) = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x}\sim P_r}\left[\log\left(\frac{P_r(\mathbf{x})}{P_\theta(\mathbf{x})}\right)\right] = \underset{\theta}{\operatorname{argmin}}\, -\mathbb{E}_{\mathbf{x}\sim P_r}\left[\log\left(P_\theta(\mathbf{x})\right)\right] \tag{1}$$

Stochastic Optimization. We start with some random parameters $\theta^{(0)}$ and for each observed sample $\mathbf{x}^{(t)}$ we move in the direction opposed to the gradient in order to minimize our cost function.

$$KL\left(P_r \,||\, P_\theta\right) \approx \sum_{\mathbf{x}\sim P_r} \log P_\theta(\mathbf{x}) \approx \log P_\theta\left(\mathbf{x}^{(t)}\right) \tag{2}$$

$$\theta^{(t+1)}_{X|\mathbf{Y}=\mathbf{y}} \leftarrow \theta^{(t)}_{X|\mathbf{Y}=\mathbf{y}} + \eta \cdot \nabla_{\theta_{X|\mathbf{Y}=\mathbf{y}}} \log P_\theta\left(\mathbf{X}=\mathbf{x}^{(t)}\right) \tag{3}$$

Since the joint probability $P_\theta(\mathbf{X})$ is just a product of all CPDs, its logarithm becomes a sum.

$$\log P_\theta(\mathbf{X}) = \sum_{X\in\mathbf{X}} \log P_\theta\left(X \mid Par(X)\right) \tag{4}$$

For some specific parameter $\theta_{X|\mathbf{Y}=\mathbf{y}}$:

$$\nabla_{\theta_{X|\mathbf{Y}=\mathbf{y}}} \log P_\theta(\mathbf{X}=\mathbf{x}) = \nabla_{\theta_{X|\mathbf{Y}=\mathbf{y}}} \log P_\theta(X=x) = \begin{cases} \frac{\sigma'\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right)}{\sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right)} = 1-\sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right) & \text{if } x=1 \\ \frac{-\sigma'\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right)}{1-\sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right)} = -\sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right) & \text{if } x=0 \end{cases} = x - \sigma\left(\theta_{X|\mathbf{Y}=\mathbf{y}}\right) \tag{5}$$