# پروژه فاز 1


## پرهام گیلانی – 400101859

## صدرا خنجری – 400101107


## بخش تئوری

1. توابع پتانسیل ($\psi$) در مدل میدان مارکوف (MRF) نشان‌دهنده تأثیر متقابل بین گره‌های یک گروه خاص (clique) در گراف است. این توابع نقش مهمی در توزیع احتمال مشترک متغیرها دارند و توزیع کلی را به صورت زیر فاکتوریزه می‌کنند:

$$p(x) = \psi_C(x_C) \prod_{C \in Cliques} \frac{1}{Z}$$

که در آن:

- Z یک ثابت نرمال‌سازی است.
- $\psi_C(x_c)$ توابع پتانسیل برای کلیک C است.
- $x_c$ مقادیر متغیرهای گره‌های کلیک C است.

این توابع معمولاً مثبت و غیرمنفی هستند و امکان محاسبه احتمال‌های شرطی را با استفاده از قانون مارکوف فراهم می‌کنند.

**مثال ساده**

فرض کنید گراف شامل سه گره $A, B, C$ است که به صورت زنجیره‌ای به هم متصل‌اند:
A با B و B با C متصل است.
فرض می‌کنیم توابع پتانسیل به شکل زیر تعریف شده‌اند:

$$\psi_{AB}(A,B) = e^{-\frac{AB}{2}}, \psi_{BC}(B,C) = e^{-0.3BC}$$

مقادیر ممکن برای $A, B, C$ باینری هستند، یعنی هرکدام می‌توانند 0 یا 1 باشند.

توزیع احتمال کلی به صورت زیر محاسبه می‌شود:

$$p(A,B,C) = \frac{1}{Z}\,\psi_{AB}(A,B)\psi_{BC}(B,C)$$

که در آن $Z$ ثابت نرمال سازی است و از مجموع همه مقادیر ممکن محاسبه می‌شود.

**محاسبات با مقادیر عددی**

تمام ترکیبات ممکن برای $A,B,C$

| A | B | C | $\psi_{AB}(A,B)$ | $\psi_{BC}(B,C)$ | $\psi_{AB}(A,B)\psi_{BC}(B,C)$ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | $e^0 = 1$ | $e^0 = 1$ | $1 \times 1 = 1$ |
| 0 | 0 | 1 | $e^0 = 1$ | $e^0 = 1$ | $1 \times 1 = 1$ |
| 0 | 1 | 0 | $e^0 = 1$ | $e^0 = 1$ | $1 \times 1 = 1$ |
| 0 | 1 | 1 | $e^0 = 1$ | $e^{-0.3} \approx 0.74$ | $1 \times 0.74 = 0.74$ |
| 1 | 0 | 0 | $e^0 = 1$ | $e^0 = 1$ | $1 \times 1 = 1$ |
| 1 | 0 | 1 | $e^0 = 1$ | $e^0 = 1$ | $1 \times 1 = 1$ |
| 1 | 1 | 0 | $e^{-0.5} \approx 0.61$ | $e^0 = 1$ | $1 \times 0.64 = 0.64$ |
| 1 | 1 | 1 | $e^{-0.5} \approx 0.61$ | $e^{-0.3} \approx 0.74$ | $0.64 \times 0.74 \approx 0.45$ |

$$Z = \sum_{A,B,C} \psi_{AB}(A,B)\psi_{BC}(B,C) = 1 + 1 + 1 + 0.74 + 1 + 1 + 0.64 + 0.45 = 7.8$$

برای مثال:

$$p(1,1,1) = \frac{1}{Z}\psi_{AB}(1,1)\psi_{BC}(1,1) = \frac{0.64 \times 0.74}{7.8} = 0.058$$

**نتیجه‌گیری**

این روش نشان می‌دهد چگونه توابع پتانسیل به کمک مقادیر عددی به محاسبه احتمال‌های میدان مارکوف کمک می‌کنند. اگر نیاز به محاسبات دقیق‌تر باشد، می‌توان تمامی ترکیبات را مشابه روش بالا محاسبه کرد.

