Subject 20A_Q3 Undirected Graphical Model / Date

Attendants Prebloc2 April 16th

Boltzmann Machine

$x \in \{0,1\}^N$ observation

$y \in \{0,1\}^M$, $z \in \{0,1\}^K$ } (latent variables)

$$P(x,y,z) = \frac{1}{Z} \exp(-E(x,y,z))$$

where $$Z = \sum_x \sum_y \sum_z \exp(-E(x,y,z))$$

and $$E(x,y,z) = \sum_{ijk} W_{ijk} x_i y_j z_k$$

Subject 20A_03 Date 2 April (Th

Attendants

MCMC

⇨ 直接サンプリングが難しい確率分布の代わりに
それを近似するサンプル列を生成する

⇓

周辺分布、同時分布や期待値などの積分計算を近似するために
用いられる

e.g. Gibbs Sampling は統計的推定やベイズ推定の手法として
頻繁に用いられる → Metropolis Hasting 法

Variational Bayes や EM アルゴリズムのような
統計的推定法のための決定論的な方法の代替法

導出：同時確率 $(x_1 \cdots x_n)$ から確率変数
$X = \{x_1 \cdots x_n\}$ を k サンプル得たい

n 次元の i 番目のサンプルを $X^{(i)} = \{x_1^{(i)} \cdots x_n^{(i)}\}$ とする

step0：初期値 $X^{(0)}$ を設定

step1：条件付き確率 $p(x_j | x_1^{(i)} \cdots x_{j-1}^{(i)}, x_{j+1}^{(i)}, \cdots x_n^{(i)})$
から i 番目のサンプル $x_j^{(i)}$ をサンプリング

step2：これ以下なら step1 を loop

Subject

Date 2.5

Attendants

$$p(x_j \mid x_1 \cdots x_{j-1}, x_{j+1}, \cdots x_n)$$

$$= \frac{p(x_1 \cdots x_n)}{p(x_1 \cdots x_{j-1}, x_{j+1} \cdots x_n)} = \frac{p(x_1 \cdots x_n)}{\int p(x_1 \cdots x_n)\, dx_j}$$

$$= \frac{1}{Z} p(x_1 \cdots x_n)$$

$$\propto p(x_1 \cdots x_n)$$

Subject 20A_Q3 Date 3

Attendants April 17th

これらのサンプル集合は全ての変数に
関する同時分布を近似する

サンプリング初期化の設定にEMアルゴ or ランダムに決めて
burn-in period を導即

また n番目ごとの値だけ用いるみたいな方法もある

自己相関を減少させる方法として

- 焼きなまし法 (Simulated Annealing)
- Collapsed Gibbs Sampling
- Block Gibbs Sampling

Subject 20A_Q3 Date 4

Attendants Boltzmann Machine April 17th

1. 状態の組み合せ $x = \{x_1 \cdots x_n\}$ によって値が変化する
   エネルギー関数Eを定義

2. Eが小さい状態ほど生起しやすいような
   $p(x|\theta)$ を定義

一般的なエネルギー関数

$$E(x,\theta) = -\sum_{i=1} b_i x_i - \sum_{(i,j)\in\mathcal{E}} w_{ij}\, x_i x_j$$

バイアス項

$\mathcal{E}$: エッジ集合

b_i : ユニットバイアス

w_{ij} : $x_i x_j$ の重み

$\mathcal{X}$: $\{x_1 \cdots x_n\}$

ユニットの状態の確率分布 $p(x|\theta)$

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp(-E)$$

この $Z(\theta)$ は分配関数

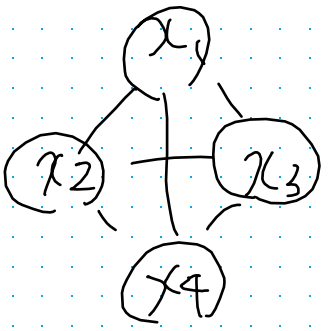
$$Z(\theta) = \sum_{x} \exp(-E) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} \exp(-E)$$

