機械学習ゼミ
⑤共クラスタリング

2018年7月7日（土）

五百藏夏穂

概要

関係データ学習

* スペクトラルクラスタリング
* 確率的ブロックモデル/無限関係モデル
* 行列分解

本ゼミでの説明は，
“関係データ学習” 石黒勝彦，林浩平（著）
に準拠します．

関係データ学習

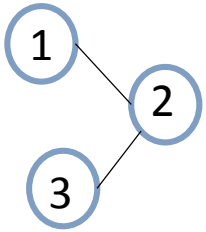
関係データとは

関係データ　複数のデータの間に観測，定義される関係に着目したデータのこと（SNSユーザー間，購買データ，etc）

有向関係
無向関係　起点によって関係が変わるか否か

単一ドメイン
複数ドメイン　オブジェクトが同じ種類か否か

ex（単一：SNSユーザー関係，複数：購買データ）



グラフ表現
$V = \{1,2,3\}$

0	1	0
1	0	1
0	1	0

行列表現
$X = (x_{i,j})\ i,j = 1,2,3$

対称関係データ：単一ドメインかつ無向関係
スペクトラルクラスタリング

非対称関係データ：それ以外
確率的ブロックモデル/無限関係モデル，行列分解

関係データ学習

タスク

＊予測　　：関係データ学習を通してみ観測のデータの値を推定
→アイテム推薦問題

＊知識抽出　：関係データの特性を解析，モデル化し，知見や知識の情報抽出
→コミュニティ抽出，クラスタリング

タスク

関係データの種類		知識抽出	予測
二項関係	対称	スペクトラルクラスタリング	行列分解
	非対称	確率的ブロックモデル 無限関係モデル	
多項関係			テンソル分解

スペクトラルクラスタリング

関係データ学習におけるクラスタリング

対象：オブジェクト（ノード）Vのクラスタリング

類似性：各オブジェクトi, j間の“関係”に基づく分類

※対象：エッジ（リンク），類似性：オブジェクト自身が持つ特徴量
に基づくクラスタリングはここでは本ゼミでは対象外とします．

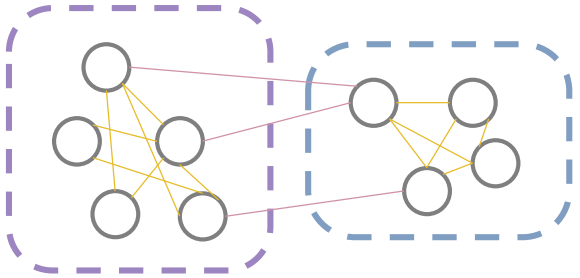
機械学習	教師あり	・分類	（決定木，NN, …etc）
		・回帰分析	(線形回帰, …etc)
	教師なし	・クラスタリング	(k-means, …etc)
		・共起分析	(協調フィルタリング, …etc)
		・次元圧縮	(主成分分析，　特異値分解, …etc)
	強化学習	・Q学習	
		・TD学習	
	etc		

コミュニティ検出と密結合グラフ

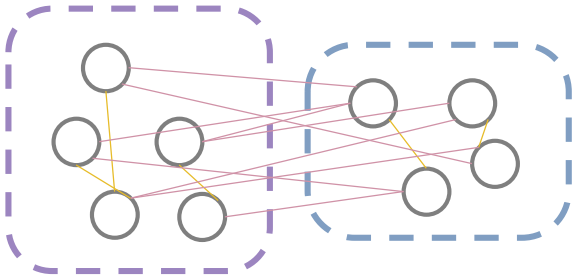
密結合クラスタ：クラスタ内部で密な結合，
クラスタ外部と疎な結合

疎結合クラスタ：クラスタ内部で疎な結合，
クラスタ外部と疎な結合

密結合クラスタグラフ



疎結合クラスタグラフ



➡ コミュニティ検出では前者
どのクラスタにどのオブジェクトを帰属させるかを計算

スペクトラルクラスタリング

グラフカット

疎・密結合の発見や判断に用いる特徴量 **グラフカット**

$$cut(P) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in P_k, j\in \bar{P}_k} x_{i,j}$$

P_k: k番目クラスタに所属する頂点集合

$\bar{P}_k$: k番目クラスタに所属しない頂点集合

$x_{i,j}$:異なるクラスタに所属する頂点i, j間の重み

頂点間のつながりの弱い部分で分割すると考えて

⇒ **カット最小化問題**として密結合クラスタを抽出

スペクトラルクラスタリング

グラフラプラシアンの固有値に着目し,
固有値の値によってグラフのノード間の連結性が推定できる

⇒ 固有値分解とk-meansの併用でグラフを構成する
頂点のクラスタリングを行う手法

スペクトラルクラスタリング

①非正規化グラフラプラシアンによるスペクトラルクラスタリング

グラフラプラシアン

観測データ行列：隣接行列

次数行列：

$$D = (d_{i,j}), \qquad i,j = 1,2,\cdots N$$

$$d_{i,j} = \begin{cases} \sum_{j=1}^{N} x_{i,j} & i = j \\ 0 & i \neq j \end{cases}$$

グラフラプラシアン：

$$L = X - D$$

観測データ行列X

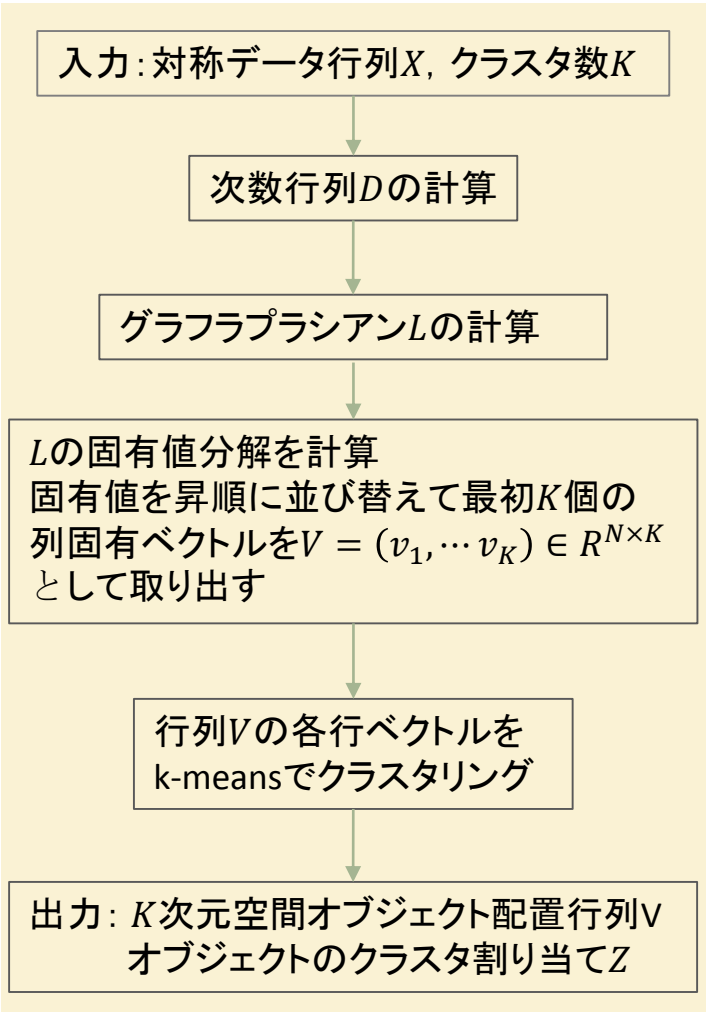
	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0	1	0	1	0
$i=2$	1	0	0	1	0
$i=3$	0	0	0	0	1
$i=4$	1	1	0	0	0
$i=5$	0	0	1	0	0

