クレジット：

Mathematics and Informatics Center 文科系のための線形代数・解析II
2020 藤堂 眞治・松尾 泰・藤原 毅夫

§ 主成分分析

多次元 データ　$\vec{x}^{(1)} = \begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_m^{(1)} \end{bmatrix} \in \mathbb{R}^n$ …　$\vec{x}^{(p)} = \begin{bmatrix} x_1^{(p)} \\ \vdots \\ x_n^{(p)} \end{bmatrix} \in \mathbb{R}^n$

(例)　入学試験

$x_i^{(1)}$：国語の点数．$x_i^{(2)}$：数学の点数，…　　$i = 1 \sim n$　受験生 ID

主成分分析　：　多次元データから情報をできるだけ失なわないように 情報の収約 を行う

例) 入学試験；合計点（重みつき）　$y_i = \sum_{a=1}^{p} w^{(a)} x_i^{(a)}$　一つの指標で優劣をつける

重み　：　文系：国語．　理系：数学

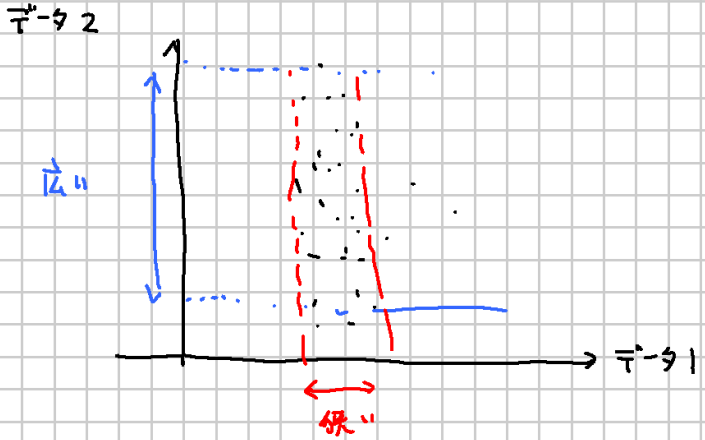
- 肥満度　：　BMI $=$ (体重)/(身長)^2　：　一つのデータで判定

$$\log(\mathrm{BMI}) = 1 \times \log(\text{体重}) - 2 \log(\text{身長})$$

（1, −2 が重み）

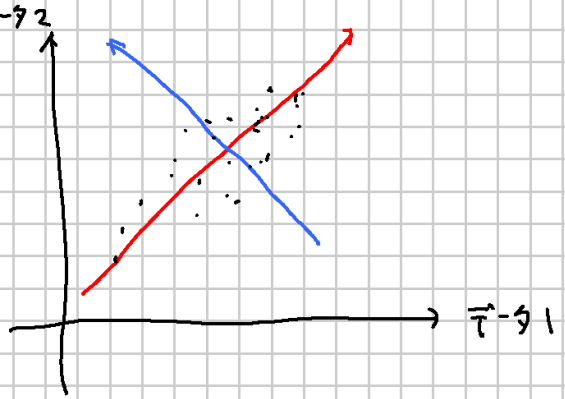
"重み" をどのようにつけるのか？

▷ 情報が多いデータを選ぶためには？　“情報が多い”とは



一般的に分散（データの広がり）が大きいデータの方が情報が多いと考えられる

各データの線形結合をとることにより　より分散が大きなデータを見い出すことができる



第一成分：射影した分散が最も大きくなる方向

第二成分：〃　2番目に〃

統計学より

データが1つの場合

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

偏差 $x_i - \bar{x}$

分散 $\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $\quad = \mathrm{Var}(x)$

$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$

複数のデータがある場合 $x_i^{(a)}$ $a = 1, \cdots, p$, $i = 1 \cdots n$

各データの平均 $\bar{x}^{(a)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(a)}$

分散 $\sigma_{aa} = \frac{1}{n}\sum_{i=1}^{n}(x_i^{(a)} - \bar{x}^{(a)})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^{(a)})^2 - (\bar{x}^{(a)})^2$ $\quad = \mathrm{Var}(x^{(a)})$

共分散 $\sigma_{ab} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(a)} x_i^{(b)} = \mathrm{Cov}(x^{(a)}, x^{(b)})$

相関係数 $\rho_{ab} = \sigma_{ab} / \sqrt{\sigma_{aa}\,\sigma_{bb}}$

$-1 \leq \rho_{ab} \leq 1$ $\quad \rho_{ab} > 0$ 正の相関 $\quad \rho_{ab} < 0$ 負の相関

$X = \left[x_i^{(a)} - \bar{x}^{(a)} \right]_{1 \le i \le n,\ 1 \le a \le p}$ $n \times p$ 各データの偏差を行列の形にまとめたもの

