1.主成分分析についてまとめてください。（自分が理解できていることを採点者に伝えてください。）

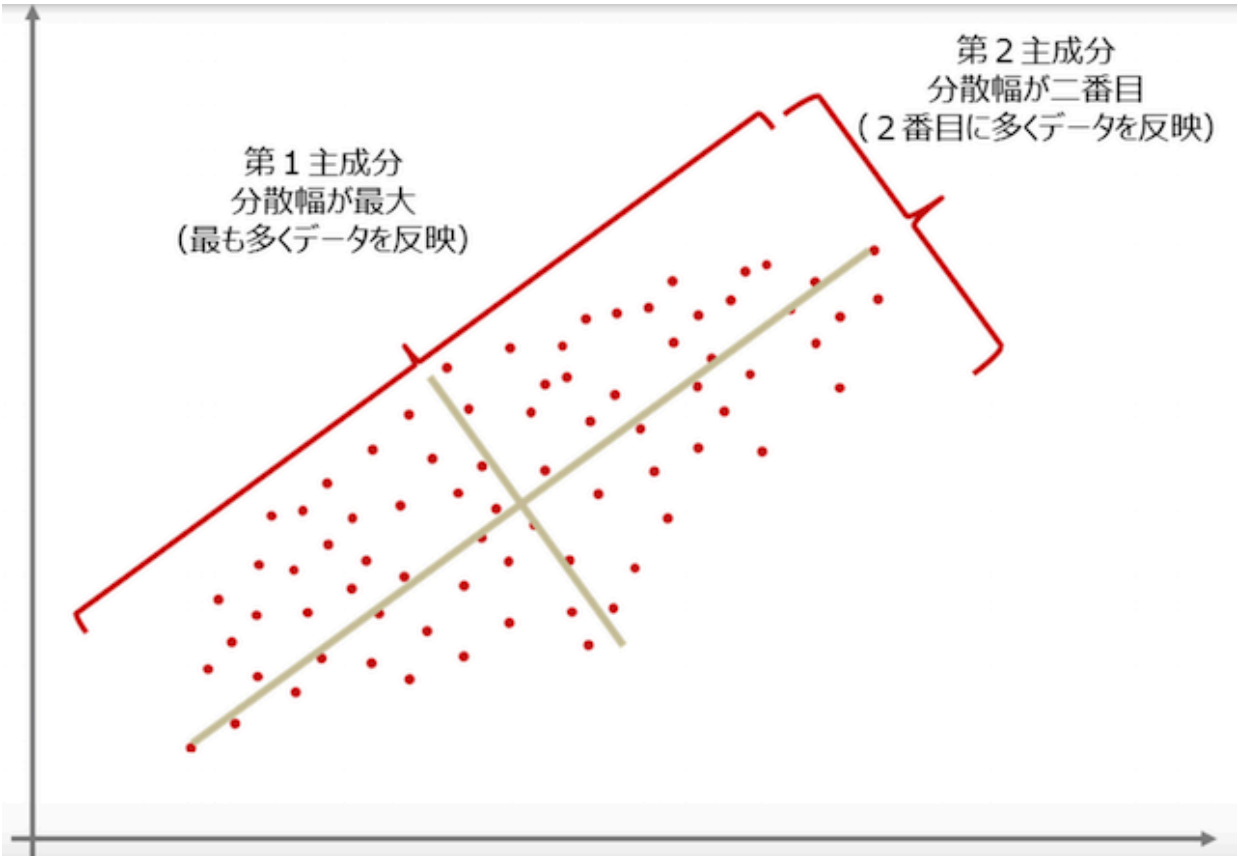
【回答】

主成分分析（principal component analysis:PCA）とは教師なし学習の一つです。 データの分散（ばらつき）が大きいところ（主成分）を見つけることにより、評価軸（次元）を削減することが可能です。データの分散が大きいところが大事で、小さいところは気にしないようにします。

主成分分析の概念は、「説明変数１」、「説明変数２」、・・・「説明変数P」から「目的変数」を導き出すことです。例えば、x軸に「観客動員数」、y軸に「DVD売上枚数」という散布図があるとします。以下の図をイメージしてください。その場合、**「観客動員数」「DVD売上枚数」が説明変数**であり、左下から右上に伸びている線を「総合人気度」とした場合、この**「総合人気度」が目的変数**となります。