次数行列D

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	2	0	0	0	0
$i=2$	0	2	0	0	0
$i=3$	0	0	1	0	0
$i=4$	0	0	0	2	0
$i=5$	0	0	0	0	1

グラフラプラシアンL

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	2	-1	0	-1	0
$i=2$	-1	2	0	-1	0
$i=3$	0	0	1	0	-1
$i=4$	-1	-1	0	2	0
$i=5$	0	0	-1	0	1

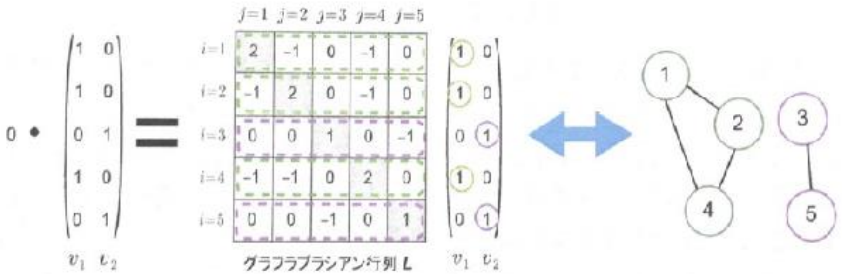


スペクトラルクラスタリング

①非正規化グラフラプラシアンによるスペクトラルクラスタリング

固有値分解によるクラスタ抽出

$$L = V\Lambda V^T \quad \text{または} \quad \lambda_k v_k = L v_k$$



Lは半正定値行列なので
$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$

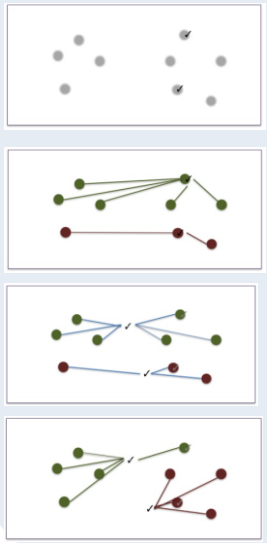
この時定数ベクトルの固有値は
0なので最小値λ_1=0を取る

定数ベクトルの要素を分割した
ベクトルを取ることで，K個のクラスタを
持つグラフでは最小K個の固有値0
（理想状態）

クラスタ割り当ての計算

列固有ベクトル$V = (v_1, \cdots v_K)$をN個の
K次元ベクトルととらえてk-meansを適用

クラスタ割り当て$Z = \{z_i = k\}$が求まる



※補足：k-means

K個の中心を決める
↓
最も近くの
中心クラスタに割り当て
↓
クラスタ内平均から
クラスタ中心を更新
↓
中心が変化しなくなったら
終了

スペクトラルクラスタリング

②正規化グラフラプラシアンによるスペクトラルクラスタリング

正規化カット

カット最小化の傾向として
大きなクラスタと，孤立し多数の単頂点クラスタ
ができやすい

⇒ 各クラスタの大きさに制約をかける
ように**正規化**を行ったカットを用いる

レシオカット

$$\mathrm{RatioCut} = \frac{1}{2}\sum_{k=1}^{K}\frac{\sum_{i\in P_k, j\in \bar{P}_k} x_{i,j}}{|P_k|} = \sum_{k=1}^{K}\frac{\mathrm{cut}(P_k, \bar{P}_k)}{|P_k|}$$

・各クラスタのサイズ（頂点数）で割り正規化
→クラスタが大きいとカットが小さくなる
・非正規化グラフラプラジアンのアルゴリズムで解ける

ノーマライズドカット

$$\mathrm{NCut} = \frac{1}{2}\sum_{k=1}^{K}\frac{\sum_{i\in P_k, j\in \bar{P}_k} x_{i,j}}{\mathrm{vol}(P_k)} = \sum_{k=1}^{K}\frac{\mathrm{cut}(P_k, \bar{P}_k)}{\mathrm{vol}(P_k)} \qquad \mathrm{vol}(P_k) = \sum_{i,j\in P} x_{i,j}$$

・クラスタ内のリンクの重みの総和で割る
→クラスタが大きくかつ相互に密に結合しているとカットが小さくなる
・正規化グラフラプラシアンに基づくスペクトラルクラスタリングに帰着

スペクトラルクラスタリング

②正規化グラフラプラシアンによるスペクトラルクラスタリング

対称正規化ラプラシアン

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

入力：対称データ行列X，クラスタ数K

次数行列Dの計算

グラフラプラシアンL_{sym}の計算

L_{sym}の固有値分解を計算
固有値を昇順に並び替えて最初K個の列固有ベクトルを$V = (v_1, \cdots v_K) \in R^{N \times K}$として取り出す

Vの各行ベクトルのノルムを1に正規化

行列Vの各行ベクトルを
k-meansでクラスタリング

出力：K次元空間オブジェクト配置行列V
オブジェクトのクラスタ割り当てZ

比較

＊コミュニティ（密結合）が対象であれば，密結合も考慮したノーマライズドカットの方が望ましい

＊L_{sym}はk-means前の正規化の分，計算コストが大きいため，L_{rw}の方が望ましい

酔歩正規化ラプラシアン

$$L_{rw} = D^{-1} L$$

入力：対称データ行列X，クラスタ数K

次数行列Dの計算

グラフラプラシアンL_{rw}の計算

L_{rw}の固有値分解を計算
固有値を昇順に並び替えて最初K個の列固有ベクトルを$V = (v_1, \cdots v_K) \in R^{N \times K}$として取り出す