${}^tX \cdot X$: グラム行列

$$({}^tX \cdot X)_{ab} = \sum_{i=1}^{n} \left(x_i^{(a)} - \bar{x}^{(a)}\right)\left(x_i^{(b)} - \bar{x}^{(b)}\right) = \begin{cases} n\,\mathrm{Var}(x^{(a)}) & : a = b \\ n\,\mathrm{Cov}(x^{(a)}, x^{(b)}) & : a \neq b \end{cases}$$

グラム行列の固有値方程式　$\vec{w}$: $p \times 1$ ベクトル

$${}^tXX \cdot \vec{w}^{(a)} = \lambda^{(a)} \vec{w}^{(a)}$$

$w^{(a)}$: 固有ベクトル　$\lambda^{(a)}$: 固有値

$$\vec{w}^{(a)} = \begin{pmatrix} w_1^{(a)} \\ \vdots \\ w_n^{(a)} \end{pmatrix}$$

※ $\lambda^{(a)}$ は 0 以上の実数

※ $\lambda^{(a)} \neq \lambda^{(b)}$ のとき $\vec{w}^{(a)} \perp \vec{w}^{(b)}$

☺ 固有値方程式の転置は $\left({}^t(AB) = {}^tB\,{}^tA \text{ に注意する}\right)$

$${}^t\vec{w}^{(a)}\,{}^tXX = \lambda^{(a)}\,{}^tw^{(a)}$$

$${}^tw^{(a)}\,{}^tX X\, w^{(b)} = {}^tw^{(a)}\left({}^tX X\, w^{(b)}\right) = \lambda^{(b)}\,{}^tw^{(a)} w^{(b)}$$

$$= \left({}^tw^{(a)}\,{}^tX X\right) w^{(b)} = \lambda^{(a)}\,{}^tw^{(a)} w^{(b)}$$

$$(\lambda_a - \lambda_b)\,\vec{w}^{(a)} \cdot \vec{w}^{(b)} = 0 \iff \lambda_a \neq \lambda_b \text{ のとき } \vec{w}^{(a)} \cdot \vec{w}^{(b)} = 0$$

※ $\vec{\tilde{w}}^{(a)} = \frac{1}{\sqrt{\|\vec{w}^{(a)}\|^2}} \vec{w}^{(a)}$ とおくと $\vec{\tilde{w}}^{(a)} \cdot \vec{\tilde{w}}^{(b)} = \delta_{ab}$ 正規直交基底

$\vec{\tilde{w}}^{(a)}$ を改めて $\vec{w}^{(a)}$ と書く

※ $$y_i^{(a)} = \sum_{b=1}^{p} (x_i^{(b)} - \bar{x}^{(b)}) w_b^{(a)}$$ データの重みつきの和

各 a に対して $\vec{y}^{(a)} = \begin{bmatrix} y_1^{(a)} \\ \vdots \\ y_n^{(a)} \end{bmatrix}$ の平均は 0 , 分散は $\frac{1}{n}\lambda^{(a)}$

$a \neq b$ $(\lambda^{(a)} \neq \lambda^{(b)}$ を仮定する$)$ $Cov(y^{(a)}, y^{(b)}) = 0$

☺ $\vec{y}^{(a)} \cdot \vec{y}^{(b)} = {}^t\vec{w}^{(a)} {}^tX \cdot X \vec{w}^{(b)} = \lambda_a {}^t\vec{w}^{(a)} \cdot \vec{w}^{(b)} = \lambda_a \delta_{ab}$

$$Var(y^{(a)}) = \frac{1}{n}\sum_{i=1}^{n} (y_i^{(a)})^2 = \frac{1}{n} \vec{y}^{(a)} \cdot \vec{y}^{(a)} = \frac{1}{n}\lambda_a$$

◎ 固有値を $\lambda^{(1)} > \lambda^{(2)} > \cdots > \lambda^{(n)} > 0$ のように順番に並べる

主成分 $\vec{y}^{(1)}$: 最も分散が大きくなるデータ

第2成分 $\vec{y}^{(2)}$: 分散が2番目に大きなデータ

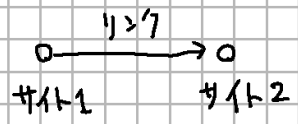
§ Google 行列と Page rank

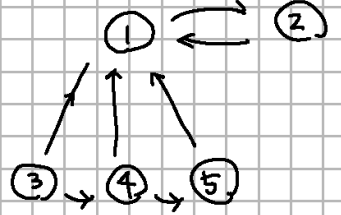
インターネット 検索 エンジン

- 条件にあてはまるサイトの中で 重要なものほど 上に出すように 工夫している

"重要性"の目安 : Page rank

※ ネットでは ユーザは 各サイトのリンクをたどって サイトを移動している





リンクは 左のような ダイアグラムで表記できる

※ どのサイトが最も 読まれるのか?