主成分分析の注意点は4つ

①主成分分析における目的変数は、実在する変数ではなくて、”想像の産物”である



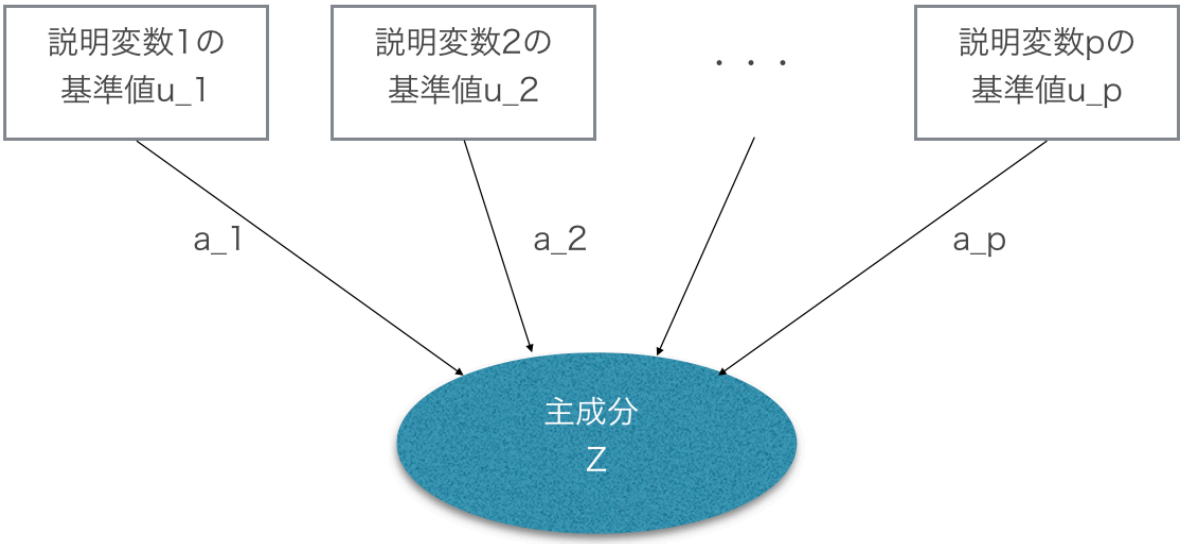
②主成分分析の計算方法には、分析対象のデータを、「基準化する」場合と「基準化しない」場合の2種類ある

今回の例では、「観客動員数（人数）」と「DVD売上枚数（枚数）」で違う単位のため、基準化（単位を揃える）必要があります。

③主成分分析の構造を式と図で表すと以下となる

zは主成分、u_1は説明変数1の基準値、u_2は説明変数2の基準値、u_pは説明変数pの基準値である。a_1やa_2は各説明変数が、主成分に与える影響の度合いです。重みのようなものです。

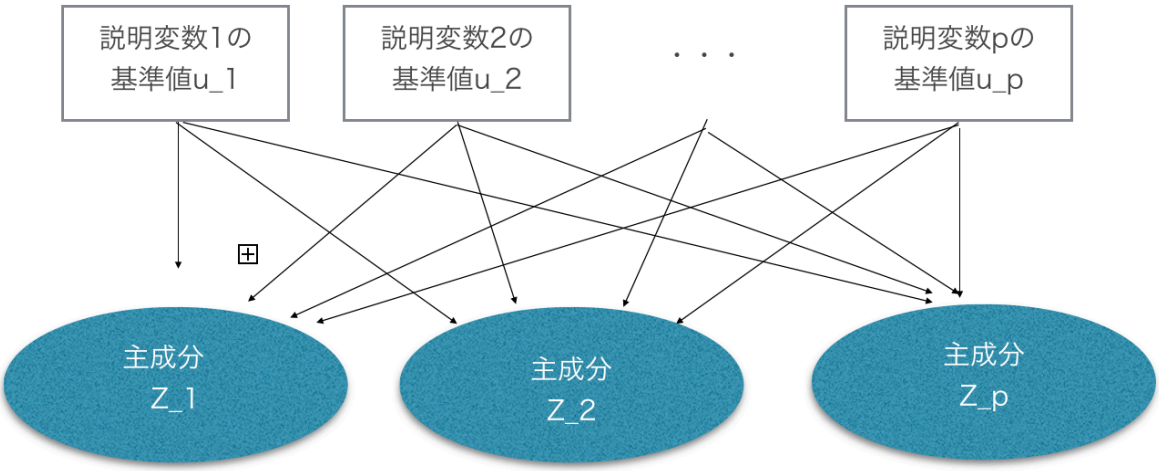
$$式 \quad : \quad z = a_1 u_1 + a_2 u_2 \cdots a_p u_p$$