行列Vの各行ベクトルを
k-meansでクラスタリング

出力：K次元空間オブジェクト配置行列V
オブジェクトのクラスタ割り当てZ

ノーマライズドカットをグラフラプラシアンを用いた二次形式に書き換え，それを最小化する解を制約付きで解くと導出できる

関係データXの重みに従って酔歩すると想定すると
遷移確率行列は$\mathrm{T} = D^{-1}X$
ノーマライズドカット最小化＝クラスタ間遷移確率最小化

スペクトラルクラスタリング

②実データ適用例

Zachary's Karate Club network data
空手クラブのメンバーの交友関係は，レッスン料の違いから2つの集団に分裂

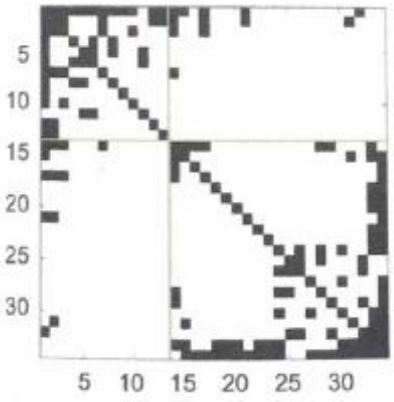
N=34, K=2として各手法を適用

実際の結果と一致しなかったもの

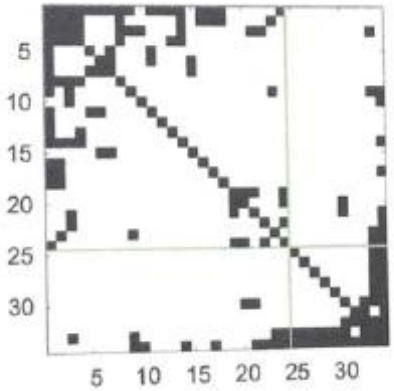
非正規化/酔歩正規化：4/34

対称正規化：7/34

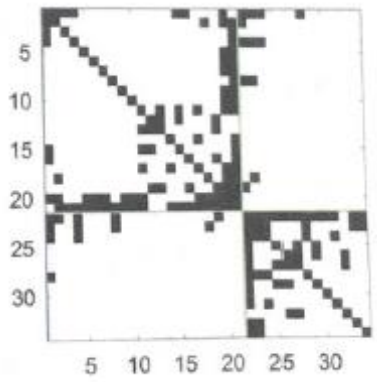
非正規化



対称正規化



酔歩正規化



確率的ブロックモデル

確率的ブロックモデルと無限関係モデル

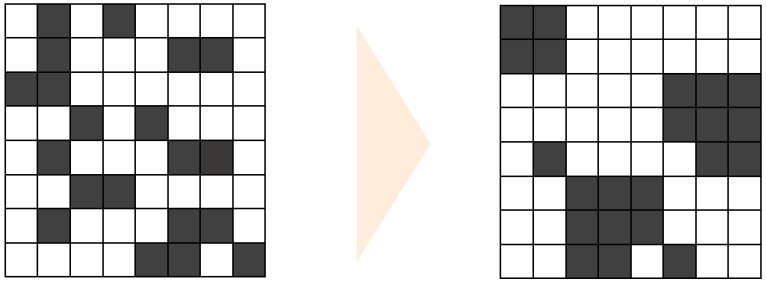
スペクトラルクラスタリングでは，対象関係データの密結合クラスタに対象が限られていた

➡ 非対称関係データにも適応可能で
密結合以外も抽出可能なデータクラスタリング手法

＊潜在的なブロック構造を仮定

＊確率的アプローチ

関係データに潜むブロック構造を推定し，この時，
この構造によりどのような観測データが確率的に得られるのか？



確率的ブロックモデル
Stochastic Block Model（SBM）

→ 無限関係モデル
Infinite Relational Model（IRM）

クラスタ数Kが
自動的に求まる

確率的ブロックモデル

確率的生成モデル

SBMもIRMもこの一種である

系を構成する各変数の値が依存関係に従って順番に生成される過程を
確率で表現するモデル

ex)サイコロの目(100回振って出た目を記録)

100回サイコロを振る　　:変数をサンプリング
1回ごとにサイコロの形状に従って　:パラメータで規定される確率分布
ランダムに目の値が決まる　　:観測値が生成される

$$x_i|\pi \sim P(\pi), i = 1,2, \dots 100$$

$X = (x_1, \cdots x_{100})$:観測値
$\pi = (\pi_1, \cdots \pi_d \pi_D)$: 面の出やすさパラメータ
D: 面数　$\sum_d \pi_d = 1$

メリット
＊確率分布の形状やパラメータが全て求まれば，実際に値を生成できる
＊事前知識を確率分布の形状や種類として反映できる

確率モデルのベイズ推定

観測データから確率分布の形状やパラメータ値を推定
本書で説明されるベイズ推定では，ベイズ事後分布を求める

$$p(Z, \{\theta\}|X) \propto p(X|Z, \{\theta\}|)p(Z, \{\theta\})$$

周辺化

$$p(Z|X) \propto \int p(X|Z, \{\theta\}|)p(Z, \{\theta\})\, d\{\theta\}$$

確率的ブロックモデル

確率的ブロックモデルの定式化

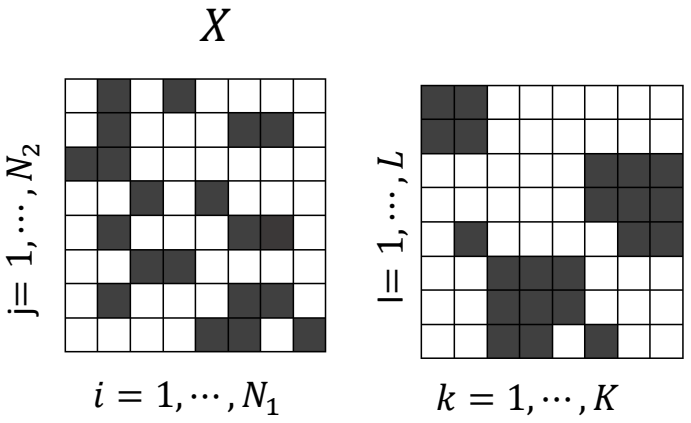
観測データ: $X = \{x_{i,j}\}\ \ x_{i,j} \in \{1,0\}$