- 多くのサイトから リンクされている
- 重要なサイトから リンクされると より重要になる

※ リンクを表す行列 $A = [a_{ij}]$ j番目のサイト n_jコのリンク $i_1, \cdots, i_{n_j}$

$$a_{ij} = \begin{cases} \frac{1}{n_j} & i = [i_1, \cdots, i_{n_j}] \\ 0 \end{cases}$$

上のグラフの場合は

$$A = \begin{bmatrix} 0 & 1 & \frac{1}{3} & \frac{1}{2} & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

重要性の目安

ページランク

$$\vec{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_N \end{bmatrix}$$

N×1 ベクトル

N：サイトの数

$$p_i = \sum_j A_{ij}\, p_j$$

i番目のページの価値 $= \sum_{\text{リンク元ページ}} \dfrac{\text{リンク元ページの価値}}{\text{リンク元ページのリンク数}}$

行列表示：$\vec{p} = A\vec{p}$ ：$\vec{p}$ は A の固有値 1 に対する固有ベクトル

規格化 $\sum_{i=1}^{N} p_i = 1$, $p_i \geq 0$

例 でとりあげた 行列の場合

$$\vec{p} = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

1と2は同等の価値

3,4,5は 価値がない

極端すぎる？

偶然 ③④⑤ を見ることもありうる..

Google 行列 $G = [g_{ij}]$ $\quad g_{ij} = \alpha\, a_{ij} + (1-\alpha)\dfrac{1}{N}$

ランダムにサイトを訪れる効果

例 $\alpha = 0.8$ $\quad p = \begin{bmatrix} 0.35 \\ 0.32 \\ 0.04 \\ 0.18 \\ 0.11 \end{bmatrix}$

Google $\alpha \sim 0.85$

Google 行列の性質

- 各行列要素は正
- 任意の j に対して $\sum_{i=1}^{N} a_{ij} = 1$

固有値 1 の固有ベクトルが存在し unique であること

⇑

Perron - Frobenius の定理

▷ 行列要素が全て正の行列 G に対して

▷ ペロン・フロベニウス固有値 r

A の固有値は $|\lambda_i| < r$

⇐

▷ 固有値 r に対する固有ベクトル（ペロンベクトル）の成分は全て正

▷ それ以外の固有値に対する固有ベクトルは負の成分を持つ

Google 行列の場合

- $\vec{p}$ の成分は全て正
- 固有値 1

} $\vec{p}$ はペロンベクトル

- $\vec{p}$ 以外の固有ベクトルは負の成分を持つ

Google行列のサイズ　N：全世界のサイト数　1兆？　超巨大な行列

全ての固有ベクトル，固有値を求めるのは現実的でない

（固有値を求める方程式　$\det(\lambda I - G) = 0$　項の数 $N!$ ）

Perron－Frobeniusの定理を用いて少い計算量で p のみを得たい．

固有値方程式　$G\vec{u}_i = \lambda_i \vec{u}_i$　　$\lambda_1 = 1$　$\vec{u}_1 = \vec{p}$　$|\lambda_i| < 1$　$i \neq 1$

$$\Lambda = \begin{pmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}$$ 対角行列　　$U = [\vec{u}_1, \cdots, \vec{u}_N]$　固有ベクトルを並べたもの

$$GU = U\Lambda \quad \Rightarrow \quad G = U\Lambda U^{-1}$$

※ 任意のベクトル $\vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix}$ 条件 $v_i > 0$　$\sum_{i=1}^{N} v_i = 0$

に対して

$$\lim_{n \to \infty} G^n \vec{v} = \vec{p}$$

☺ $G^n = (U\Lambda U^\top)^n = U\Lambda \cancel{U^\top U}\Lambda U^\top \cdots \cancel{U^\top U}\Lambda U^\top = U\Lambda^n U^\top$

一方 $\Lambda = \begin{pmatrix} 1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_N \end{pmatrix}$ $|\lambda_i| < 1$ より

$$\lim_{n\to\infty} \Lambda^n = \lim_{n\to\infty} \begin{pmatrix} 1 & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_N^n \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & \ddots \\ & & & & 0 \end{pmatrix}$$

$U = [\underset{\vec{p}}{\vec{u}_1}, \vec{u}_2, \cdots, \vec{u}_n]$ $U^\top = \begin{bmatrix} \vec{w}_1 \\ \vdots \\ \vec{w}_N \end{bmatrix}$ より

$$\lim_{n\to\infty} G^n = [\vec{u}_1, \cdots, \vec{u}_n] \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \vec{w}_1 \\ \vdots \\ \vec{w}_N \end{bmatrix} = \underset{N\times 1}{\vec{u}_1} \cdot \underset{1\times N}{\vec{w}_1} \text{行列}$$

$\therefore \lim_{n\to\infty} G^n \vec{v} = \underset{=\vec{p}}{\vec{u}_1} (\vec{w}_1 \cdot \vec{v}) \propto \vec{p}$

比例係数が1であること $\sum_{i=1}^N (Gv)_i = \sum_{i,j} G_{ij} v_j = \sum_j v_j$ $\left(\sum_{i=1}^N G_{ij} = 1\right)$