④主成分の個数について、主成分は説明変数の個数だけ求まる

説明変数がp個なら主成分はp個求まる。そのうち、「総合人気度」に相当するものは、第1主成分になります（ちなみに各主成分は直交する）。また、主成分分析では、第1主成分と第2主成分だけを求めて、それを２次元の点グラフで表すのが一般的です。

$$\begin{aligned} z_1 &= a_{11} u_1 + a_{12} u_2 + \cdots + a_{1p} p u_p \\ z_2 &= a_{21} u_1 + a_{22} u_2 + \cdots + a_{2p} p u_p \\ &\cdots\cdots \\ z_p &= a_{p1} u_1 + a_{p2} u_2 + \cdots + a_{pp} p u_p \end{aligned}$$



以下例をもとに、主成分分析を行う手順（流れ）を記載します。

<<課題>>

ラーメン店を、麺、具、スープをそれぞれ５段階で評価しているデータがあるとします。このデータに対して主成分分析を行って、「ラーメン店の総合評価」を行います。

主成分分析の流れ

①主成分と主成分得点を求める
step1：変数ごとに基準化する
step2：相関行列を求める
step3：固有値λと固有ベクトルを求める
step4：step3における以下をもとに点グラフを描く
最大の固有値に対応する固有ベクトル
最大から2番目の固有値に対応する固有ベクトル
step5：step4より、第１主成分と第2主成分を確認する
step6：各個体の第1主成分における座標と第2主成分における座標を、つまり各個体の第1主成分得点と第2主成分得点を求める
step7：step6における第1主成分得点と第2主成分得点をもとに点グラフを描く

②分析結果の精度を確認する
③分析結果を検討する

2.主成分分析について素人にも分かるように簡潔に説明してください。

【回答】

野球選手の身長や体重、打率、出場試合数、出身高校などの様々なデータをもとにその選手の「試合成績」との比較を行いたいとします。
その時に、何を基準に比較すれば良いか、説明変数（＝打率など）がたくさんあり、選ぶのは難しいです。このような時に、主成分分析を行うと、いくつかある説明変数についてまとめられた新しい軸を作成することが出来ます。

この出来あがった新たな評価軸であれば理解しやすい図になります。このように、多数ある評価軸を2軸まで減らして把握することに優れている手法が主成分分析と言います。

3.主成分分析について数式を用いて説明してください。

> また、以下のキーワードを用いて説明すること
> 共分散行列
> 固有値問題

【回答】

<分散と共分散>
第1主成分を射影する軸の方向はデータが最もばらつく方向にります。 このデータのばらつきの指標に分散：V_xがあります。式で書くといかになります。

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

1〜n個までの「データx_i」からそれぞれ「平均$\bar{x}$」引いたもの（これを偏差といいます）の2乗を全部足し合わせてnで割った数が分散となります。平均から離れているデータが多いほど分散が大きくなります。逆に一定の値をとるようなデータであれば分散は0に近づいていきます。また、$(x_i - \bar{x})^2$は2乗をとっていますので、xが実数である限りは分散は0以上となります。

分散はxという1つのデータのみ扱いましたが、今度は別のデータyが登場します。この2つから共分散S_{xy}を計算します。

$$S_{xy}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

分散の式に似ていますが、共分散はx と yの偏差のかけ合わせた数を足し合わせています。同じ数の2乗の足し合わせだった分散とは違って共分散は正の値も負の値もどちらもとることができます。共分散は2つのデータがどのような傾向にあるかを簡単に示す指標になっています。例えばxが正の値をとるとき、yが負の値をとるとすると共分散は負の値となります。両方とも正の値をとる場合、分散は正の値となるわけです。グラフにすると傾きが正の分布は共分散は正、傾きが負の分布は共分散が負になりやすいといえます。

<分散共分散行列>

高次元のデータを考えたときに分散、共分散の組み合わせがたくさんになってしまい書くのが面倒になりますので行列にまとめてしまおうというのが分散共分散行列です。わかりやすくする為、ここではx と yの2次元データを考えていきましょう。まずはこんな行列を定義します。

$$X = \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{bmatrix}$$

x と yの偏差を1〜nまで並べていっています。その行列にこんなことをします。

$$\begin{aligned}\frac{1}{n}X^TX &= \frac{1}{n}\begin{bmatrix} x_1 - \bar{x} & \dots & x_n - \bar{x} \\ y_1 - \bar{y} & \dots & y_1 - \bar{y} \end{bmatrix}\begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} (x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 & (x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y}) \\ (x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y}) & (y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2 \end{bmatrix} \\ &= \begin{bmatrix} V_x & S_{xy} \\ S_{xy} & V_y \end{bmatrix}\end{aligned}$$