行(第一ドメイン)インデックス: $i = 1, \cdots, N_1$

列(第二ドメイン)インデックス: $j = 1, \cdots, N_2$

行クラスタインデックス: $k = 1, \cdots, K$

列クラスタインデックス: $l = 1, \cdots, L$



$\pi_1 | \alpha_1 \sim \mathrm{Dirichlet}(\alpha_1)$

$\pi_2 | \alpha_2 \sim \mathrm{Dirichlet}(\alpha_2)$

π:オブジェクトのクラスタへの混合割合パラメータ

α:ハイパラメータ

ex)K面のサイコロをランダムに生成する分布

$z_{1,i}=k | \pi_1 \sim \mathrm{Discrete}(\pi_1)$

$z_{2,j}=l | \pi_2 \sim \mathrm{Discrete}(\pi_2)$

各ドメインのクラスタの混合割合に従ってオブジェクトのクラスタ割り当てをサンプリング

ex)K面のサイコロに従って目を出す

$\theta_{k,l} | a_0, b_0 \sim \mathrm{Beta}(a_0, b_0)$

θ:クラスタ間関係強さ(2値関係なのでベータ分布)

$x_{i,j} | \{\theta_{k,l}\}, z_{1,i}, z_{2,j} \sim \mathrm{Bernoulli}(\theta_{z_{1,i}, z_{2,j}})$

z:割り当てを表す隠れ変数

観測行列の生成

確率的ブロックモデル

周辺化ギブスサンプラー（CGS）による推論

$$z_i^{(s)} \sim p(z_i|X, Z^{/(i)}) \propto \int p\left(X^{(i)}\middle|X^{/(i)}, z_i, Z^{/(i)}\ \{\theta\}\right) p\left(z_i, \{\theta\}|X^{/(i)}, Z^{/(i)}\right) d\{\theta\}$$

※確率モデルは隠れ変数z，パラメータ集合$\{\theta\}$，観測データXからなるが，
結果的に必要なのは隠れ変数の事後分布だけなのでパラメータは周辺化して推論

s：サンプリング回数
$X^{(i)}$：z_iに依存する観測量
$X^{/(i)}$：$X - X^{(i)}$

事後分布から生成されたとみなされる確率変数のサンプリングを多数行い，
サンプルの経験分布をもって事後分布の推定値を計算する
CGSでは，1回のサンプリングである一つの変数だけのサンプリングを順番に行う．

$z_i^{(s)}$はs回目の事後分布サンプリング結果

↓ $z_i \leftarrow z_i^{(s)}$：既知の値として保存

$z_j\ (i \neq j)$のサンプリングに映る

確率的ブロックモデル

CGSによるSBMの推論式

第一・第二ドメインで対象なので

$$p(z_{1,i}=k|X,Z_1^{/(i)},Z_2)\propto$$

$$\int p\left(X^{(i)}|X^{/(i)},z_{1i=k},Z_1^{/(i)},Z_2,\pi_1,\pi_2,\{\theta_{k,l}\}\right)\cdot p\left(z_{1i=k},\pi_1,\pi_2,\{\theta_{k,l}\}|X^{/(i)},Z_1^{/(i)},Z_2\right)d\pi_1 d\pi_2 d\{\theta_{k,l}\}$$

z_{1i}に依存する関係データ行列の要素に対する尤度

z_{1i}及びパラメータのz_{1i}以外の情報が所与の時の事後分布

$$p(z_{1,i}=k|X,Z_1^{/(i)},Z_2)\propto$$

$$\int p(z_{1,i}=k|\pi_1)p(\pi_1|Z_1^{\overline{i}})d\pi_1\cdot\int\prod_{l=1}^{L}\prod_{j=1}^{N_2}[p(x_{i,j}|\theta_{k,l})]^{\mathrm{II}(z_{2,j}=l)}\ p(\theta_{k,l}|X^{/(i)},Z_1^{/(i)},Z_2)\ d\{\theta_{k,l}\}$$

π_1を周辺化した時のクラスタ割り当て$z_{1,i}$の確率

第二ドメインのクラスタlごとにクラスタ間関係強さ$\theta_{k,l}$を周辺化した時の観測データの尤度の影響

生成モデルに従って計算

$$p(z_{1,i}=k|X,Z_1^{/(i)},Z_2)\propto\widehat{\alpha_{1,k}}\frac{\Gamma(\hat{a}_{k,l}+\hat{b}_{k,l})}{\Gamma(\hat{a}_{k,l})\Gamma(b_{k,l})}\times\frac{\Gamma(\hat{a}_{k,l}+\sum_{j=1}^{N_2}x_{i,j}\,\mathrm{II}(z_{2,j}=l))\,\Gamma(\hat{b}_{k,l}+\sum_{j=1}^{N_2}(1-x_{i,j})\,\mathrm{II}(z_{2,j}=l))}{\Gamma(\hat{a}_{k,l}+\hat{b}_{k,l}+\sum_{j=1}^{N_2}\mathrm{II}(z_{2,j}=l))}$$

$$\hat{a}_{k,l}=a_0+\hat{n}_{k,l}^{(+)}\quad\hat{b}_{k,l}=b_0+\hat{n}_{k,l}^{(-)}$$

確率的ブロックモデル

CGSによるSBMの推論式(十分統計量)

定義

$$m_{1,k} = \sum_{i=1}^{N_1} \mathrm{II}\left(z_{1,i} = k\right)$$

:クラスタkに所属するオブジェクト数

$$n_{k,l}^{(+)} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{i,j}\,\mathrm{II}\left(z_{1,i} = k\right)\mathrm{II}\left(z_{2,j} = l\right)$$

$$n_{k,l}^{(-)} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} (1 - x_{i,j})\,\mathrm{II}\left(z_{1,i} = k\right)\mathrm{II}\left(z_{2,j} = l\right)$$

:ブロック(k,l)において$x = 1(0)$となる要素数

$z_{1,i}$のサンプリング時に

$z_{1,i}$のクラスタ割り当てを解除してた時の一つ抜き十分統計量

$$\hat{m}_{1,k} = \sum_{i' \neq i, i'=1}^{N_1} \mathrm{II}\left(z_{1,i'} = k\right) = m_{1,k} - \mathrm{II}\left(z_{1,i} = k\right)$$

$$\hat{n}_{k,l}^{(+)} = \sum_{i' \neq i, i'=1}^{N_1}\sum_{j=1}^{N_2} x_{i',j}\,\mathrm{II}\left(z_{1,i'} = k\right)\mathrm{II}\left(z_{2,j} = l\right) = n_{k,l}^{(+)} - \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} x_{i,j}\,\mathrm{II}\left(z_{2,j} = l\right)$$

$$\hat{n}_{k,l}^{(-)} = \sum_{i' \neq i, i'=1}^{N_1}\sum_{j=1}^{N_2} (1 - x_{i',j})\,\mathrm{II}\left(z_{1,i'} = k\right)\mathrm{II}\left(z_{2,j} = l\right) = n_{k,l}^{(-)} - \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} (1 - x_{i,j})\,\mathrm{II}\left(z_{2,j} = l\right)$$

