主成分分析 (PCA: Principal Component Analysis)

清家大嗣

平成 30 年 9 月 21 日

1　主成分分析について

主成分分析とは、n 次元のベクトルデータ $\boldsymbol{x}_n$ を m 次元ベクトルデータに圧縮することである。圧縮するために、主成分分析ではいくつかの自明ではない仮定を置く。最初に、各データ $\boldsymbol{x}_n$ を、そのデータ群の平均値 $\bar{\boldsymbol{x}} = \sum_{n=1}^{N} \boldsymbol{x}_n$ を引いた値で再定義する。つまり、

$$\boldsymbol{x}_{n,\mathrm{new}} \equiv \boldsymbol{x}_{n,\mathrm{original}} - \bar{\boldsymbol{x}} \tag{1}$$

である。この $\boldsymbol{x}_{n,\mathrm{new}}$ を改めて $\boldsymbol{x}_n$ として利用する。つまり、主成分分析では平均値の情報が失われることになる。言い換えると、主成分分析は平均値が 0 の n 次元データを、同じく平均値が 0 の m 次元データに落とす手法である。平均値 ($n \to \infty$ で真の平均値が得られる) がデータセットに与える影響がないという仮定は、まったく非自明であるため主成分分析利用者はそのことについて注意する必要がある。

その非自明の前提の下、平均値を差し引いた n 次元データは $\boldsymbol{x}_n$ は n 個の正規直交基底によって表現されていると言え、同じく m 個の正規直交基底で表現する方法を考える。つまり、データセットをよく表現する互いに直交した n 次元ベクトル $\boldsymbol{w}_i$ ($|\boldsymbol{w}_i| = 0$) ($i = 1, 2, ..., m$) を探す問題について議論する。データセット全体について考える必要があるため、N 個の n 次元データ列ベクトルを列方向に並べた行列 $\boldsymbol{X}$ を $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$ を定義する。ここで、$\boldsymbol{w}_i$ が $\boldsymbol{x}_n$ をよく表現するとは、どのようなことか、$n = 2$ の場合について考えることで直感的に理解する。図 1 は、基底ベクトル $\boldsymbol{w}$ ($|\boldsymbol{w}| = 1$) と 1 つのデータ $\boldsymbol{x}_n$ との間で内積を取り、そのスカラー値に基底ベクトル $\boldsymbol{w}$ をかけている図である。

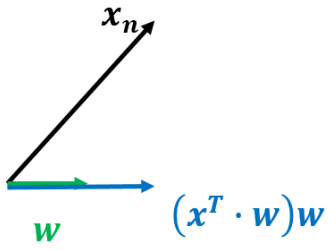

図 1: $n = 2$ の場合、$\boldsymbol{x}_n$ と $\boldsymbol{w}$ の内積

つまり、このベクトル $(\boldsymbol{x}_n^T \boldsymbol{w})\boldsymbol{w}$ が大きいほどよくデータを表現していると言える [1]。従って、このベクトルの大きさの絶対値 $\boldsymbol{x}_n^T \cdot \boldsymbol{w}$ を全サンプルに対して足した場合に最大にするような $\boldsymbol{w}$ を探すことを目的とする。つまり、次式を最大にする $\boldsymbol{w}$ を見つければよい。

[1] 単純な平方完成によって, 誤差ベクトル $\boldsymbol{x}_n - a\boldsymbol{w}$ のノルムを最小にする a は内積 $\boldsymbol{x}_n \boldsymbol{w}$ となる

$$|\boldsymbol{X}^T\boldsymbol{w}|^2 = (\boldsymbol{w}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{w}\Big(\boldsymbol{X}\boldsymbol{X}^T\Big)\boldsymbol{w} \tag{2}$$

ここで、2 つの証明のアプローチがある。(1) $\boldsymbol{w}$ が $|\boldsymbol{w}| = 1$ という拘束条件を満たしながら上式の最大値を求めている。従って、ラグランジュの三定係数法を用いることができる。(2) 実対称行列の性質を利用して、固有値分解を行い、そこから二次形式の最小・最大値を得る。この両方について説明したい。また、表記を簡単にするため 2 式を $XX^T \equiv 1/N \cdot XX^T$ と再定義し、$A \equiv 1/NXX^T$ とする。これは、2 次形式の対象行列を共分散行列とするためと、A による 2 次形式の最小・最大化問題にして見通しをよくするためである。

また、この 2 つの手法では次の実対象行列の性質をよく利用している。

- 実対称行列の固有値は実数となる。つまり、固有ベクトルも実ベクトルとなる。
- 実対称行列の固有値の異なる任意の 2 つの固有ベクトルは常に直交する。

1.1 ラグランジュの未定係数法による手法

2 式より、共分散行列 A の定義から、次のような関数 L を定義する。ここで、$|\boldsymbol{w}| = 1$ という拘束条件を利用している。

$$L(\boldsymbol{w}) = \boldsymbol{w}^T A\boldsymbol{w} - \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1) \tag{3}$$

