work-010

PCAの簡単な導出を行う。データ集合 $\{\boldsymbol{x}_n\}\ n = 1,...,N$ を考える。$\boldsymbol{x}$ は D次元のデータ(特徴量がD個)であり、$D < N$ を仮定する。

$$\boldsymbol{x}_i = x_{i1}\boldsymbol{e}_1 + x_{i2}\boldsymbol{e}_2 + \cdots + x_{iD}\boldsymbol{e}_D$$

ここで、$\{\boldsymbol{e}_j\}$ は正規直交基底とする。 以下で第一主成分を求める。

まず、D次元単位ベクトル $\boldsymbol{u}_1$ を考える。第一主成分とは、「データの分散が最大となる方向」であるので、$\boldsymbol{u}_1^T\boldsymbol{x}$ の分散が最大となるような$\boldsymbol{u}_1$を求めればよい。

1. $\boldsymbol{u}_1^T\boldsymbol{x}$ はどのような量か。説明せよ。

分散は

$$\mathrm{Var}(\boldsymbol{u}_1^T\boldsymbol{x}) = \frac{1}{N}\sum_{i}^{N}\{\boldsymbol{u}_1^T\boldsymbol{x}_i - \boldsymbol{u}_1^T\overline{\boldsymbol{x}}\}^2 = \boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u_1}$$

となる。

ここで、$\overline{\boldsymbol{x}}$ 、$\boldsymbol{\Sigma}$ はそれぞれ $\boldsymbol{x}$ の平均、共分散行列を表す。

2. $\overline{\boldsymbol{x}}$、$\boldsymbol{\Sigma}$ を書け。
3. $\mathrm{Var}(\boldsymbol{u}_1^T\boldsymbol{x}) = \boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u_1}$ の計算過程を書け。

$\mathrm{Var}(\boldsymbol{u}_1^T\boldsymbol{x})$ を最大にする $\boldsymbol{u}_1$を求める。ラグランジュの未定乗数法を用いると$\boldsymbol{u}_1$の満たすべき条件式が以下になることがわかる。

$$\boldsymbol{\Sigma}\boldsymbol{u_1} = \lambda_{\boldsymbol{1}}\boldsymbol{u_1}$$

4. ラグランジュの未定乗数法を用いて上式を導け。
5. 求めた$\boldsymbol{u}_1$の各成分はどのような意味を持つか？
6. $\boldsymbol{u}_1$ 方向の分散を求めよ。

以上が、第一主成分を求める手続きである。上記のような方法を繰り返すことで、第二、第三主成分も同様に求めることができる。

求めた主成分を用いて各$\boldsymbol{u}_i$方向にデータ$\boldsymbol{x}_n$を射影する。具体的には以下のように表現される。

$$
\begin{aligned}
y_{n1} &= \boldsymbol{u}_1^T \boldsymbol{x}_n \\
y_{n2} &= \boldsymbol{u}_2^T \boldsymbol{x}_n \\
&\vdots \\
y_{ni} &= \boldsymbol{u}_i^T \boldsymbol{x}_n \\
&\vdots \\
y_{nd} &= \boldsymbol{u}_d^T \boldsymbol{x}_n
\end{aligned}
$$

行列とベクトルで表現すれば、

$$
\boldsymbol{y}_n = U^T \boldsymbol{x}_n
$$

となる。

7. 上式はどのような意味を持つか。説明せよ。