更新

$$m_{1,k} = \hat{m}_{1,k} + \mathrm{II}\left(z_{1,i} = k\right)$$

$$n_{k,l}^{(+)} = \hat{n}_{k,l}^{(+)} + \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} x_{i,j}\,\mathrm{II}\left(z_{2,j} = l\right)$$

$$n_{k,l}^{(-)} = \hat{n}_{k,l}^{(-)} + \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} (1 - x_{i,j})\,\mathrm{II}\left(z_{2,j} = l\right)$$

対象となる変数の差し引きだけで，1回のサンプリングに対応する更新式計算ができる！

無限関係モデル

無限関係モデルの定式化

無限個のクラスタの存在を仮定した確率的生成モデル
潜在クラスタ数K, Lを自動的に決定できる

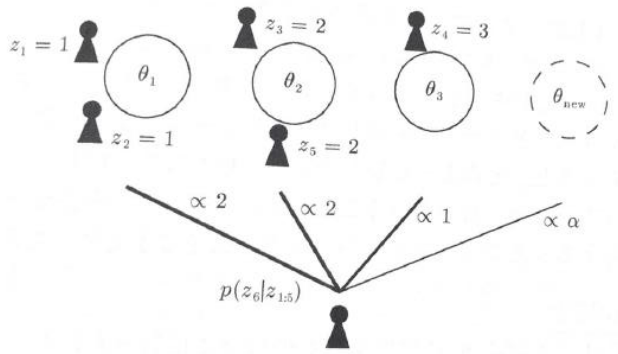
$$Z_1 | \alpha_1 \sim \mathrm{CRP}(\alpha_1)$$
$$Z_2 | \alpha_2 \sim \mathrm{CRP}(\alpha_2)$$

CRPは無次元ディリクレ分布と無次元離散分布を組み合わせたもの

$$\theta_{k,l} | a_0, b_0 \sim \mathrm{Beta}(a_0, b_0)$$

$$x_{i,j} | \{\theta_{k,l}\}, z_{1,i}, z_{2,j} \sim \mathrm{Bernoulli}(\theta_{z_{1,i}, z_{2,j}})$$

Chinese Restaurant Process(CRP)
ノンパラメトリックベイズで用いられる確率過程



＊各テーブルに着席できる客数に上限はない
＊中華料理店に配置できるテーブル数に上限はない

⇨ 入店した客をどのテーブルに配置するか決める

$n+1$人目の客のテーブル選択確率

$$P(z_{n+1} = k | z_{1:n}, \alpha) \propto \begin{cases} \sum_{i=1}^{n} \mathrm{II}(z_i = k) & k \in \{1, 2, \cdots, K\} \\ \alpha & k = K + 1 \end{cases}$$

無限個のコンポーネントを持つ混合分布を正しく表現
実際のサンプリングではN+1個のテーブルと，有限になるので実装可能

無限関係モデル

CGSによるIRMの推論式

$$p(z_{1,i}=k|X,Z_1^{/(i)},Z_2)\propto$$
$$\int p\left(X^{(i)}|X^{/(i)},z_{1i=k},Z_1^{/(i)},Z_2,\{\theta_{k,l}\}\right)\cdot p\left(z_{1i=k},\{\theta_{k,l}\}|X^{/(i)},Z_1^{/(i)},Z_2\right)d\{\theta_{k,l}\}$$

$$p(z_{1,i}=k|X,Z_1^{/(i)},Z_2)\propto$$
$$p\left(z_{1,i}=k\middle|Z_1^{/(i)}\right)\int\prod_{l=1}^{L}\prod_{j=1}^{N_2}[p(x_{i,j}|\theta_{k,l})]^{\mathbb{I}\left(z_{2,j}=l\right)}\ p(\theta_{k,l}|X^{/(i)},Z_1^{/(i)},Z_2)\ d\{\theta_{k,l}\}$$

場合分けして計算

$$p(z_{1,i}=k\in\{1,2,\cdots,K\}(\text{既存クラスタ})|X,Z_1^{/(i)},Z_2)\propto$$
$$\hat{m}_{1,k}\prod_{l=1}^{L}[\frac{\Gamma(\hat{a}_{k,l}+\hat{b}_{k,l})}{\Gamma(\hat{a}_{k,l})\Gamma(b_{k,l})}\times\frac{\Gamma(\hat{a}_{k,l}+\sum_{j=1}^{N_2}x_{i,j}\,\mathbb{I}\left(z_{2,j}=l\right))\,\Gamma(\hat{b}_{k,l}+\sum_{j=1}^{N_2}(1-x_{i,j})\,\mathbb{I}\left(z_{2,j}=l\right))}{\Gamma(\hat{a}_{k,l}+\hat{b}_{k,l}+\sum_{j=1}^{N_2}\mathbb{I}\left(z_{2,j}=l\right))}]$$

$$p(z_{1,i}=k+1\ (\text{新規クラスタ})|X,Z_1^{/(i)},Z_2)\propto$$
$$\alpha_1\prod_{l=1}^{L}[\frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)}\times\frac{\Gamma(a_0+\sum_{j=1}^{N_2}x_{i,j}\,\mathbb{I}\left(z_{2,j}=l\right))\,\Gamma(b_0+\sum_{j=1}^{N_2}(1-x_{i,j})\,\mathbb{I}\left(z_{2,j}=l\right))}{\Gamma(a_0+b_0+\sum_{j=1}^{N_2}\mathbb{I}\left(z_{2,j}=l\right))}]$$

無限関係モデル

CGSによるIRMの推論式(十分統計量)

更新

$$m_{1,k} = \hat{m}_{1,k} + \mathrm{II}\left(z_{1,i} = k\right)$$

$$n_{k,l}^{(+)} = \hat{n}_{k,l}^{(+)} + \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} x_{i,j}\,\mathrm{II}\left(z_{2,j} = l\right)$$

$$n_{k,l}^{(-)} = \hat{n}_{k,l}^{(-)} + \mathrm{II}\left(z_{1,i} = k\right)\sum_{j=1}^{N_2} (1 - x_{i,j})\,\mathrm{II}\left(z_{2,j} = l\right)$$