2^n 通りの加算

Subject 20A_Q3 Date 5

Attendants Boltzman Machine April 17th

1. 対数尤度を考える
2. 〃 をパラメータ微分して grad 導出
3. GDで θ を update



$$\ln L(\theta) = \sum_{n=1}^{N} \left(-E(x_n, \theta) - \ln Z(\theta) \right)$$

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(-E(x, \theta)\right)$$

$$E(x, \theta) = -\sum_{i=1}^{M} b_i x_i - \sum_{(i,j) \in E} w_{ij} x_i x_j$$

$$\frac{\partial \ln L(\theta)}{\partial w_{ij}} = \sum_{n=1}^{N} x_{n(i)} x_{n(j)} - N E_\theta[x_i x_j]$$

x_n の i番目の要素

$$x_1 = \begin{pmatrix} x_{1(0)} \\ x_{1(1)} \\ \vdots \\ x_{1(M)} \end{pmatrix}$$

分配関数

$$E_\theta[x_i x_j] = \sum_{x} x_i x_j p(x|\theta)$$

の厳密な計算は難しい

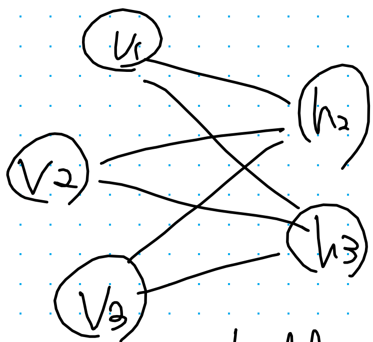
$$= \sum_{x} x_i x_j q(x) - N E_\theta[x_i x_j]$$

<データ平均> <モデル平均>

Subject 20A_Q3 Date 6

Attendants Boltzmann Machine w/ Hidden Layer April 17th

visible variable



hidden variable

$$z = \{v_1 \dots v_J, h_1 \dots h_K\}$$

$$p(z|\theta) = \frac{1}{Z(\theta)} \exp(-E(z,\theta))$$

↳ hidden layer なしと同じ形式

$p(z|\theta) = p(h, v|\theta)$ の h を周辺化する必要

$$p(v|\theta) = \sum_h p(v, h|\theta)$$

$$\frac{\partial \ln L(\theta)}{\partial w_{ij}} \propto \langle z_i z_j \rangle_{data} - \langle z_i z_j \rangle_{model}$$

Assumption each v_i, h_j doesn't have edges between v_i 同士, h_j 同士.

$$E(v, h, \theta) = -\sum_{i=1} a_i v_i - \sum_{i=1} b_i h_i - \sum_{(i,j)\in\mathcal{E}} w_{ij} x_i x_j$$

Subject: JoA-Q3 Date: 7

Attendants: Boltzman Machine April 17th

$$p(h|v,\theta) = \prod_j p(h_j|v,\theta)$$ ← conditional distribution

$$p(h_j=1|v,\theta) = \sigma\left(b_j + \sum_i w_{ij} v_j\right)$$

$$\frac{1}{N}\frac{\partial \ln L(\theta)}{\partial w_{ij}} = \sum_{n=1}^{N} v_n(i)\, p(h_i=1|v_n)$$

$$- \sum_{v,h} v_i h_i\, p(v,h|\theta)$$

こいつが不変

← Gibbs Sampling で近似

ユニット i の条件つき確率

$$p(x_i|x_{-i},\theta) = \frac{\exp\left\{\left(b_i + \sum_{j\in N} w_{ij} x_j\right) x_i\right\}}{1+\exp\left\{\left(b_i + \sum_{j\in N} w_{ij} x_j\right) x_i\right\}}$$

あるユニット x_i 以外の x_i につながるユニットが分かれば (x_{-i}) 以下を計算できる

Subject 2.A_Q3 Date 8.

