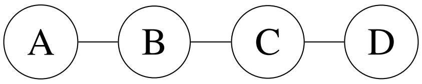
Graph Models and Boltzmann Machine

Arian Bastani - 400100073

Hossein Zarinkooh - 400109995

Q1. Potential functions represent the compatibility between random variables within a clique. These functions assign a non-negative score to each possible outcome in a clique, indicating the relative compatibility of that configuration.

Example: A path graph with 4 nodes:



Each node represents a random variable whose state space is {0,1}. The maximal cliques are: {A,B}, {B,C} and {C,D}. We can define a potential functions as followed and assign it to all of the maximal cliques:

$$\psi(X,Y) = \begin{cases} 2 & X = Y \\ 1 & \text{O.W.} \end{cases}$$

Q2. 3 examples of potential functions are provided:

$$\psi_1(x_i, x_j) = e^{-\beta|x_i - x_j|}$$

$$\psi_2(x_i, x_j) = \alpha \frac{1}{(x_i - x_j|)^2}$$

$$\psi_2(x_i) = \begin{cases} e^{\theta} & x_i = 1 \\ e^{-\theta} & x_i = 0 \end{cases}$$

β, α and θ are respectively the adjustable parameters of each function.

Q3. Gradient ascent is a method for maximizing a function. The idea is to take repeated steps in the direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest ascent.

Q4.

$$\frac{\partial \ln L(\theta|v)}{\partial\theta} = -\frac{\sum_h e^{-E(v,h)}\frac{\partial E(v,h)}{\partial\theta}}{\sum_h e^{-E(v,h)}} + \frac{\sum_{v,h} e^{-E(v,h)}\frac{\partial E(v,h)}{\partial\theta}}{\sum_{v,h} e^{-E(v,h)}}$$

$$p(v,h) = \frac{1}{Z}e^{-E(v,h)}$$

$$\Rightarrow \frac{\partial \ln L(\theta|v)}{\partial\theta} = -\frac{\sum_h p(v,h)\frac{\partial E(v,h)}{\partial\theta}}{p(v)} + \sum_{v,h}\frac{1}{Z}e^{-E(v,h)}\frac{\partial E(v,h)}{\partial\theta}$$

$$\Rightarrow \frac{\partial \ln L(\theta|v)}{\partial\theta} = -\sum_h p(h|v)\frac{\partial E(v,h)}{\partial\theta} + \sum_{v,h} p(v,h)\frac{\partial E(v,h)}{\partial\theta}$$

Q5. Here are some applications of Boltzman machine in ML:

1. Feature Learning

RBMs can be used to learn key features from unlabeled data, representing the data in a way that captures underlying similarities.

2. Generative Modeling

It can be used for generating new samples similar to the training data. Therefore it can be used for generative task and also recovering missing data points.

3. Recommendation Systems

RBMs can be used to model user preferences and make recommendations.

Q6.

$$p(v,h) = \frac{1}{Z}\exp\left(\sum_i\sum_j w_{ij}h_iv_j + \sum_j b_jv_j + \sum_i c_ih_i\right)$$

$$= \frac{1}{Z}\exp\left(\sum_j b_jv_j\right)\prod_{i=1}^{n}\exp\left(\left[\sum_j w_{ij}v_j + c_i\right]h_i\right)$$

$$\Rightarrow p(v) = \frac{1}{Z}\exp\left(\sum_j b_jv_j\right)\prod_{i=1}^{n}\left(1+\exp\left(\left[\sum_j w_{ij}v_j + c_i\right]h_i\right)\right)$$

$$p(v,H_i=1) = \frac{1}{Z}e^{\sum_j b_jv_j}e^{\sum_j w_{ij}v_j + c_i}\prod_{\substack{k=1\\k\neq i}}^{n}\left(1+\exp\left(\left[\sum_j w_{kj}v_j + c_k\right]h_k\right)\right)$$

$$\Rightarrow p(H_i=1|v) = \frac{p(v,H_i=1)}{p(v)} = \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)$$

Due to the symmetry of the RBM's energy function, we can derive the conditional probability P(v = 1 | h) by exchanging the roles of v and h, and using the corresponding biases (b_i instead of c_i).