もし$z_{1,i} = k + 1$ならば
ハットつきの値を全て0にする

出力方法

サンプリング実現値を用いて事後分布を計算

最初の適当なB期間のサンプルを棄却
その後M個置きのサンプル実現値だけを抽出(T=$\frac{S-B}{M}$個)

$$p\left(z_{1,i} = k \middle| X\right) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{II}\left(z_{1,i}^{(t)} = k\right) \qquad p\left(z_{2,j} = l \middle| X\right) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{II}\left(z_{2,j}^{(t)} = l\right)$$

具体的なクラス割り当て

事後分布最大化(MAP)に基づく

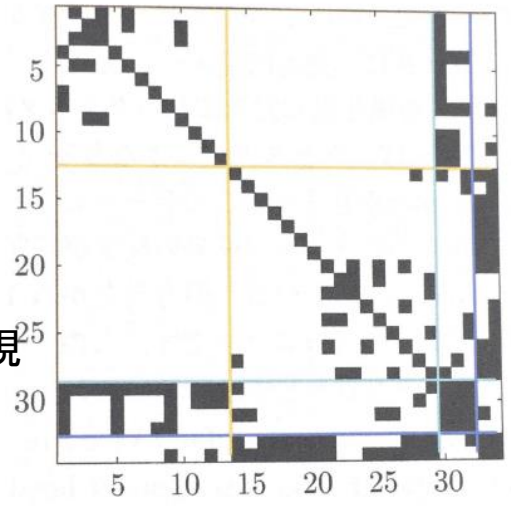
$$c_{1,i} = \mathrm{argmax}\, p\left(z_{1,i} = k \middle| X\right)$$

$$c_{2,j} = \mathrm{argmax}\, p\left(z_{2,j} = l \middle| X\right)$$

無限関係モデル

②実データ適用例

＊Zachary's Karate Club network data



4つに分裂
現実と分裂結果の一致度は
スペクトラルクラスタリングより良い値に.

➡ 人間の思い込みに反するがよりよく関係データを表現

＊Enron社内のEメールデータ

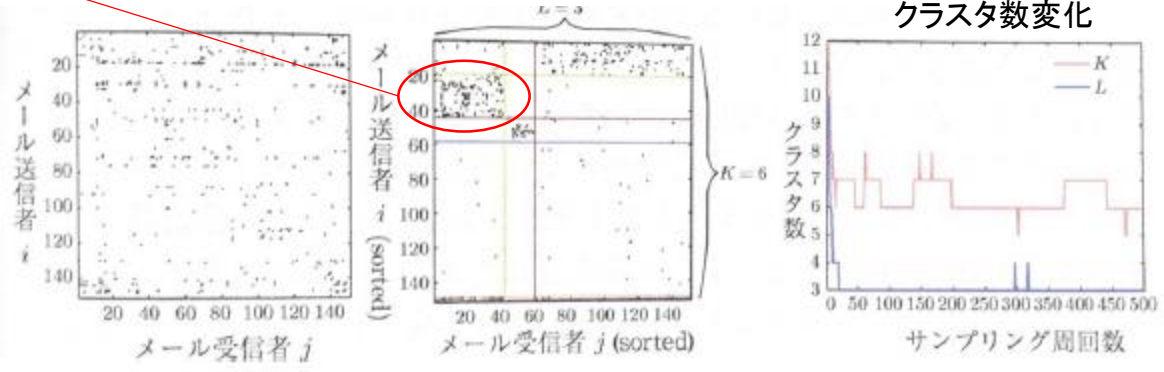
メール送信者(K), 受信者(L)の二値データ

K=6, L=3クラスタに分割

経営層コミュニティ

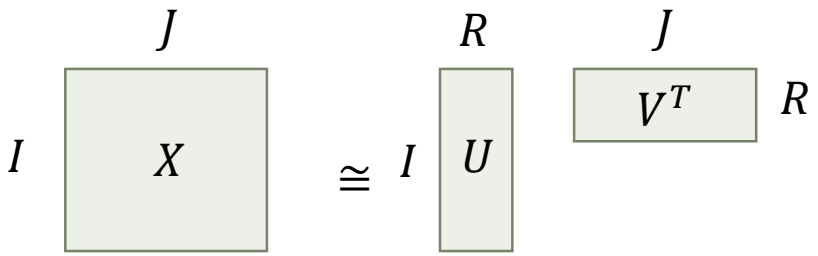
➡ 真のクラスタ構造がわからないデータでもクラスタリングを実行可能

関係データ行列　IRM適用結果　繰り返し回数ごとのクラスタ数変化



行列分解

単純行列分解



例として映画の推薦を考えます

I: 顧客 J:映画

要素の値が評価値となる行列

$$x_{ij} \cong u_{i,action} v_{j,action} + u_{i,horror} v_{j,horror}$$

より一般に$R \in N$個の評価軸があり，すべての顧客，映画について評価値が得られているとする

$u_{i,:} \in \mathrm{R}$:顧客iの嗜好をまとめたもの

$v_{j,:} \in \mathrm{R}$:映画jの成分をまとめたもの

XのランクR行列分解

$$X \cong U V^T$$

因子行列

$U \in \mathrm{R}^{I \times R}$:行方向にまとめた

$V \in \mathrm{R}^{J \times R}$:列方向にまとめた

目的関数

$$E = X - U V^T$$

Xと$U V^T$の近似誤差を最小化する問題ととらえる

2乗誤差最小化問題

$$\min_{U,V} \frac{1}{2} \|E\|_{\mathrm{Fro}}^2 = \min_{U,V} \frac{1}{2} \|X - U V^T\|_{\mathrm{Fro}}^2$$

行列分解

l^2正則行列分解

Xに大きいノイズが加わっていた時などは過学習の危険がある
→正則化（制約を付加し，解の範囲を狭める）

$$\min_{U,V} \frac{1}{2}\|X - U\,V^T\|_{\mathrm{Fro}}^2 + \frac{\lambda}{2}\left(\|U\|_{\mathrm{Fro}}^2 + \|V\|_{\mathrm{Fro}}^2\right)$$

$\lambda \geq 0$は正則化強さを決定する係数

非負行列分解

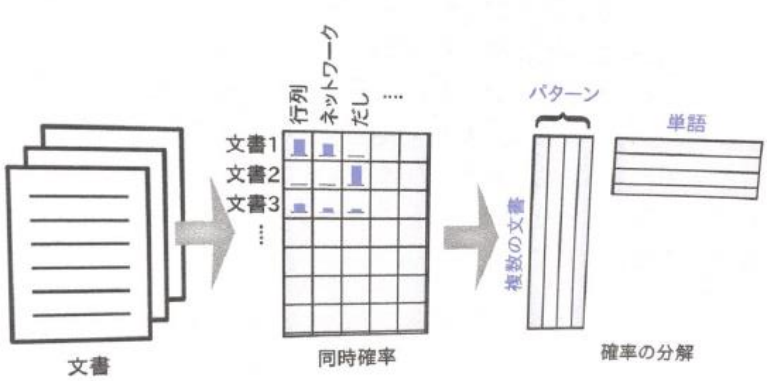
テキストの単語数データなど，非負値しかとらないもの

$$\min_{U \in R_+^{I\times R}, V \in R_+^{J\times R}} \frac{1}{2}\|X - U\,V^T\|_{\mathrm{Fro}}^2$$