2. توابع پتانسیل در میدان مارکوف به انواع مختلفی تقسیم می‌شوند که می‌توان به موارد زیر اشاره کرد:

**توابع پتانسیل گره‌ای (Node Potential)**

این توابع به متغیرهای منفرد در گره‌ها اعمال می‌شوند. به عنوان مثال، برای یک گره A، می‌توان تعریف کرد:

$$\psi_A(A) = e^{-\alpha A} \, , \alpha = constant$$

اگر A فقط مقادیر 0 یا 1 بگیرد و $\alpha = 0.7$:

$$\psi_A(0) = e^{-0\times 0.7} = 1 \, , \psi_A(1) = e^{-1\times 0.7} \approx 0.5$$

**توابع پتانسیل یالی (Edge Potential)**

این توابع تأثیر متقابل بین دو گره متصل را نشان می‌دهند. برای مثال، گره‌های متصل A و B:

$$\psi_{AB}(A,B) = e^{-\beta AB} \, , \beta = constant$$

اگر A,B مقادیر 0 یا 1 بگیرند و $\beta = 0.5$:

$$\psi_{AB}(0,0) = e^{-\frac{0\times 0}{2}} = 1 \, , \psi_{AB}(0,1) = e^{-\frac{0\times 1}{2}} = 1$$
$$\psi_{AB}(1,0) = e^{-\frac{1\times 0}{2}} = 1 \, , \psi_{AB}(1,1) = e^{-\frac{1\times 1}{2}} \approx 0.61$$

**توابع پتانسیل گروه خاص (Clique Potential)**

این توابع برای گروه‌های خاص (cliques) در گراف تعریف می‌شوند و تأثیر متقابل بیش از دو گره را نشان می‌دهند. به عنوان مثال، برای یک گروه خاص شامل سه گره: $A, B, C$

$$\psi_{ABC}(A,B,C) = e^{-\gamma(A+B+C)} \, , \gamma = constant$$

اگر $A, B, C$ مقادیر 0 یا 1 بگیرند و $\gamma = 0.2$:

$$\psi_{ABC}(0,0,0) = e^{-0.2\times(0+0+0)} = 1 \, , \psi_{ABC}(0,0,1) = e^{-0.2\times(0+0+1)} \approx 0.82$$
$$\psi_{ABC}(0,1,1) = e^{-0.2\times(0+1+1)} \approx 0.67 \, , \psi_{ABC}(1,1,1) = e^{-0.2\times(1+1+1)} \approx 0.55$$

توابع پتانسیل می‌توانند برای گره‌ها، یال‌ها، یا گروه‌های خاص تعریف شوند.

در هر مورد، مقدار آن‌ها به متغیرهای گره‌ها و ثابت‌های مشخص وابسته است.

محاسبات عددی این توابع به ما کمک می‌کند تا ارتباط متغیرها در میدان مارکوف را بهتر درک کنیم.

3. روش **MLE** پارامترهای مدل مارکوف را با بیشینه‌کردن احتمال داده‌های مشاهده شده تخمین می‌زند. تابع درست‌نمایی $\ln L(\theta|S)$ شامل توابع پتانسیل و ثابت نرمال‌سازی $Z(\theta)$ است. گرادیان برای بیشینه سازی محاسبه و از روش‌هایی مثل **Gradient Ascent** استفاده می‌شود. در مثال عددی، با داده‌های $S = \{(1,1), (1,0), (0,0)\}$، ثابت $Z(w) = 3 + e^{-w}$ و $\ln L(\theta|S)$ محاسبه شده و w بهینه‌سازی می‌شود.

4. گرادیان لگاریتم درست‌نمایی برای یک مدل مارکوف با متغیر پنهان به صورت زیر تعریف می‌شود:

$$\ln L(\theta|v) = \ln p(v|\theta) = \ln\left(\frac{1}{Z}\sum_{h} e^{-E(v,h)}\right)$$

که v متغیرهای مشاهده‌شده، h متغیرهای پنهان، E(v, h) انرژی سیستم، Z ثابت نرمال‌سازی است.

$$Z = \sum_{v,h} e^{-E(v,h)}$$

**محاسبه گرادیان**

$$\ln L(\theta|v) = \ln\left(\sum_{h} e^{-E(v,h)}\right) - \ln(\sum_{v,h} e^{-E(v,h)})$$

**مشتق نسبت به $\boldsymbol{\theta}$:**

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

**گرادیان بخش اول:**

$$\frac{\partial}{\partial\theta}\ln\left(\sum_{h} e^{-E(v,h)}\right) = \frac{1}{\sum_h e^{-E(v,h)}}\sum_{h}(-\frac{\partial E(v,h)}{\partial\theta}e^{-E(v,h)})$$