Attendants Gibbs Sampling April 17th

$p(x_i|x_{-i},\theta)$ のサンプル生成

① $p_i \sim p(x_i=1|x_{-i},\theta)$ を計算

② $u_i \sim U(0,1)$

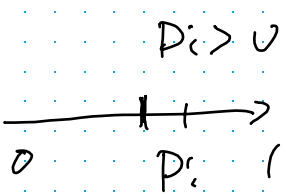
if $p_i > u_i$

$p_i = 1$?

else

$p_i = 0$?

p_i を利用してサンプリングする



p_i は大きいほど1になりやすい

これらによって生成された値は

$p(x_i|x_{-i},\theta)$ に従う.

実際のRBMでは Gibbs Sampling だと計算コスト高いので

Contrastive Divergence を用いる

Subject 20A_Q3 Date 9

Attendants Block Sampling / CD April 17th

vとhの条件付き確率を利用

vとhを交互にサンプリング

1. $v^0 \leftarrow$ random init

2. $v^0 \to h^0 \to v^1 \to h^1$ と交互にサンプリング

Contrastive Divergence

⇒ gibbs sampling の init が random ではなく

訓練サンプル ($v^0 = v_n$) を用いる

↓

1回操作する (T=1ともいう)

この方法は対数尤度でなく contrastive divergence を小さくなる最適化を行う.

これ(CD)に

- weight decay
- momentum
- sparse regularization

などが加えられる

Subject 20A_Q3 Date 10

Attendants

Persistent CD

通常のCDでは $v_0 = v_n$ とし

$v_0 \to h_0 \to v_1$ とサンプリングしていく

より10倍高速

PCDでは v_0 に前回のパラメータ更新時にサンプリングした v を用いる

深層学習勉強会 第6回

$X = \{V, H\}$

$$p(X=x|\theta) = p(V=v, H=h|\theta) = \frac{1}{Z(\theta)} \exp(-E(v,h,\theta))$$

$$p(x|\theta) = p(v|\theta) = \sum_h p(h,v|\theta)$$

h について周辺化した分布の最大推定をする

h は観測できないため

Subject 20A_Q3 Date //

Attendants

Restricted Boltzmann Machine

from p 3 f / p 66

Aizu Univ.

$$\frac{\partial L_D(\theta)}{\partial w_{ij}} = N \sum_{x,h} x_i x_j p(h|x,\theta) q_D(x)$$

$$- N \mathbb{E}_{p(V,H|\theta)}[X_i X_j]$$

$$p(h|x,\theta) = \frac{p(x,h|\theta)}{p(x|\theta)}$$

右辺 第一項は

$$N \sum_{x,h} x_i x_j p(h|x,\theta) q_D(x)$$

$$= N \cdot \frac{1}{N} \sum_{\mu=1}^{N} \sum_{h} f(x^{(\mu)}, h) p(h|x^{(\mu)}, \theta)$$

$$= \frac{1}{N} \sum_{\mu=1}^{N} \mathbb{E}_{p(h|x^{(\mu)},\theta)}[f(x^{(\mu)}, h)]$$

(再現するために computational explosion.

⇒ Gibbs-Sampling

Subject 20A_Q3 Date 12

Attendants

BM の学習 -> 最尤推定で求める

$$L_D(\theta) = \prod_{\mu=1}^{N} p(V = v^{(\mu)} | \theta)$$

$$\mathcal{L}_D(\theta) = \ln L_D(\theta) = \sum_{\mu=1}^{N} \ln p(v^{(\mu)} | \theta)$$

$$\frac{\partial \mathcal{L}_D(\theta)}{\partial w_{ij}} = \sum_{\mu=1}^{N} v_i^{(\mu)} v_j^{(\mu)} - N \mathbb{E}_{p(v|\theta)}[V_i V_j]$$

expectation data distribution

expectation of model distribution

where

$$\mathbb{E}_{p(v|\theta)}[f(v)] = \sum_{v} f(v) p(v|\theta)$$

(learning equation of Boltzmann machine)

$\frac{\partial \mathcal{L}_D(\theta)}{\partial w_{ij}} = 0$ とおく

moment matching

$$\frac{1}{N} \sum_{\mu=1}^{N} v_i^{(\mu)} v_j^{(\mu)} = \mathbb{E}_{p(v|\theta)}[V_i V_j]$$

Subject 20A-Q3 Date 13

Attendants

Gradient Ascent

↳ need to calc

$$\frac{\partial \ln p(\theta)}{\partial w_{ij}} = \sum_{u=1}^{N} u_i^{(u)} v_j^{(u)} - N \sum_{u} p(u|\theta)\,[v_i v_j]$$

↓

$$\sum_{v} u_i u_j\, p(u|\theta)$$

combination explosion.