Ex)　同時確率

$x_{i,j} = p(\text{文書}i, \text{単語}j)$

＊単語の確率分布は典型的なパターンを持つ
＊文書における単語の確率分布はパターンの組み合わせ



行列分解

アルゴリズム

目的関数の導関数を0としても，連立方程式を代数的に解くことができない

➡ 勾配法による最適化

一次交互勾配法

$K \in N$個の変数$\theta_1, \cdots \theta_K$を引数にとる関数$h(\theta_1, \cdots \theta_K) \in$Rでは

$$\theta_k^{new} = \theta_k - \eta \nabla_{\theta_k} h(\theta_1^{new}, \cdots \theta_{k-1}^{new}, \theta_k, \theta_{k+1}, \cdots, \theta_K)$$

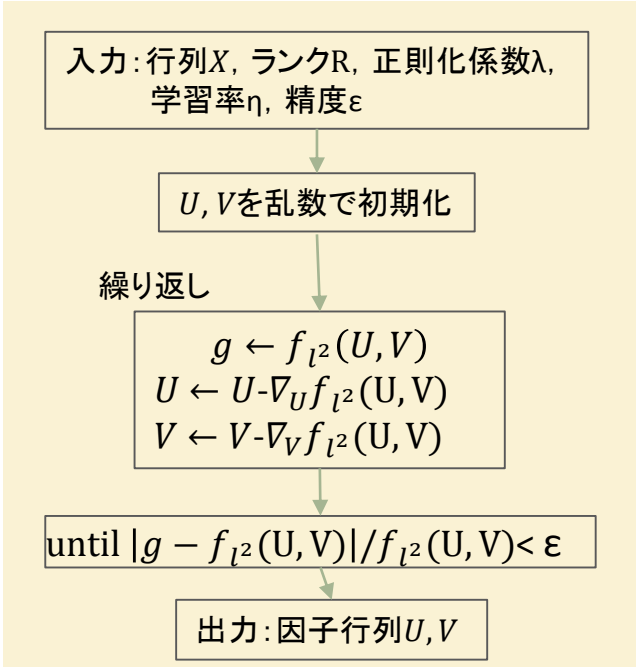
η:学習率

l^2正則行列分解の交互最適化

目的関数と連鎖率より

$$\nabla_U f_{l^2} = -XV + U(V^T V + \lambda I)$$
$$\nabla_V f_{l^2} = -X^T U + V(U^T U + \lambda I)$$

Xが密でデータが大きい場合計算量の負荷が大きい



行列分解

疑似二次交互勾配法

一次交互勾配法は実装は比較的容易であるが，収束が遅いことも．
二階微分の情報を用いて収束を早くする

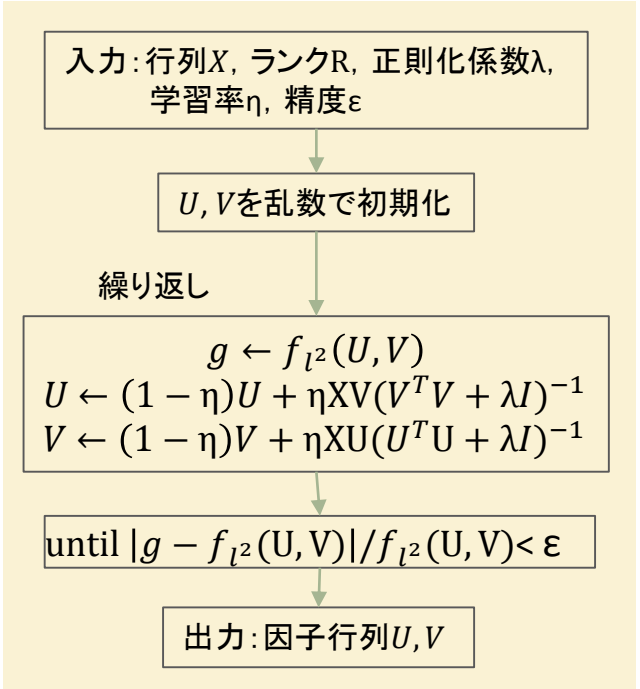
$$\theta_k^{new} = \theta - \eta\,(\nabla\nabla^{\mathrm{T}} g(\theta))^{-1}\nabla g(\theta)$$

ヘッセ行列

一次勾配法に比べて収束速度は速いが，
ヘッセ行列の逆行列を計算しなければならず，
計算量は高次元

l^2正則行列分解の交互最適化

$$U^{new} = (1-\eta)U + \eta \mathrm{XV}(V^T V + \lambda I)^{-1}$$
$$V^{new} = (1-\eta)V + \eta \mathrm{XU}(U^T \mathrm{U} + \lambda I)^{-1}$$



行列分解

交互勾配法＋射影勾配降下法（制約条件）

射影勾配降下法　非負制約などの制約を取り入れたい場合

u: 制約を満たす集合

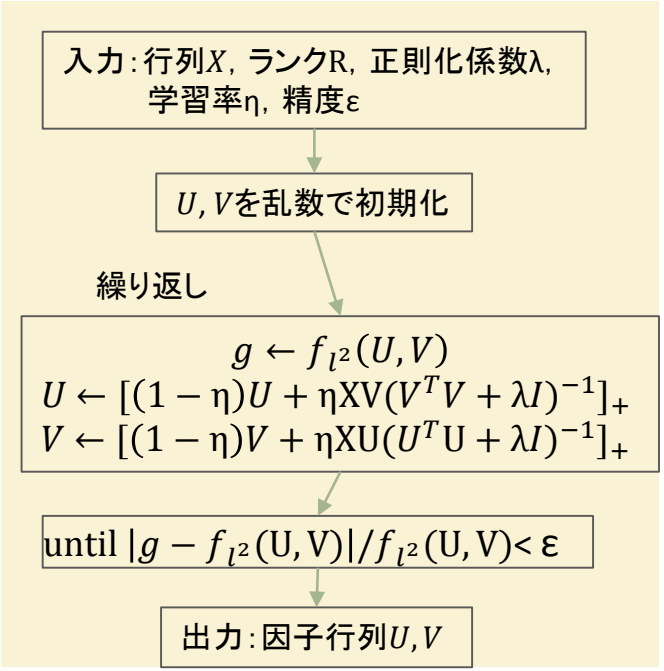
$$U^{new}=\mathrm{Proj}_u[U-\eta\nabla_U f_{l^2}(U,V)]$$
$$\mathrm{Proj}_u[U]=argmin_{Z\in u}\|U-Z\|^2_{\mathrm{Fro}}$$