این عبارت را می‌توان به صورت زیر نوشت:

$$p(h|v) = \frac{e^{-E(v,h)}}{\sum_h e^{-E(v,h)}} \rightarrow \frac{\partial}{\partial\theta}\ln\left(\sum_{h} e^{-E(v,h)}\right) = -\sum_{h} p(h|v)\frac{\partial E(v,h)}{\partial\theta}$$

**گرادیان بخش دوم:**

$$\frac{\partial}{\partial\theta}\ln\left(\sum_{v,h}e^{-E(v,h)}\right)=\sum_{v,h}p(v,h)\frac{\partial E(v,h)}{\partial\theta}\ ,p(v,h)=\frac{e^{-E(v,h)}}{Z}$$

**ترکیب دو بخش:**

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

5. کاربرد های ماشین بولتزمن:

**یادگیری ویژگی‌های پنهان :(Unsupervised Learning)** ماشین بولتزمن می‌تواند برای یادگیری ویژگی‌های پنهان در داده‌های بدون برچسب استفاده شود. با استفاده از این مدل، می‌توان ساختارهای پنهان در داده‌ها را شناسایی و ویژگی‌های مهم را استخراج کرد.

**شبکه‌های مولد :(Generative Models)** ماشین بولتزمن یک مدل مولد است که می‌تواند برای تولید داده‌های جدید مشابه داده‌های آموزشی استفاده شود. به این صورت که می‌تواند نمونه هایی شبیه به داده‌های ورودی تولید کند، به ویژه در مسائلی مانند تولید تصاویر، متن یا صدا.

**شبکه‌های عصبی پنهان :(Deep Belief Networks - DBN)** ماشین بولتزمن یکی از اجزای اصلی شبکه‌های عصبی عمیق (Deep Networks) به نام شبکه‌های باور عمیق (DBN) است. درDBN ، از چندین لایه ماشین بولتزمن برای یادگیری ویژگی‌های پیچیده تر در سطوح مختلف داده‌ها استفاده می‌شود.

**مدل‌سازی توزیع‌های پیچیده :**ماشین بولتزمن قادر است توزیع‌های پیچیده و غیرخطی را مدل سازی کند. به عنوان مثال، در مدل‌سازی توزیع‌های پنهان یا توزیع‌های احتمال در داده‌هایی که دارای روابط پیچیده‌ای هستند، مانند داده‌های تصویری یا صوتی، می‌تواند مفید باشد.

**یادگیری تقویتی :(Reinforcement Learning)** در برخی از مسائل یادگیری تقویتی، ماشین بولتزمن می‌تواند برای انتخاب عمل‌ها یا سیاست‌های بهینه استفاده شود. به ویژه در مدل‌های دارای محیط‌های پیچیده و فضای حالت بزرگ، استفاده از ماشین بولتزمن برای مدل‌سازی توزیع‌های پیچیده و بهینه‌سازی انتخاب‌ها کاربرد دارد.

**کاهش ابعاد (Dimensionality Reduction):** از ماشین بولتزمن می‌توان برای کاهش ابعاد داده‌ها و استخراج ویژگی‌های مهم و مختصر از داده‌های پیچیده استفاده کرد. به طور خاص، مدل‌های مولد مانند ماشین بولتزمن می‌توانند به شبیه‌سازی ویژگی‌های پنهان داده‌های پیچیده بپردازند و ابعاد داده را کاهش دهند.

6. توزیع مشترک احتمالات بین نرون‌های مشاهده پذیر v و نرون‌های پنهان h به صورت زیر تعریف می‌شود:

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

$$E(v,h) = -\sum_{i,j} w_{ij} h_i v_j - \sum_j b_j v_j - \sum_i c_i h_i$$

**اثبات برای $P(H_i = 1|v)$**

برای محاسبه از تعریف توزیع شرطی استفاده می‌کنیم:

$$p(H_i = 1|v) = \frac{p(H_i = 1, v)}{p(v)}$$

$$p(H_i = 1, v) = \frac{1}{Z} \sum_{h, h_i = 1} e^{-E(v,h)} \ , p(v) = \frac{1}{Z} \sum_h e^{-E(v,h)} \rightarrow$$

$$p(H_i = 1|v) = \frac{\sum_{h, h_i = 1} e^{-E(v,h)}}{\sum_h e^{-E(v,h)}}$$

$$E(v,h) = -H_i\left(\sum_j w_{ij}v_j + c_i\right) + terms\ independent\ to\ H_i$$

