


بررسی نمودار 1.

توابع توانایی، توابعی هستند که میان مارکهای گره‌های مربوط به هر clique را مشخص می‌کنند.

تجارت به توابعی به joint probabilities دارند و احتمال تمامی تغییراتی که می‌تواند رخ بدهد را می‌پوشاند. (در ضرب توابع)

توانایی Maximal clique نامی به دست می‌آید که در آن هیچ مارکوف را افزایش نمی‌دهد.

10



$$\phi(X_1, X_2) = \begin{cases} 1 & X_1=0, X_2=1 \\ 3 & X_1=1, X_2=1 \\ 1 & X_1=0, X_2=0 \\ 2 & X_1=1, X_2=0 \end{cases}$$

یک مثال ساده:

$$\phi(X_3) = \begin{cases} 2 & X_3=0 \\ 1 & X_3=1 \end{cases}$$

$X_1, X_2, X_3 \sim \text{Binary}$

15

$$\Rightarrow P(X_1, X_2, X_3) = \frac{1}{2} \phi(X_1, X_2) \phi(X_3), \quad Z = (1 \times 2) + (1) + (3 \times 2) + (3) + (2) + (1) + (4) + (2) = 24$$

$$\Rightarrow P(X_1, X_2, X_3) = \frac{\phi(X_1, X_2) \phi(X_3)}{24}$$

20

$$M - \text{Cliques} : \{ \{X_1, X_2\}, \{X_3\} \}$$

i.e. $P(1, 1, 0) = \frac{3 \times 2}{24} = 0.25$

بررسی نمودار 2.

25 یک نمونه از توابع توانایی، توابعی جدولی هستند، «مانند جدول» نامی نوعی، هر مقدار در جدول

یک بار است و ابعاد جدول معادل است با تعداد پارامترها.

(2) توانایی همواره پارامتری است، نوع تابع (ایزوس، چندجمله‌ای و...) و ضرایب آنهاست.

$$\phi(\vec{x}) = a_1 x_1 x_2 + a_2 x_2^2 + a_3 x_1 + a_4, \quad x_c = (x_1, x_2, x_3)$$

$$\phi(\vec{x}) = \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1}(\vec{x} - \vec{\mu})\right) \quad \text{«توزیع گاوسی»} \quad \text{«توزیع برادر متغیرهای تصادفی منفرد»}$$

متغیرهای این تابع تصادفی: میانگین μ (برادر $\vec{\mu}$) و ماتریس کوواریانس Σ

بررسی تئوری 3.

$$\theta_{n+1} = \theta_n + \alpha \left(\nabla L(\theta) \right)_{\theta=\theta_n} \quad \text{Gradient Ascent روشی برای بهینه کردن توابع است:}$$

اینکه در محاسبات در بازه مد نظر شروع می شود و با جمع ضرایب در بازه مد نظر با این مرحله به مرحله پیش می رود تا ضریب همگرا شود. به ضریب α ضریب یادگیری می گویند.

بررسی تئوری 4.

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

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

$$\text{پس داریم: } P(h|v) = \frac{e^{-E(v, h)}}{\sum_h e^{-E(v, h)}} \Rightarrow \frac{\partial}{\partial \theta} \ln \sum_h e^{-E(v, h)} = \sum_h \frac{\partial E(v, h)}{\partial \theta} P(h|v) \quad (1)$$

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

$$\Rightarrow \frac{\partial}{\partial \theta} - \ln \sum_{v, h} e^{-E(v, h)} = \sum_{v, h} \frac{\partial E(v, h)}{\partial \theta} P(v, h) \quad (2)$$

$$(1)(2) \Rightarrow \frac{\partial}{\partial \theta} \ln L(\theta, v) = - \sum_h P(h|v) \frac{\partial E(v, h)}{\partial \theta} + \sum_{v, h} P(v, h) \frac{\partial E(v, h)}{\partial \theta}$$

T.O

برش توری 5.

در بخش های که در مورد unsupervised می خواهیم ویژگی های استخراجی کنیم.

یادگیری ویژگی های غیر خطی برای این کار است همچنین در Deep Belief networks برای

یادگیری hierarchical ویژگی ها، مدل سازی توزیع احتمال داده های ساختار درونی دارند، در

سیستم های ترکیبی که در این رابطه تعاملی هستند و علاقه ای به محاسبات عددی در این زمینه دارند.

10

مورد استفاده از این.