$$\Rightarrow p(V_j=1|h) = \sigma(\sum_i w_{ij}h_i + b_i)$$

Q7. We use the result of Q4

$$\frac{\partial E(v,h)}{\partial w_{ij}} = -h_iv_j \Rightarrow \frac{\partial \ln L(\theta|v)}{\partial w_{ij}} = \frac{1}{l}\sum_{v\in S}(\sum_h p(h|v)h_iv_j - \sum_{v,h}p(v,h)h_iv_j)$$

$$= \frac{1}{l}\sum_{v\in S}(\mathbb{E}_{p(h|v)}[h_iv_j] - \mathbb{E}_{p(h,v)}[h_iv_j]) = \frac{1}{l}\sum_{v\in S}(v_jhp(H_i=1|v) - \mathbb{E}_{p(h,v)}[h_iv_j])$$

$$= \frac{1}{l}\sum_{v\in S}(v_j\,\sigma(\sum_j w_{ij}v_j + c_i) - \mathbb{E}_{p(h,v)}[h_iv_j])$$

$$\frac{\partial E(v,h)}{\partial b_j} = -v_j \Rightarrow \frac{\partial \ln L(\theta|v)}{\partial b_j} = \frac{1}{l}\sum_{v\in S}(\sum_h p(h|v)v_j - \sum_{v,h}p(v,h)v_j)$$

$$= \frac{1}{l}\sum_{v\in S}(v_j - \mathbb{E}_{p(v)}[v_j])$$

$$\frac{\partial E(v,h)}{\partial c_i} = -h_i \Rightarrow \frac{\partial \ln L(\theta|v)}{\partial c_i} = \frac{1}{l}\sum_{v\in S}(\sum_h p(h|v)h_i - \sum_{v,h} p(v,h)h_i)$$
$$= \frac{1}{l}\sum_{v\in S}(\mathbb{E}_{p(h|v)}[h_i] - \mathbb{E}_{p(h)}[h_i]) = \frac{1}{l}\sum_{v\in S}(\sigma(\sum_j w_{ij}v_j + c_i) - \mathbb{E}_{p(h)}[h_i])$$

Q8. It was proved in Q7

Q9. p_{data} is independent of model parameters.

$$\frac{\partial D(p_{data}||p_\theta)}{\partial \theta} = -\mathbb{E}_{p_{data}}[\frac{\partial \ln(p_\theta)}{\partial \theta}]$$
$$\frac{\partial D(p_\theta^k||p_\theta)}{\partial \theta} = \mathbb{E}_{p_\theta^k}[\frac{\partial \ln(p_\theta)}{\partial \theta}] + \frac{\partial p_\theta^k}{\partial \theta}\frac{\partial D(p_\theta^k||p_\theta)}{\partial p_\theta^k}$$

$$\Rightarrow \frac{\partial(D(p_{data}||p_\theta) - D(p_\theta^k||p_\theta))}{\partial \theta} \approx -\mathbb{E}_{p_{data}}[\frac{\partial \ln(p_\theta)}{\partial \theta}] + \mathbb{E}_{p_\theta^k}[\frac{\partial \ln(p_\theta)}{\partial \theta}]$$
$$= -\sum_h p(h|v_0)\frac{\partial E(v_0,h)}{\partial \theta} + \sum_h p(h|v_k)\frac{\partial E(v_k,h)}{\partial \theta}$$

when k is increased, p_θ^k approaches p_θ. Therefore $D(p_\theta^k||p_\theta) = 0$ and the objective of CD will be to estimate p_θ by p_{data}. Hence it's the same as ML estimator.

Q10. To compute the gradient of the log-likelihood (used in Maximum Likelihood estimation), it is computationally expensive to run the Markov Chain Monte Carlo (MCMC) process to full equilibrium. Contrastive Divergence (CD) offers a more efficient alternative.

Instead of converging to the stationary distribution, CD runs the chain for only k steps (e.g., k = 1) starting from a data sample **v**, resulting in $\mathbf{v^k}$. This sample is then used to approximate the gradient, which is subsequently used to update the model parameters.

Q11. We can use gaussian RBM to fit continues data. In this machine, we use the following energy function:

$$E(v,h) = \sum_j \frac{(v_j - \mu_j)^2}{2\sigma_j^2} - \sum_j (c_i h_i) - \sum_{i,j} \frac{v_j}{\sigma_j} h_i w_{ij}$$

We can scale our data so that: $\mu_i = 0,\ \sigma_i = 0.$