پس $p(H_i = 1|v)$ به صورت یک توضیع برنولی نتیجه میشود.

$$p(H_i = 1|v) = \sigma(\sum_j w_{ij}v_j + c_i)$$

**اثبات برای $p(v_j = 1|h)$**

اثبات $p(v_j = 1|h)$ مشابه است، با این تفاوت که توزیع شرطی بدین گونه است:

$$p(v_j = 1|h) = \frac{p(v_j = 1, h)}{p(h)} \rightarrow p(v_j = 1|h) = \sigma(\sum_j w_{ij}v_j + c_i)$$

7. ابتدا انرژی کلی سیستم را یادآوری می‌کنیم:

$$E(v,h) = -\sum_{i,j} w_{ij}h_iv_j - \sum_j b_jv_j - \sum_i c_ih_i$$

**مشتق نسبت به $w_{ij}$**

فقط ترم اول انرژی شامل $w_{ij}$ است:

$$E(v,h) = -w_{ij}h_iv_j + (terms\ independent\ of\ w_{ij}) \rightarrow$$

$$\frac{\partial E(v,h)}{\partial w_{ij}} = -h_iv_j$$

جایگذاری در فرمول کلی مشتق:

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

به صورت ساده‌شده:

$$\frac{\partial \ln L(\theta|v)}{\partial w_{ij}} = < h_i v_j >_{data} - < h_i v_j >_{model}$$

اینجا:

$< h_i v_j >_{data}$ امید ریاضی تحت توزیع شرطی $p(h|v)$

$< h_i v_j >_{model}$ امید ریاضی تحت توزیع مدل $p(v,h)$

**مشتق نسبت به $b_j$**

فقط ترم دوم انرژی شامل $b_j$ است:

$$E(v,h) = -b_j v_j + \left(terms\ independent\ of\ b_j\right) \rightarrow$$

$$\frac{\partial E(v,h)}{\partial b_j} = -v_j$$

جایگذاری در فرمول کلی مشتق:

$$\frac{\partial \ln L(\theta|v)}{\partial b_j} = -\sum_h p(h|v)\left(-v_j\right) + \sum_{v,h} p(v,h)\left(v_j\right)$$

به صورت ساده‌شده:

$$\frac{\partial \ln L(\theta|v)}{\partial w_{ij}} = < v_j >_{data} - < v_j >_{model}$$

**مشتق نسبت به $c_i$**

فقط ترم سوم انرژی شامل $c_i$ است:

$$E(v,h) = -c_i h_i + (terms\ independent\ of\ c_i) \rightarrow$$

$$\frac{\partial E(v,h)}{\partial c_i} = -h_i$$

جایگذاری در فرمول کلی مشتق:

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

به صورت ساده‌شده:

$$\frac{\partial \ln L(\theta|v)}{\partial c_i} = < h_i >_{data} - < h_i >_{model}$$

8. **تعریف تابع هدف**

$$\ln L(\theta|v) = \ln P(v|\theta)$$

$$p(v|\theta) = \frac{1}{Z}\sum_h e^{-E(v,h)} \quad , Z = \sum_{v,h} e^{-E(v,h)} \rightarrow$$

$$\ln L(\theta|v) = \ln(\sum_h e^{-E(v,h)}) - \ln Z$$

**مشتق لگاریتم درست‌نمایی نسبت به $w_{ij}$**

$$\frac{\partial}{\partial w_{ij}} ln\left(\sum_h e^{-E(v,h)}\right) = \frac{1}{\sum_h e^{-E(v,h)}} \frac{\partial}{\partial w_{ij}} \sum_h e^{-E(v,h)}$$

$$\frac{\partial}{\partial w_{ij}} \sum_h e^{-E(v,h)} = \sum_h \frac{\partial}{\partial w_{ij}} e^{-E(v,h)} = -\sum_h e^{-E(v,h)} \frac{\partial E(v,h)}{\partial w_{ij}}$$

$$\frac{\partial E(v,h)}{\partial w_{ij}} = -v_i h_j \rightarrow \frac{\partial}{\partial w_{ij}} ln\left(\sum_h e^{-E(v,h)}\right) = \frac{\sum_h e^{-E(v,h)} v_i h_j}{\sum_h e^{-E(v,h)}} \rightarrow$$

$$\frac{\partial}{\partial w_{ij}} ln\left(\sum_h e^{-E(v,h)}\right) = \sum_h p(h|v) v_i h_j$$