途中式は面倒でしたが見ごとにxとyの分散、共分散からなる行列を計算することができました。

<主成分分析>

次に、主成分分析について考えていきたいと思います。今回もx と yの2次元データを使って考えていきましょう。主成分分析は2次元データの場合、縦軸と横軸を回転させた新しい座標系にデータを変換させる処理です。では変換後のデータをzと置くとx と yの関係はこのように表すことができます。zは第1主成分なのか、第2主成分なのか今はおいておきます。

$$z_i = a_1 x_i + a_2 y_i$$

(1)

$$z_i = a_1 x_i + a_2 y_i$$

ここでおもむろにzの分散：V_zを求めてみます。この分散が最大になるようz、もとい変換に使われるパラメタのa_1、a_2を求めることが主成分分析のポイントです。V_zは以下の式で表せます。

(2)

$$
\begin{aligned}
V_z &= \frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(a_1x_i+a_2y_i)-(a_1\bar{x}+a_2\bar{y})\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{a_1(x_i-\bar{x})+a_2(y_i-\bar{y})\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{a_1^2(x_i-\bar{x})^2+2a_1a_2(x_i-\bar{x})(y_i-\bar{y})+a_2^2(y_i-\bar{y})^2\right\} \\
&= a_1^2V_x+2a_1a_2S_{xy}+a_2^2V_y
\end{aligned}
$$

V_zをxとyの分散と共分散で表すことができました。 今度はa_1とa_2についてもう少し考えてみます。a_1とa_2をそれぞれ$a_1=\cos\theta_1$と$a_2=\cos\theta_2$と置きます。θ_1とθ_2はそれぞれzに変換するための軸とx軸、y軸との角度になっています。詳しいことは省きますが、このようなa_1とa_2を方向余弦といいます。このように定義することでa_1とa_2には以下の関係というか制限ができます。

(3)

$$a_1^2+a_2^2=1$$

式(2)と式(3)を使って、V_zが最大になるa_1とa_2を求めていきます。式(2)のような方程式と式(3)のような条件式がそろうとLagrangeの未定乗数法を使って答えを出すことができます。色々出てきて頭がパンクしてしまいましたが、もう少しです。 Lagrangeの未定乗数法とは「求めたい変数の定義式：式(2) – λx条件式：式(3)」を関数Fとして変数a_1、a_2、λの偏微分が0になるようにa_1、a_2、λを求めるとV_zが最大になるという魔法のような解法です。.....お気持ちはわかります。この段階でしれっと変数が追加された憤りはひしひしと。では関数Fと偏微分を解いていきましょう。 関数F

(4)

$$F(a_1,a_2,\lambda)=a_1^2V_x+2a_1a_2S_{xy}+a_2^2V_y-\lambda(a_1^2+a_2^2-1)$$

偏微分式

(5)

$$\frac{\partial F(a_1,a_2,\lambda)}{\partial a_1}=2a_1V_x+2a_2S_{xy}-2\lambda a_1=0$$

(6)

$$\frac{\partial F(a_1,a_2,\lambda)}{\partial a_2}=2a_1S_{xy}+2a_2V_y-2\lambda a_2=0$$

(7)

$$\frac{\partial F(a_1,a_2,\lambda)}{\partial \lambda}=-a_1^2-a_2^2+1=0$$

式(7)は式(3)と同様なので用済みです。式(5)と式(6)は下のような行列式に書き換えることで知っている人なら知っている固有値問題になります。さっさとやってしまいましょう。

(8)

$$\begin{bmatrix}a_1V_x+a_2S_{xy}\\a_1S_{xy}+a_2V_y\end{bmatrix}=\begin{bmatrix}V_x & S_{xy}\\S_{xy} & V_y\end{bmatrix}\begin{bmatrix}a_1\\a_2\end{bmatrix}=\lambda\begin{bmatrix}a_1\\a_2\end{bmatrix}$$

ここから分散共分散行列の固有方程式を解くことでa_1、a_2、λが求まります。この場合は以下の式を解いていきます。

(9)

$$\left|\begin{bmatrix} V_x & S_{xy} \\ S_{xy} & V_y \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = \left|\begin{bmatrix} V_x - \lambda & S_{xy} \\ S_{xy} & V_y - \lambda \end{bmatrix}\right|$$
$$= (V_x - \lambda)(V_y - \lambda) - S_{xy}^2 \quad = 0$$

上記の通りλの2次方程式となりましたので、2つのλが求まったことになります。それぞれのλからa_1、a_2を求めることができます。2つのλのうち大きいほうが第1主成分、小さいほうが第2主成分となっており、それぞれに対応するa_1、a_2が変換軸を表すベクトルになっています。

4.主成分分析をPythonで実装してください。

【回答】

以下のファイルをご参照ください。 AIF_week2*授業課題*python実装.ipynb

5.その他、今回の授業で学んだことを記述してください。

【回答】

共分散行列について、（特に数式について）理解しているとは言い難い状況です。 継続して理解するようにしたいと思います。