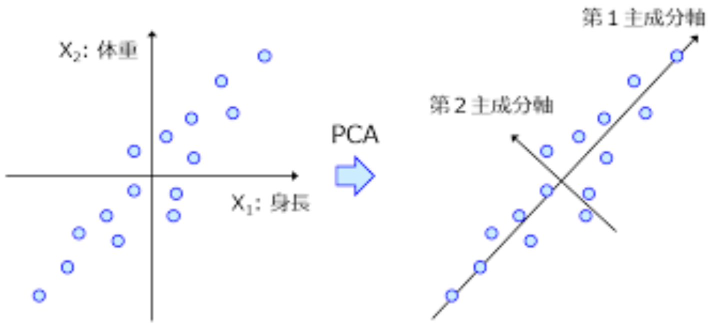
主成分分析

次元の高いデータを高いデータを寝るべく**情報** を失わないように低次元に圧縮する方法。

機械学習では**次元の呪い**と言われることがあり、次元を増やすと特徴量が増え、計算量が指数関数的に増える。そこで次元を減らす方法として主成分分析が用いられる。主成分分析は教師なし学習の一種で、データから特徴抽出と次元の圧縮を自動に行う。例えば、身長・体重という2の特徴量で表現される人のデータセットが会ったとき、その**情報量**をなるべく失わないように2次元からより低い次元、つまり１個（次元）でデータセットを表現できる。