通常の勾配法の後に，空間uからはみ出た部分を戻す

非負条件では，

$$U^{new}=[U-\eta\nabla_U f_{l^2}(U,V)]_+$$

$[U]_+ : u_{i,j} > o$の時$u_{i,j}$，$u_{i,j} \leqq o$の時0

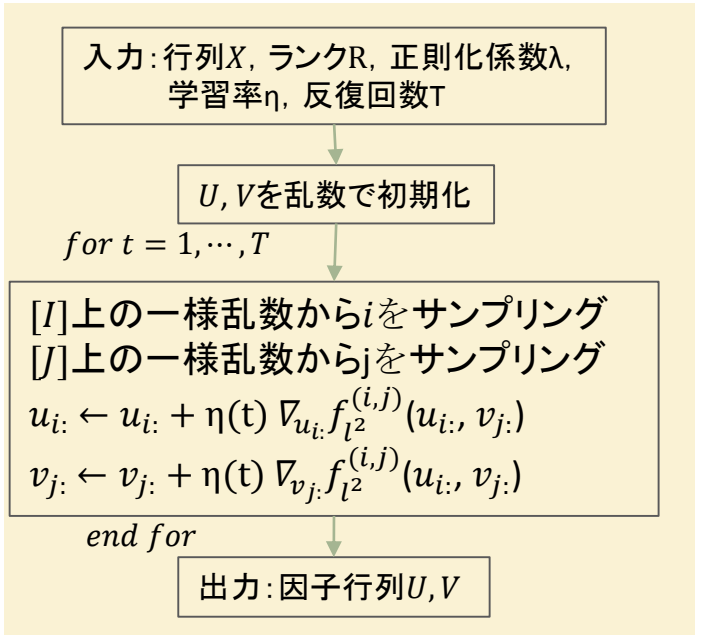


行列分解

確率勾配降下法

Xが密でI, Jがお真理に大きいとメモリに乗りきらないことがある
少数のサンプルで勾配を近似的に計算し最適化を行う方法.

⇒ 目的関数fがサンプルnのみに依存する関数f_n
の和で書けるとき，代わりにf_nの勾配を使う



要素ごと

$$\nabla_{u_{i:}} f_{l^2}^{(i,j)} = -x_{i,j} v_{j:} + u_{i:}\left(v_{j:}^T v_{j:} + \frac{\lambda}{J}\right)$$

$$\nabla_{v_{j:}} f_{l^2}^{(i,j)} = -x_{i,j} u_{i:} + v_{j:}\left(u_{i:}^T u_{i:} + \frac{\lambda}{I}\right)$$

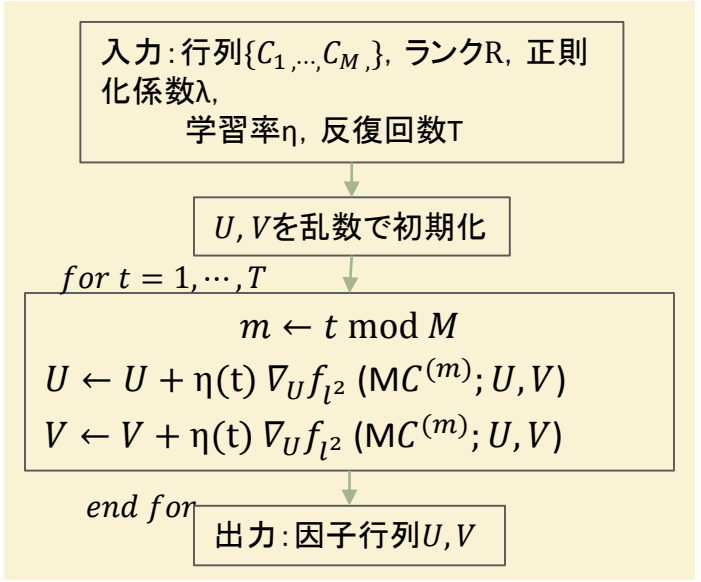
行列ごと

共起行列のようなXが複数の行列の和で与えられており
要素へのアクセスにもコストがかかる場合（ex. 単語の共起行列）

単純行列分解での目的関数を
$f_{std}(X; U, V) = \frac{1}{2}\|X - U V^T\|_{\mathrm{Fro}}^2$と表記すると

$$f_{std}(X; U, V) = P_m\left[f_{std}\left(\mathrm{MC}^{(m)}; U, V\right)\right] + M^2\sigma^2$$

$\sigma^2 = P_m\left\|C^{(m)} - P_m[C^{(m)}]\right\|_{\mathrm{Fro}}^2$, $P_m[\cdot] = \frac{1}{M}\sum_{m=1}^{M}(\cdot)$



行列分解

欠損値がある場合の行列分解

欠損値を除外

欠損部分を完全に未知だとし，学習の際に一切使わない方法

Xの観測部分のインデックス集合Ω

目的関数：$\sum_{(i,j)\in\Omega}(x_{i,j}-u_{i:}^{T}v_{j:})^{2}$

勾配：$\nabla_{U}f_{l^2}=-\left((X-UV^{T})*M\right)V+\lambda U$

マスク $m_{i,j}=\begin{cases}1 & (i,j)\in\Omega\\ 0 & \text{それ以外}\end{cases}$

欠損値を補完

Xの欠損値を何らかの形で補完した行列を$\hat{\mathrm{X}}$とする

目的関数：$\left\|\hat{\mathrm{X}}-U\,V^{T}\right\|_{\mathrm{Fro}}^{2}$

補完の仕方はいくつかあるが，交互最適化と組み合わせる場合，

U,Vをランダムに初期化した後t+1回目の反復においてt回目の予測値で補完

U^{t+1}の更新：i, j ∉ Ωにおいて$\hat{x}_{i,j}=(u_i^{(t)})^{T}v_j^{(t)}$

V^{t+1}の更新：i, j ∉ Ωにおいて$\hat{x}_{i,j}=(u_i^{(t+1)})^{T}v_j^{(t)}$

参考文献

＊「関係データ学習」 石黒勝彦・林浩平　講談社

＊「続・わかりやすい　パターン認識」 石井健一郎・上田修功　オーム社