**مشتق بخش دوم:**

$$\frac{\partial}{\partial w_{ij}} \ln Z = \frac{1}{Z} \frac{\partial Z}{\partial w_{ij}}$$

$$\frac{\partial Z}{\partial w_{ij}} = \sum_{v,h} \frac{\partial}{\partial w_{ij}} e^{-E(v,h)} = -\sum_{v,h} e^{-E(v,h)} \frac{\partial E(v,h)}{\partial w_{ij}}$$

$$\frac{\partial E(v,h)}{\partial w_{ij}} = -v_i h_j \rightarrow \frac{\partial Z}{\partial w_{ij}} = \sum_{v,h} e^{-E(v,h)} v_i h_j \rightarrow$$

$$\frac{\partial}{\partial w_{ij}} \ln Z = \frac{\sum_{v,h} e^{-E(v,h)} v_i h_j}{Z} = \sum_{v,h} p(v,h) v_i h_j$$

**ترکیب دو بخش**

$$\frac{\partial}{\partial w_{ij}} \ln L(\theta|v) = \sum_{h} p(h|v) v_i h_j - \sum_{v,h} p(v,h) v_i h_j$$

$$= < v_i h_j >_{data} - < v_i h_j >_{model}$$

اینجا:

$< h_i v_j >_{data}$ امید ریاضی تحت توزیع شرطی $p(h|v)$

$< h_i v_j >_{model}$ امید ریاضی تحت توزیع مدل $p(v,h)$

9. **تابع کلاسبندی متقاطع کلبرگ-لیبلر:(KL Divergence)**

برای دو توزیع احتمال P و Q، $D_{KL}(P||Q)$ به صورت زیر تعریف می‌شود:

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

**توزیع‌های مربوطه:**

$P_{data}$ توزیع داده‌های واقعی.

$P_{\theta}$ توزیع مدل بولتزمن پارامتری.

$P_k^{\theta}$ توزیع حاصل از k مرحله نمونه‌برداری گیبس.(Gibbs Sampling)

**تابع $CD_k(\theta, v_0)$**

$$CD_k(\theta, v_0) = -\sum_{h} P(h, v_0) \frac{\partial E(v_0, h)}{\partial \theta} + \sum_{h} P(h, v_k) \frac{\partial E(v_k, h)}{\partial \theta}$$

که $v_k$ خروجی حاصل از k مرحله نمونه برداری گیبس است.

**اثبات معادل بودن $D_{KL}(P_{data}||P_{\theta}) - D_{KL}(P_k^{\theta}||P_{\theta})$ با $CD_k(\theta, v_0)$**

**مرحله 1: مشتق $D_{KL}(P_{data}||P_{\theta})$**

$$D_{KL}(P_{data}||P_{\theta}) = \sum_{v} P(x) \log \frac{P_{data}(v)}{P\_\theta(v)} \rightarrow$$

$$\frac{\partial}{\partial \theta} D_{KL}(P_{data}||P_{\theta}) = -\sum_{v} P_{data}(v) \frac{\partial}{\partial \theta} \ln P_{\theta}(v) \ , P_{\theta}(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)}$$

$$\rightarrow \frac{\partial}{\partial \theta} \ln P_{\theta}(v) = \sum_{h} P(h|v) \frac{\partial E(v,h)}{\partial \theta} - \sum_{v,h} P(h, v) \frac{\partial E(v,h)}{\partial \theta}$$