برش توری 6.

$$E(v, h) = - \sum_{i=1}^n \sum_{j=1}^m w_{ij} h_i v_j - \sum_{j=1}^m b_j v_j - \sum_{i=1}^n c_i h_i, \quad P(v, h) \propto \exp(E(v, h))$$

$$P(h_i = 1 | v) = \frac{P(v, h_i = 1)}{P(v)} = \frac{\exp(\sum_{j=1}^m w_{ij} v_j - \sum_{j=1}^m b_j v_j - c_i)}{\exp(\sum_{j=1}^m w_{ij} v_j - \sum_{j=1}^m b_j v_j - c_i) + \exp(\sum_{j=1}^m w_{ij} v_j - \sum_{j=1}^m b_j v_j - c_i)}$$

تقسیم بر صورت

$$= \frac{1}{\exp(\sum_{j=1}^m w_{ij} v_j + c_i) + 1} \triangleq \sigma(\sum_{j=1}^m w_{ij} v_j + c_i)$$

این تابع سیگموئید است

$$P(v_j = 1 | h) = \frac{P(v_j = 1, h)}{P(v_j = 1, h) + P(v_j = 0, h)} = \frac{\exp(-\sum_{i=1}^n w_{ij} h_i - \sum_{i=1}^n c_i h_i - b_j)}{\exp(-\sum_{i=1}^n w_{ij} h_i - \sum_{i=1}^n c_i h_i - b_j) + \exp(-\sum_{i=1}^n c_i h_i)}$$

25

$$= \frac{1}{1 + \exp(\sum_{i=1}^n w_{ij} h_i + b_j)} = \sigma(\sum_{i=1}^n w_{ij} h_i + b_j)$$

بررسی تستی 7.

از بسخ سوال چهارم استفاده می کنیم:

$$\frac{\partial}{\partial \omega_{ij}} \ln(L(\theta|V)) = - \sum_h P(h|V) \frac{\partial E}{\partial \omega_{ij}} + \sum_{h,v} P(v,h) \frac{\partial E}{\partial \omega_{ij}}$$

$$5 \quad \frac{\partial E}{\partial \omega_{ij}} = h_i v_j \Rightarrow \frac{\partial}{\partial \omega_{ij}} \ln(L(\theta|V)) = + \sum_h P(h|V) h_i v_j - \sum_{h,v} P(v,h) h_i v_j$$

$$10 \quad \frac{\partial}{\partial b_j} \ln L(\theta|V) \Rightarrow \frac{\partial E}{\partial b_j} = v_j \Rightarrow \frac{\partial}{\partial b_j} \ln L(\theta|V) = \sum_h P(h|V) v_j - \sum_{v,h} P(v,h) v_j$$

$$15 \quad \frac{\partial}{\partial c_i} \ln L(\theta|V) = \sum_h P(h|V) c_i - \sum_{v,h} P(v,h) c_i$$

بررسی تستی 8.

15 اگر برای هر داده $v \in S$ داریم $\frac{\partial}{\partial \omega_{ij}} \ln L(\theta|V)$ و $\frac{\partial}{\partial b_j} \ln L(\theta|V)$ و $\frac{\partial}{\partial c_i} \ln L(\theta|V)$ را می توانیم به دست آوریم.

احتمال $v_i \in S$ ، $i = 1, 2, 3, \dots, l$

$$20 \quad \frac{\partial}{\partial \omega_{ij}} L(\theta|V_{v_i}) = \sum_h P(h|V) h_i v_j - \sum_{v,h} P(v,h) h_i v_j$$

برای هر تعریف امید ریاضی به صورت زیر در دسترس:

$$E[X] = \sum_{h,v} P(h,v) X$$

25 در نتیجه داریم:

$$\frac{\partial}{\partial \omega_{ij}} L(\theta|V_{v_i}) = E[h_i v_j | P(h|V)] - E[h_i v_j | P(h,v)]$$

$$\Rightarrow \frac{\partial}{\partial \omega_{ij}} L(\theta|V) = \frac{1}{l} \sum_{v \in S} E[h_i v_j | P(h|V)] - E[h_i v_j | P(h,v)] = \langle h_i v_j \rangle_{data} - \langle h_i v_j \rangle_{model}$$

$P(h|V)$: h در از روی V است.

