1　主成分分析概要

v_0 個の変数からなる N 個のデータセットがあったとき、それを v_0 次元空間の点とみて扱い、さらに適当な (以下の詳細で述べる)v_p 次元の部分空間に射影して v_p 個の変数に情報を削減する方法。

特に分散をなるべく大きくなるように射影 (=圧縮) させるので、データが区別をしやすくなる。

2　主成分分析詳細

前提として、v_0 個の変数からなる N 個のデータがあったとき、それらを v_0 次元空間の N 個のベクトルとみなす。

これら N 個のベクトルを「射影したときに分散が最大になるような方向 (ベクトル)」を見つけたい。

測定値のオリジナルのデータを x^*_{np} としておく。p は変数の次元、n はデータ数。

また各変数に対する平均をひいて、

$$x_{np} = x^*_{np} - \overline{x}_p \quad (n = 1, 2, ...N \;\; p = 1, 2, ..., P)$$

また、測定データに対し行列 X を以下で定める。

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix}$$

射影する先のベクトルを z_1 として、

$$z_1 = \sum_{p=1}^{P} w_{p1} x_p$$

と表現しておく。ここで、x_p というのはもともとの変数から由来しているので、普通のベクトル空間の言葉でいえば、x_p は標準基底のこと。射影する先のベクトル z_1 をもとの標準基底を使って表そうとしている。

$$\boldsymbol{w} = \begin{pmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{P1} \end{pmatrix}$$

としておく。

n 番目のサンプルの成分を

$$x_n = (x_{n1}, x_{n2}, \dots, x_{nP})$$

としておくと、対応する第一主成分 z_1 の値 t_{n1} は、

$$\begin{aligned} t_{n1} &= \sum_{p=1}^{P} w_{p1} x_{np} \\ &= \boldsymbol{x}_n \boldsymbol{w}_1 \end{aligned}$$

となる。

N 個のサンプルに対してこの t_{n1} を並べて、

$$\boldsymbol{t}_1 = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{N1} \end{pmatrix}$$

とおくと、

$$\boldsymbol{t}_1 = \boldsymbol{X}\boldsymbol{w}_1$$

となる。$\boldsymbol{t}_1$ の平均値、$\bar{\boldsymbol{t}}_1$ は

$$\begin{aligned} \bar{\boldsymbol{t}}_1 &= \frac{1}{N}\sum_{n=1}^{N} t_{n1} \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{w}_1 \\ &= \frac{1}{N}\sum_{n=1}^{N} \left(\sum_{p=1}^{P} w_{p1} x_{np} \right) \\ &= \frac{1}{N}\sum_{p=1}^{P} w_{p1} \left(\sum_{n=1}^{N} x_{np} \right) \end{aligned}$$

となり、x_{np} は標準化してあったので、$\sum_{n=1}^{N} x_{np} = 0$ だから、$\bar{\boldsymbol{t}}_1 = 0$ である。
z_1 の分散 $\sigma_{z_1}^2$ は、

$$\begin{aligned} \sigma_{z_1}^2 &= \frac{1}{N-1}\boldsymbol{t}_1^T \boldsymbol{t}_1 \\ &= \frac{1}{N-1}\left(\boldsymbol{X}\boldsymbol{w}_1\right)^T \left(\boldsymbol{X}\boldsymbol{w}_1\right) \\ &= \boldsymbol{w}_1^T \boldsymbol{V} \boldsymbol{w}_1 \\ &\geqq 0 \end{aligned}$$

となっている。
ここで、$\boldsymbol{V}$ は共分散行列になっていて、

$$\boldsymbol{V} = \frac{1}{N-1}\boldsymbol{X}^T \boldsymbol{X}$$

もともとの $\boldsymbol{X}$ の定義に沿って計算をすれば、

$$\begin{aligned} v_{ij} &= \frac{1}{N-1}\sum_{n=1}^{N} x_{ni} xnj \\ &= \frac{1}{N-1}\sum_{n=1}^{N} (x_{ni}^* - \overline{x}_i)(x_{nj}^* - \overline{x}_j) \end{aligned}$$

となっており、$v_{ij} = v_{ji}$ であり、つまり $\boldsymbol{V} = \boldsymbol{V}^T$ となっている。

ここで、$z_1 = \sum_{p=1}^{P} w_{p1} x_p$ という前提で話を進めてきており、さらに z_1 が単位ベクトル、つまり

$$\sum_{p=1}^{P} w_{p1}^2 = 1$$

の元で、$\sigma_{z_1}^2$ を最大にするような z_1 を求めたいのだが、これを求めるのにラグランジュの未定乗数法と呼ばれる方法を使う。

ラグランジュの未定乗数法による詳細な計算は省くが (このあたり証明とか若干忘れてるので復習しておきます。。。)、それを使って問題を変形していくと、

$$\sigma_{z_1}^2 \text{を最大にしたい} \Leftrightarrow \boldsymbol{V} w = \lambda w \text{ となる } \lambda \text{ を求めたい}$$

と、$\boldsymbol{V}$ に関する固有値問題に置き換わる。

V は実対象行列なので、求めようと思えば、固有値も固有ベクトルも変数の数と同じだけ求まるが、全部使ってはせっかく次元削減しようとしている意味がないので、通常は 2 次元とか 3 次元とかに制限して利用する。

固有値がそのまま分散に対応しているので (証明は略)、必要に応じて大きい固有値とそれに対応する固有ベクトルを選び、実際のデータはその固有ベクトルに射影してデータ分析を行う。

3　今回の授業で学んだこと、とその他

jupyter 常でまだ十分に数式を使いこなせない (出せないフォントとかがある) ので Tex でかきました。

なるべく情報を落とさずに、かつ情報を圧縮したいと両立をするアイデアが勉強になりました。感想になってしまうけど。