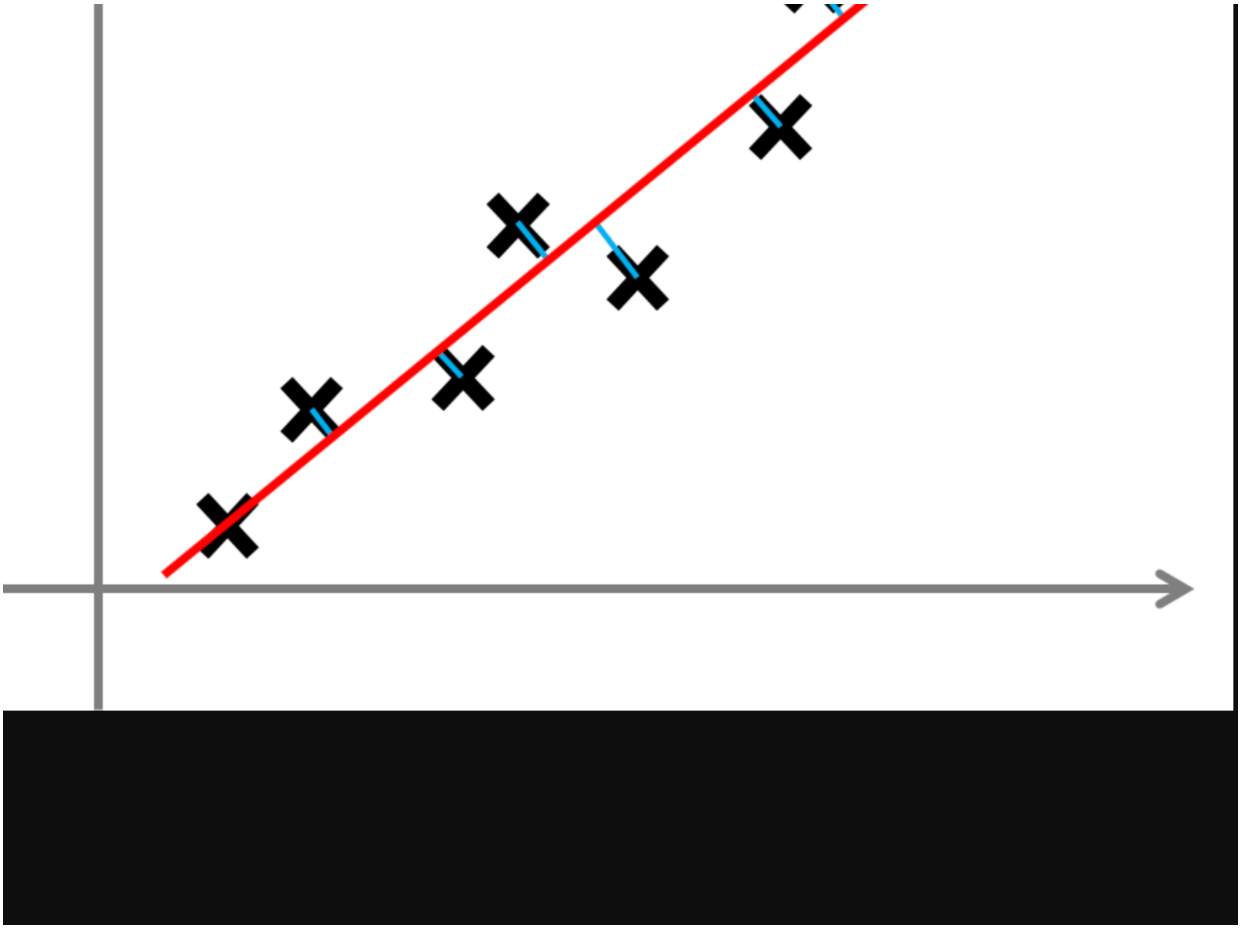


以上のように２次元のデータを１次元に圧縮する場合、どんな直線を引くと最も**情報**が失われないか。ここでいう情報を失はないとは下図の青い線分の長さを最小にするということ。

射影先の直線(ベクトル)を求める作業において「集団を最もよく表現するベクトルと」とは **青い線分の長さの２乗和を最小化するもの。(=射影先のベクトルと各データの距離)**

この値（誤差）が**小さい**ほど、それぞれのデータをよく**保存している**という。

残された**一次元データがなるべくバラついている、つまり分散(残された１次元データがなるべくバラついている。)の値が大きいほどよい**。なぜなら、バラついているということは、各データ点１つ１つの違いをより多く情報として保っていることになるからである。逆に言えば、バラツキの少ない方向というのは、各データが共通して持っている自明な情報なので削除してしまっても問題ない。



数式を用いて説明しなさい

ある方向の単位ベクトルを$\vec{e} = \begin{pmatrix} e_x \\ e_y \end{pmatrix}$と表す。データを$\vec{e}$の方向に射影して、1次元データとなった後の分散値の表式を求める。例えば、$\vec{a}_1$を$\vec{e}$に射影した値は$\vec{a}_1 \cdot \vec{e}$のようにベクトルの内積を取ることで簡単に計算でき、これを用いて分散(var)を定義通りに計算すると、

$$\mathrm{var} = \frac{(\vec{a}_1 \cdot \vec{e})^2 + (\vec{a}_2 \cdot \vec{e})^2 + (\vec{a}_3 \cdot \vec{e})^2}{3}$$

となる(平均値は0となっていることに注意)。これを行列の積の形に分解して表現してみると、次に示す式の一行目のようになる。実は、この分解した行列がそれぞれ$(X\vec{e})$とその転置行列$(X\vec{e})^{\mathrm{T}}$になっている(実際に計算してみれば確認できる)。二行目から三行目では行列の性質を用いている。こうすると$X^{\mathrm{T}}X$という並びが出現する。これは上で計算した共分散行列Σである。

$$
\begin{aligned}
\mathrm{var} &= \frac{1}{3}\begin{pmatrix} \vec{a}_1 \cdot \vec{e} & \vec{a}_2 \cdot \vec{e} & \vec{a}_3 \cdot \vec{e} \end{pmatrix} \begin{pmatrix} \vec{a}_1 \cdot \vec{e} \\ \vec{a}_2 \cdot \vec{e} \\ \vec{a}_3 \cdot \vec{e} \end{pmatrix} \\
&= \frac{1}{3}(X\vec{e})^{\mathrm{T}}(X\vec{e}) \\
&= \frac{1}{3}\vec{e}^{\mathrm{T}} X^{\mathrm{T}} X\vec{e} \\
&= \vec{e}^{\mathrm{T}} \Sigma \vec{e}
\end{aligned}
$$

このvarの表式はデータが多成分(つまり高次元)、多データ化しても一般的に成り立つはずである。なぜなら、行列Xやベクトル$\vec{e}$の縦成分と横成分数をデータ次元数とデータ数に併せて拡張すれば同じような計算が行えるからである。

varが最大になる方向を求める

さて、主成分分析ではこのvarを最大化するような$\vec{e}$を求めることが目的であった。また、$\vec{e}$には単位ベクトル、つまりノルムが1であるという制約条件も付いている。このように、ある制約条件のもとで、関数の最大(最小)を決定する時には、「ラグランジュの未定乗数法」というものがよく用いられる。

ラグランジュの未定乗数法

:

　変数x,yについて、「$g(x,y)=0$」の制約条件下で「$f(x,y)$」という関数を最大(小)化するのは、

$L(x,y,\lambda) \equiv f(x,y) - \lambda g(x,y)$

　という関数を定義した時に、

$\frac{\partial L}{\partial x} = 0, \frac{\partial L}{\partial y} = 0, \frac{\partial L}{\partial \lambda} = 0$

　を満たすような解(x,y)である。

ラグランジュの未定乗数法を用いると、

$$
\begin{aligned}
L(e_x, e_y, \lambda) &\equiv \mathrm{var} - \lambda(e_x^2 + e_y^2 - 1) \\
&= \begin{pmatrix} e_x & e_y \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} - \lambda(e_x^2 + e_y^2 - 1)
\end{aligned}
$$

と定義した関数$L(e_x, e_y, \lambda)$に対して、

$$
\begin{aligned}
\frac{\partial L(e_x, e_y, \lambda)}{\partial e_x} = 0 \\
\frac{\partial L(e_x, e_y, \lambda)}{\partial e_y} = 0
\end{aligned}
$$

が成立するようなe_x, e_yを見つければよい(λについての偏微分は一旦無視する)。 $\frac{\partial L(e_x, e_y, \lambda)}{\partial e_x}$, $\frac{\partial L(e_x, e_y, \lambda)}{\partial e_y}$を計算すると、

$$
\begin{aligned}
\frac{\partial L(e_x, e_y, \lambda)}{\partial e_x} &= \begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} + \begin{pmatrix} e_x & e_y \end{pmatrix} \Sigma \begin{pmatrix} 1 \\ 0 \end{pmatrix} - 2\lambda e_x \\
&= \begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} + \left[\begin{pmatrix} e_x & e_y \end{pmatrix} \Sigma \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right]^{\mathrm{T}} - 2\lambda e_x \\
&= \begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} + \begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma^{\mathrm{T}} \begin{pmatrix} e_x \\ e_y \end{pmatrix} - 2\lambda e_x \\
&= 2 \begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} - 2\lambda e_x \\
\frac{\partial L(e_x, e_y, \lambda)}{\partial e_y} &= 2 \begin{pmatrix} 0 & 1 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} - 2\lambda e_y
\end{aligned}
$$

であり、これらが0になるので
$$
\begin{aligned}
\begin{pmatrix} 1 & 0 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} &= \lambda e_x \\
\begin{pmatrix} 0 & 1 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} &= \lambda e_y
\end{aligned}
$$

と式変形出来る。2つの方程式をまとめると、 $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} = \lambda \begin{pmatrix} e_x \\ e_y \end{pmatrix}$

一番左の行列はただの単位行列なので省略して、 $\Sigma \begin{pmatrix} e_x \\ e_y \end{pmatrix} = \lambda \begin{pmatrix} e_x \\ e_y \end{pmatrix}$

これは共分散行列Σに対する固有方程式となっている

(上で無視していた$\frac{\partial L(e_x, e_y, \lambda)}{\partial \lambda} = 0$、つまり$e_x^2 + e_y^2 = 1$というベクトルの正規化条件は、**固有値問題**を解いた後の固有ベクトルを拡大縮小することで調節出来る。) 実際にデータを主成分分析する際には、データから**共分散行列**を生成し、固有ベクトルを計算、全データを射影する、といった作業を行えば良い。 上で最終的に導出した固有方程式は2×2の行列についてのものなので、固有値λと固有ベクトル$\vec{e}$は2種類出現する。これは簡単に決めることが出来る。手に入れた固有方程式$\Sigma \vec{e} = \lambda \vec{e}$の表式から、分散varをさらに計算してみる。

$$
\begin{aligned}
\mathrm{var} &= \vec{e}^{\,\mathrm{T}} \Sigma \, \vec{e} \\
&= \vec{e}^{\,\mathrm{T}} \lambda \, \vec{e} \\
&= \lambda \, (\vec{e}^{\,\mathrm{T}} \cdot \vec{e}) \\
&= \lambda
\end{aligned}
$$

つまり、固有値λは分散varの値そのものだったのである。主成分分析では分散varを最大化することが目的であったので、固有値λは2つの内大きい方を選択すれば良い

その他、今回の授業で学んだことを記述してください。

次元を圧縮して、過学習を防いだり・計算量を指数関数的に減らせるPCAは実用的だなぁと思いました。