$P(h,v)$: احتمال که مدل ارائه می کند

T.O

برای مقایسه

$$D_{KL}(P_{data} \parallel P_{\theta}) = \sum_{h,v} p_{data}(h,v) \log(P_{data}(h,v)) - \sum_{h,v} p_{data}(h,v) \log(P_{\theta}(h,v))$$

$$D_{KL}(P_{\theta}^k \parallel P_{\theta}) = \sum_{v,h} P_{\theta}^k(v,h) \log(P_{\theta}(v,h)) - \sum_{v,h} P_{\theta}^k(v,h) \log(P_{\theta}(v,h))$$

$$D_{KL,m} = \sum_{v,h} P_{\theta}^k(v,h) \log(P_{\theta}(v,h)) - \sum_{h,v} p_{data}(h,v) \log(P_{\theta}(h,v)) - \sum_{v,h} P_{\theta}^k \log P_{\theta}^k$$

مقدار $D_{KL,m}$ ثابت است و در این مورد تغییر نمی‌کند.

$$D_{KL,m} = H(P_{\theta}^k, P_{\theta}) - H(P_{data}, P_{\theta}) - H(P_{\theta}^k) + \text{constant}$$

$$P_{\theta}^k \sim P_{\theta} \Leftrightarrow H(P_{\theta}^k, P_{\theta}) = H(P_{\theta}^k)$$

$$CD_K$$

برای مقایسه CD_K از CD_K استفاده می‌کنیم.

$$\langle v, h, j \rangle_{data} = \langle v, h, j \rangle_{model}$$

$$CD_K$$

$$CD_K = E_{data} + E_{P_{\theta}} = (\langle v, h, j \rangle_{data} - \langle v, h, j \rangle_{model})$$

$$CD_K$$

بررسی آموزش 10

الگوریتم داده شده، به ازای مقدار اولیه داده، به ازای هر داده، به ترتیب زیر عمل می کنند:

5 لایه $z^{(0)}$ را برابر لایه $z^{(1)}$ داده قرار می دهیم: $z^{(0)} = z^{(1)}$

به ترتیب از الگوریتم Gibbs sampling استفاده می کنند تا به $z^{(k)}$ و $h^{(k)}$ برسند

سپس با استفاده از تخمین $contrastive-divergence$ و $gradient descent$ θ را

مقادیر منفی θ را به 0 می رسانند و $z^{(k)}$ و $h^{(k)}$ را $update$ می کنند.

این کار را برای هر $data$ (sample data) انجام می دهیم.

15 به ازای یک $z^{(k)}$ ، انتظار داریم که خروجی نمونه ای در برداری را بتواند بازسازی کند $z^{(0)} \sim z^{(k)}$

یعنی وزن های W مناسب به نحوی انتخاب شده باشند که خروجی $z^{(k)}$ بتواند $z^{(0)}$ را حدس زد.

تخمین در نهایت انتظار داریم که توزیع مدل حاصل به تدریج داده نزدیک شود و این به دلیل عمودیت های

20 ماده "Gibbs sampling" کامل به تدریج داده map می شود.

در یک $z^{(k)}$ به ازای هر i ، برای $k(m+n)$ برای $z^{(k)}$ $update$ mn را

25 روی فرم های اعمال می شود $n_y(k(m+n) + mn)$

که خیلی کوچکتر از 2^n حالت در رایانه ها می باشد.

پیش‌توری 17.

است. در این روش، برای داده‌ای پیوسته ممکن است، تغییراتی را با توزیع گوسی مدل

می‌کنیم. تغییراتی را با همان توزیع با هم می‌زنیم. تابع از روش جدید به شکل زیر تعریف می‌شود:

$$f(x, h) = \sum_{i=1}^n \frac{(x_i - b_i)}{2\sigma_i^2} \cdot f(x_i, h) \quad \text{و} \quad E(x, h)$$

2. به نام تغییراتی را با هم می‌زنیم. تابع از روش جدید به شکل زیر تعریف می‌شود:

این تغییراتی را با هم می‌زنیم. تابع از روش جدید به شکل زیر تعریف می‌شود:

از الگوریتم MCMC می‌توان برای داده‌های پیوسته هم استفاده کرد. فقط با این تفاوت که در Gibbs sampling

نمونه‌ها را با توزیع گوسی (استفاده می‌شود)، همچنین از متدهای پیوسته برای برآورد کردن پارامترها استفاده می‌شود.

برای تغییراتی را با هم می‌زنیم. تابع از روش جدید به شکل زیر تعریف می‌شود:

17. به نام تغییراتی را با هم می‌زنیم. تابع از روش جدید به شکل زیر تعریف می‌شود:

20

25