ここで、3 式を最大・最小にする $\boldsymbol{w}$ の条件は $\boldsymbol{w}$ で微分した値が 0　となることが必要条件となるため、

$$(A + A^T)\boldsymbol{w} = 2\lambda\boldsymbol{w} \tag{4}$$

を満たす。共分散行列 A は実対称行列でもあるので、次式が得られる。

$$A\boldsymbol{w} = \lambda\boldsymbol{w} \tag{5}$$

これを元の 2 式に $1/N$ をかけた式に代入すると

$$\frac{1}{N}\boldsymbol{X}^T\boldsymbol{w}|^2 = \boldsymbol{w}^T A\boldsymbol{w} = \boldsymbol{w}^T A\boldsymbol{w} = \boldsymbol{w}^T(\lambda\boldsymbol{w}) = \lambda \tag{6}$$

実対称行列の固有値 λ は常に実数となるため、固有値最大に対応する固有ベクトルを用いればよい (固有ベクトルは互いに直交するので、まさに $\boldsymbol{w}_i$ $(i = 1, 2, ..., m)$ の定義として適切、仮に n_k 個の重解を持つ固有値が選ばれた場合、その固有ベクトル空間内の任意の n_k 個の 1 次独立なベクトルに対してグラムシュミットの直交化を用いる)。

1.2 対称行列の固有値分解を利用した手法

n 変数で、かつ二次形式の行列 A が実対称行列になる 2 次形式の最大・最小を求める問題について考えたい。実は、実対称行列 A は重解も含めると n 個の固有値を持ち、その最小値と最大値を λ_1, λ_n とすると, 任意の大きさが 1 の n 次元ベクトル $\boldsymbol{x}$ に対して、$\lambda_1 \leq \boldsymbol{x}^* A\boldsymbol{x} \leq \lambda_n$ が成立する。これは次のようにして、証明できる。

任意の n 次元ベクトル $\boldsymbol{x}$ は、実対称行列 A の互いに直交する n 個の大きさ 1 の固有ベクトルを正規直交基底としてユニークに表現できる (実対称行列の性質のファイルを参照, 実対称行列 A は、n 次元空間の正規直交基底となるような n 個の固有ベクトルを持つ)。従って、適当な定数をかけることで $\boldsymbol{x} = \sum_k r_k \boldsymbol{x'}_k$ ($\boldsymbol{x'}_k$ は互いに直交な固有ベクトル, $\sum_k |r_k|^2 = 1$) とできる。つまり、次式のようになる。

$$\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} = (\sum_k r_k \boldsymbol{x'}_k)^* \boldsymbol{A} (\sum_k r_k \boldsymbol{x'}_k) \tag{7}$$

$$= (\sum_k r_k \boldsymbol{x'}_k)^* (\sum_k \lambda_k r_k \boldsymbol{x'}_k) \tag{8}$$

$$= (\sum_k \lambda_k |r_k|^2) \tag{9}$$

つまり、

$$\lambda_1 \leq \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} \leq \lambda_n \tag{10}$$

となる。ここで用いた重要な性質は、実対称行列 A から適当に n 個の固有ベクトルを選び正規化すると、正規直交基底を得られるということである。これを次のようにして証明する。$|A - I\lambda|\boldsymbol{x} = \boldsymbol{0}$ を利用して固有ベクトルを求める際に、代数学の基本定理から、複素解と重解を含めて、

$$|A - \lambda I| = (\lambda - \lambda_1)^{n_1} \cdot ... \cdot (\lambda - \lambda_p)^{n_p} = 0 \tag{11}$$

と表せる。ここで、等号が成立する場合は固有値に対応した固有ベクトルの係数のみが 0 でない場合である。従って、値が大きい固有値から順にその固有値に対応した固有ベクトルを選んでいく (重複度 l が 2 以上の場合、その固有ベクトル空間から互いに直交する固有ベクトルを n_k 個選択する) のが主成分分析ということである。

また、2 式は $|\boldsymbol{w}| = 1$ を満たした状態で 2 式を最大化するような $\boldsymbol{w}$ を探していた。この式は、次のように変形することで、各成分と主成分分析によって得られた主成分との誤差を最小化する問題と一致していることも確認できる。このことは、[1] の 5.1 式において主成分と自己符号化器との関連を説明することに用いられている。

$$\begin{aligned}
\boldsymbol{w} &= \max_{\boldsymbol{w} \in \mathcal{R}^n, |\boldsymbol{w}|=1} \boldsymbol{w} \left(\boldsymbol{X} \boldsymbol{X}^T \right) \boldsymbol{w} \\
&= \max_{\boldsymbol{w} \in \mathcal{R}^n, |\boldsymbol{w}|=1} \sum_{n=1}^{N} \left((\boldsymbol{x}_n^T \cdot \boldsymbol{w}) \boldsymbol{w} \right)^T (\boldsymbol{x}_n^T \cdot \boldsymbol{w}) \boldsymbol{w} \\
&= \min_{\boldsymbol{w} \in \mathcal{R}^n, |\boldsymbol{w}|=1} \sum_{n=1}^{N} |\boldsymbol{x}_n|^2 - \left((\boldsymbol{x}_n^T \cdot \boldsymbol{w}) \boldsymbol{w} \right)^T (\boldsymbol{x}_n^T \cdot \boldsymbol{w}) \boldsymbol{w} \\
&= \min_{\boldsymbol{w} \in \mathcal{R}^n, |\boldsymbol{w}|=1} \sum_{n=1}^{N} |\boldsymbol{x}_n - (\boldsymbol{x}_n^T \cdot \boldsymbol{w}) \boldsymbol{w}|^2
\end{aligned} \tag{12}$$

参考文献

[1] MLP シリーズ 「深層学習」