ボーダーが 2^n 個の項の和で計算する

(d は v の次元数 $v_1 \cdots v_d$ の d.)

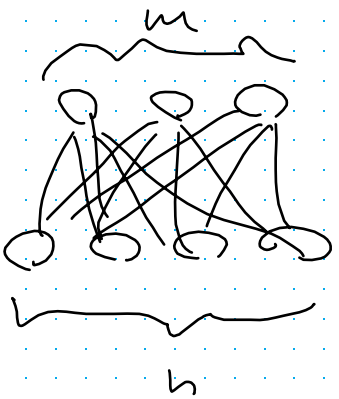
Subject 20A_Q3 Date 14

Attendants RBM : Restricted Boltzman Machine

↳ which is defined on complete bipartite graph

(完全二部グラフ)

as BM w/ hidden (latent variable



H hidden layer

V visible layer

as limitation,

there is no connection between same layer.

n個の要素を持つ観測

データ点の集合を用いて学習する

Subject 20A_03 Date 15

Attendants

$$p(h|v,\theta) = \frac{p(v,h|\theta)}{\sum_h p(v,h|\theta)} = \frac{p(v,h|\theta)}{p(v|\theta)}$$

同様に

$$p(v|h,\theta) = \prod_{j\in H} p_{sb}(h_j|\lambda_j)$$

$$= \prod_{j\in H} p_{sb}(h_j|\lambda_j)$$

$$\hookrightarrow p_{sb}(h_j|\lambda_j) = \frac{\exp(\lambda_j x_j)}{1+\exp(\lambda_j)}$$

sigmoid belief

Training of RBM

$$p(x|\theta) = p(v|\theta) = \sum_h p(h,v|\theta)$$

$$= \frac{1}{Z(\theta)} \exp\left(\sum_{i\in V} b_i v_i + \sum_{j\in H} h_j (1+\exp(\lambda_j)) \right)$$

↑

可視変数のみの 周辺分布

Subject JOA_Q3 Date 1.6

Attendants

RBMも同様に組み合わせ爆発する

⇔ contrastive divergence などの近似が必要

gibbs samplingはたくさんの反復回数Tが必要

⇓

小さいTでもうまくパラメータの更新量求めたい

step 1

v をランダム初期化ではなく訓練サンプルにする

⇓

hidden layerを計算

の probability

これと $U \sim U(0,1)$ を比べて

$U > prob \rightarrow 1$

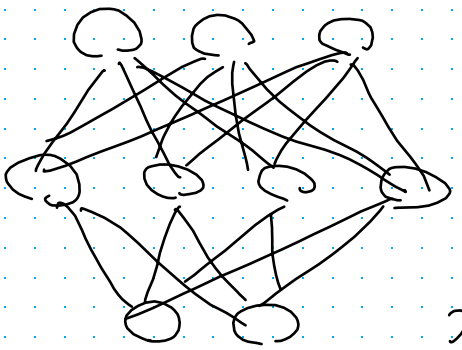
$U \leq prob \rightarrow 0$

Subject PreDoc2 2024 - Q3 Date 20

Attendants RBM April 15th

Q1. Describe Block-Gibbs Sampling Strategy for RBM

Graph G



} latent variables

At first, we assume that there are no connections between the same layers.

So, now, we can divide each node into two group as follows

$g_1 = \{x, z\}$

$g_2 = \{y\}$

$x = \{x_1, x_2 \dots\}$

$y = \{y_1, y_2 \dots\}$

$z = \{z_1, z_2 \dots\}$

by spliting G into two part as bipartite graph.