**مرحله 2: مشتق $\boldsymbol{D_{KL}(P_k^{\theta}||P_{\theta})}$**

$$\boldsymbol{D_{KL}(P_k^{\theta}||P_{\theta})} = \sum_v P_k^{\theta}(v) \log \frac{P_k^{\theta}(v)}{P_{\theta}(v)} \rightarrow$$

$$\frac{\partial}{\partial \theta} \boldsymbol{D_{KL}(P_k^{\theta}||P_{\theta})} = -\sum_v P_k^{\theta}(v) \frac{\partial}{\partial \theta} \ln P_{\theta}(v) \rightarrow$$

$$\frac{\partial}{\partial \theta} D_{KL}(P_k^{\theta}||P_{\theta}) = \sum_v P_k^{\theta}(v) (\sum_{v',h} P(h,v') \frac{\partial E(v',h)}{\partial \theta} - \sum_h P(h|v) \frac{\partial E(v,h)}{\partial \theta})$$

**مرحله 3: ترکیب دو مشتق**

با ترکیب نتایج، مشتق ترکیبی برای اختلاف دو KL Divergence به صورت زیر است:

$$\frac{\partial}{\partial \theta} (D_{KL}(P_{data}||P_{\theta}) - D_{KL}(P_k^{\theta}||P_{\theta})) =$$

$$-\sum_v P_k^{\theta}(v) (\sum_h P(h|v) \frac{\partial E(v,h)}{\partial \theta} - \sum_{v',h} P(h,v') \frac{\partial E(v',h)}{\partial \theta})$$

این معادله دقیقاً با تعریف $CD_k(\theta, v_0)$ مطابقت دارد، زیرا:

بخش اول :($P_{data}$) توزیع داده.

بخش دوم :($P_k^{\theta}$) توزیع تخمینی حاصل از نمونه‌برداری گیبس.

پس تابع $CD_k(\theta, v_0)$ معادل با $D_{KL}(P_{data}||P_{\theta}) - D_{KL}(P_k^{\theta}||P_{\theta})$ است.

## 10.شرح الگوریتم k-step Contrastive Divergence

الگوریتم CD برای یادگیری مدل‌های انرژی مانند Restricted Boltzmann Machine (RBM) طراحی شده و بر نمونه‌برداری زنجیره مارکوف (MCMC) مبتنی است:

درCD ، به جای استفاده از تعداد زیادی گام مارکوف برای همگرا شدن به توزیع پایدار، از تعداد محدودی گام k استفاده می‌شود.

مراحل اصلی الگوریتم :

- **شروع با داده واقعی** :($v_0$)نمونه‌هایی از داده‌های آموزشی به مدل ورودی داده می‌شوند.
- **نمونه‌برداری مارکوف** :از حالت اولیه $v_0$، با استفاده از توزیع شرطی مدل $P(h|v)$ و $P(v|h)$ ، k گام نمونه برداری می‌شود تا $v_k$ به دست آید.
- **محاسبه گرادیان‌ها**:

$$\Delta\theta \propto E_{data}[\nabla E(v_0, h) - E_{model}[\nabla E(v_k, h)]]$$

که در آن $E(v, h)$ انرژی مدل است.

- **به‌روزرسانی پارامترها** :گرادیان محاسبه‌شده برای بهینه‌سازی پارامترها استفاده می‌شود.

**پیچیدگی الگوریتم و توجیه آن**

**پیچیدگی زمانی**:

الگوریتم CD برای هر گام k نمونه‌برداری، به محاسبه شرطی $P(h|v)$ و $P(v|h)$ نیاز دارد. این محاسبات به تعداد واحدهای نورونی در لایه‌های پنهان و نمایان بستگی دارد. پیچیدگی کلی برابر است با $O(k \times (m \times n))$ که m تعداد نورون‌های لایه نمایان و n تعداد نورون‌های لایه پنهان است.

**مزایا**:

برخلاف Gibbs Sampling کامل که به تعداد زیادی گام برای همگرا شدن نیاز دارد، CD با مقدار کوچک (مثلا $k = 1$) سرعت اجرا را به طور قابل توجهی افزایش می‌دهد و تقریب ساده‌ای ارائه می‌دهد که در عمل برای یادگیری پارامترها مؤثر است.

**معایب**:

با k کوچک، همگرایی کامل به توزیع مدل تضمین نمی‌شود و انتخاب مقدار k تأثیر مستقیمی بر کیفیت یادگیری دارد.