Subject 20A-Q3 Date 21 April (fri)

Attendants

Given g_1, it is easy to compute the conditional probability of g_2

$$p(g_2 \mid g_1, \theta) = \prod_{i=1} p(g_{2i} \mid g_1, \theta) = \frac{p(x, y, z)}{p(x, z)} \quad \cdots *1$$

and vice-versa

$$p(g_1 \mid g_2, \theta) = \prod_{j=1} p(g_{1j} \mid g_2, \theta) = \frac{p(x, y, z)}{p(y)} \quad \cdots *2$$

where

$$*1 = \frac{p(x, y, z)}{\sum_x \sum_z p(x, y, z)} = \frac{\exp(-\sum_{ijk} w x y z)}{\sum_x \sum_z \left(\exp(-\sum_{ijk} w x y z) \right)}$$

$O(2^{N+K})$

$$*2 = \frac{p(x, y, z)}{\sum_y p(x, y, z)} = \frac{\exp(-\sum_{ijk} w x y z)}{\sum_y \left(\exp(-\sum_{ijk} w x y z) \right)}$$

$O(2^N)$

Subject 20A_Q2 Date 22 April / 8th

Attendants

Block Gibbs Sampling Algorithm

step 1 initialize $x^{[0]}, y^{[0]}, z^{[0]}$ at random

$t = 1$

step t if $(1 < t < k)$

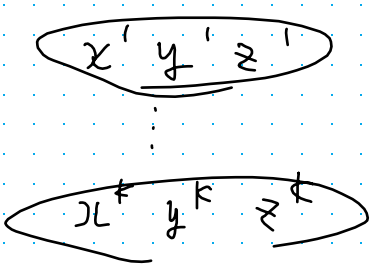
$x, z^{[t+1]} \sim p(x, z \mid y^{[t]}, \theta)$ by using #1

$y^{[t+1]} \sim p(y \mid x^{[t+1]}, \theta)$ by using #2

$t = t + 1$

Step K

return



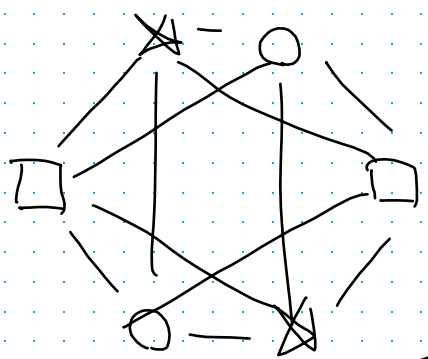
} k sample.

Subject 20A_Q3 Date 23 April 18th

Attendants

cost of block gibbs sampling

Multiparticle Graph (多粒子グラフ)



△ : x

○ : y

□ : z

$$P(x_1 \mid y, z) = P(x_1 \mid x_{-1}, y, z)$$

$$= \frac{P(x, y, z)}{P(x_{-1}, y, z)} \quad \text{of } p(x, y, z)$$

So $$P(x_1 = 0 \mid y, z) = \frac{\tilde{p}(x_1 = 0 \mid y, z)}{\tilde{p}(x_1 = 0 \mid y, z) + \tilde{p}(x_1 = 1 \mid y, z)}$$

$$\tilde{p}(x_1 = 1 \mid y, z) = \tilde{p}(X_1 = 1, X_{-1} = 0, y = y_t, z = z_t)$$

where $p = \frac{1}{Z}\tilde{p}$

Subject 20A_Q3 Date 24 April / 9th

Attendants

for calculation

$$\tilde{p}(X = x_1, x_{-1}, y = y_1, z = z_1)$$

but we have to wait mixing time

$$p(x,y,z) = \frac{1}{Z}\exp(-E(x,y,z))$$

$$\tilde{p}(x,y,z) = \exp(-E(x,y,z))$$

$$E(x,y,z) = \sum_{ijk} W_{ijk}\, x_i y_j z_k$$

$x \in \{0,1\}^N$

$y \in \{0,1\}^M$

$z \in \{0,1\}^K$

it costs $O(N \times M \times K)$

$$Z = \sum_x \sum_y \sum_z \exp(-E(x,y,z))$$

it costs $O(2^{M+N+K})$

we can save this cost by Block gibbs sampling