**انتظارات از مدل خروجی**

**بازسازی داده‌ها** :مدل باید بتواند داده‌هایی مشابه داده‌های آموزشی ورودی تولید کند.

**نزدیکی توزیع‌ها** :توزیع داده‌های تولیدی مدل به توزیع داده‌های آموزشی نزدیک تر می‌شود.

**کاهش خطای بازسازی** :تفاوت بین داده اصلی و داده بازسازی‌شده (reconstruction error) باید در طول زمان کاهش یابد.

**تولید نمونه‌های جدید** :مدل توانایی تولید داده هایی که شبیه داده‌های آموزشی باشند را پیدا می کند.

11. ماشین‌های بولتزمن (BM) و مدل‌های مشابه مانند ماشین‌های بولتزمن محدود(RBM) ، به طور سنتی برای کار با داده‌های باینری طراحی شده‌اند. در این مدل‌ها، متغیرهای نمایان (v) و پنهان (h) به صورت باینری در نظر گرفته می‌شوند و تابع انرژی مدل به این صورت است که به طور طبیعی برای داده‌های باینری بهینه می‌شود.
فرمول انرژی در این مدل‌ها معمولاً به شکل زیر است:

$$E(v,h) = -\sum_{i,j} w_{ij} h_i v_j - \sum_j b_j v_j - \sum_i c_i h_i$$

استفاده از ماشین‌های بولتزمن برای داده‌های پیوسته (که مقادیر آن‌ها می‌توانند هر عددی از یک بازه خاص باشند) نیاز به برخی تغییرات در مدل و نحوه محاسبات دارد. در اینجا چند پیشنهاد برای استفاده از این مدل‌ها برای داده‌های پیوسته آورده شده است:

**تغییر در توزیع‌ها:**

برای داده‌های باینری، توزیع‌های برنولی معمولاً برای مدل سازی استفاده می‌شود. اما برای داده‌های پیوسته، باید از توزیع‌های پیوسته مانند توزیع گوسی استفاده کرد و برای داده‌های پیوسته، متغیرهای v و h می‌توانند از توزیع‌های نرمال (گوسی) به‌جای توزیع‌های باینری پیروی کنند.

به‌عنوان مثال، فرض کنید که $v_i$ و $h_j$ به‌جای مقادیر باینری، از مقادیر پیوسته با توزیع نرمال $v_i \sim N(0, \sigma_v^2)$ پیروی کنند.

**تغییر در تابع انرژی:**

برای داده‌های پیوسته، می توان تابع انرژی را به گونه‌ای تغییر داد که شامل متغیرهای پیوسته باشد. برای مثال، تابع انرژی به صورت زیر در نظر گرفته می‌شود:

$$E(v,h) = -\sum_{i,j} w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

این تابع می‌تواند به‌گونه‌ای تغییر یابد که متغیرهای $v_i$ و $h_j$ به جای مقادیر باینری، مقادیر پیوسته بگیرند.

**استفاده از الگوریتم‌های نمونه‌برداری مناسب:**

برای داده‌های باینری، از نمونه برداری گیبز برای تولید نمونه‌های جدید استفاده می‌شود. برای داده‌های پیوسته، به دلیل اینکه توزیع‌های گوسی داریم، باید از نمونه برداری گوسی یا روش‌هایی مانند نمونه برداری از زنجیره‌ای گوسی مارکوف استفاده کرد.

(Gibbs sampling for Gaussian distributions)

می‌توان از الگوریتم‌هایی مانند Contrastive Divergence (CD) و Gibbs Sampling برای داده‌های پیوسته نیز استفاده کرد، اما به طور خاص برای داده‌های پیوسته، باید روش‌های نمونه برداری را تغییر داد تا با توزیع‌های پیوسته سازگار شوند.

**تغییرات مورد نیاز در الگوریتم‌ها:**

در الگوریتم Contrastive Divergence، اگر داده‌های ورودی به صورت پیوسته باشند، به جای استفاده از توزیع برنولی، باید از توزیع‌های گوسی برای نمونه برداری استفاده کرد. در الگوریتم Gibbs Sampling، در هر گام، برای متغیرهای پیوسته، از نمونه برداری گوسی استفاده می‌شود، نه نمونه برداری باینری.

پس ماشین‌های بولتزمن می‌توانند برای داده‌های پیوسته با تغییرات در توزیع‌ها و الگوریتم‌های نمونه برداری استفاده شوند. به طور خاص، استفاده از توزیع‌های گوسی به جای توزیع‌های باینری و تغییر در الگوریتم‌های Gibbs Sampling و Contrastive Divergence برای داده‌های پیوسته ضروری است.