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。）

主成分分析(principal component analysis:PCA)とは、与えられたデータの傾向から機械学習によって自動的に特徴量を見つけ出し、その特徴をよく表す低次元データへと次元圧縮をする手法の一つです。

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。

肥満度を測るBMIという指標で説明します。 BMIは（体重）÷（身長 $)^2$という計算式で導出されるものですが、元のデータは（身長、体重）という２成分を持つデータであったのに対し、BMIはただ１成分の数値となっています。これを次元の圧縮といい、その結果、BMIは情報を削ぎ落としたにも関わらず、肥満度という特徴を表すのに十分な機能を持っています。このようにデータに適切な処置を行えば、情報量の削減と特徴の抽出とを同時に行い、人間がデータの関係性を把握しやすくすることができるようになります。以上のようなデータに行う処置が主成分分析です。

主成分分析について数式を用いて説明してください。

二つの変量x、yを持つサンプルデータ（点）が多数あるとする。例えば、x軸が国語の点数、y軸が数学の点数で、点は一人の生徒を表す場合等である。 一人一人の違いを見るとき、x軸の国語の側から見るのと（国語の点数を比較する）、y軸の数学の側から見る事が出来る。 しかし二つの方向から眺めて一眼で違いを判断することは難しい。またサンプルデータの次元が増えればなおさらのことである。もし生徒の散らばり方の情報を最も保持した新しい一つ次元を減らした軸から見れば、一人一人の違い（各点の離れ方）は一目瞭然である。主成分分析は、この様な線を引く。それは座標軸の回転と考える事も出来る。
このように主成分分析とは、新たな座標軸における各点の座標が最もばらけるように（違いが分かるように）、低次元の座標軸を引くことである。そのような座標軸を引くことを計算で求める。 ここで新しい座標軸の単位ベクトルを $(a, b) = (cos\theta, sin\theta)$ とする。 まず、元の座標でデータの中心化を行い分散を求める。中心化とは、其々の軸の平均値に原点を移動する。分散とは、各データの値$(x_1, x_2, \cdots)$から全データの平均値$\overline{x}$を引いた値（偏）差の2乗の総和をデータの個数で割ったもの。従って、中心化させる事で各データの値がそのまま偏差となって計算が簡単になる。偏差の2乗を取るのは、原点を挟んでプラスとマイナスがあるので、相殺されるのを防ぐためである。 そこで、ある点k(中心化後の座標をx_k、y_kとする）と新しい座標軸の単位ベクトルと内積をD_kとすると、$D_k = ax_k + by_k (a = cos\theta、b = sin\theta)$となる。その2乗は以下のようになる。

$$(D_k)^2 = (ax_k + by_k)^2 = a^2 x_k^2 + b^2 y_k^2 + 2abx_k y_k$$

分散（var）を求めるには、全てのデータの２乗の和を求めてデータの個数 (n) で割ればいい。以下のように式が展開されるが、a^2, b^2は$(a, b) = (cos\theta, sin\theta)$ で定数であるため括り出せる。

$$var = \frac{1}{n}\sum_{k=1}^{n} D_k^2 = \frac{1}{n}\sum_{k=1}^{n}(ax_k + by_k)^2 = \frac{1}{n}\sum_{k=1}^{n}(a^2 x_k^2 + b^2 y_k + 2abx_k y_k)$$

$$= a^2 \frac{1}{n}\sum_{k=1}^{n} x_k^2 + b^2 \frac{1}{n}\sum_{k=1}^{n} y_k^2 + 2ab\frac{1}{n}\sum_{k=1}^{n} x_k y_k \cdots (A)$$

(A)式のa、b以外のところはそれぞれ、回転前の座標軸のx座標の分散（中心化後なので平均0の分散）、y座標の分散、x座標y座標の共分散となっている。これを以下のように定める。

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2 \Rightarrow S_x : x\text{座標の分散}$$

$$\frac{1}{n}\sum_{k=1}^{n} y_k^2 \Rightarrow S_y : y\text{座標の分散}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_k y_k \Rightarrow S_{xy} : x, y\text{座標共分散}$$

また、$a = cos\theta$、$b = sin\theta$から、$a^2 + b^2 = 1$の制約もある。 この制約の中で、分散varの最大値を求めるために、ラグランジュの未定係数法を用いる。 この方法によれば、以下のように関数を作り、Gの最大値を与えるa、b、λを求めれば、Fの最大値を与えるa、bも求まることが分かっている。

$$F(a,b) = S_x a^2 + S_y b^2 + S_{xy} 2ab$$
$$C(a,b) = a^2 + b^2 - 1 = 0$$
$$G(a,b,\lambda) = F(a,b) - \lambda C(a,b)$$

これを解くには、Gをa、b、λで其々偏微分して、=0と置いた連立方程式を作る。

$$G(a,b,\lambda) = F(a,b) - \lambda C(a,b) = S_x a^2 + S_y b^2 + S_{xy} 2ab - \lambda(a^2 + b^2 - 1)$$
$$\frac{\partial G}{\partial a} = 2S_x a + 2S_{xy} b - 2\lambda a = 0$$
$$\frac{\partial G}{\partial b} = 2S_y b + 2S_{xy} a - 2\lambda b = 0$$
$$\frac{\partial G}{\partial \lambda} = -a^2 - b^2 + 1 = 0$$

上記の偏微分した式をまとめると

$$S_x a + S_{xy} b = \lambda a \quad \cdots (1)$$
$$S_y b + S_{xy} a = \lambda b \cdots (2)$$
$$a^2 + b^2 = 1$$

$$\begin{pmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

この式は共分散行列

$$\begin{pmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{pmatrix}$$

の固有方程式になっている。 以上から、λは共分散行列の固有値、(a、b)はその固有ベクトルになっているのでそれを求めればよい。 固有ベクトルは通常、a、bの比しか求められないが、ここでは$a^2 + b^2 = 1$の制約があるので、各固有値に対するa、bは一意に決まる。

分散varを最大化するa、b、λはラグランジュの未定係数法により、以下の(1)(2)のように求められた。(1)×a＋(2)×bと置くと、これも＝0となる。 これを整理して、$a^2 + b^2 = 1$の条件を使うと

$$S_x a + S_{xy} b - \lambda a = 0 \quad \cdots (1)$$
$$S_y b + S_{xy} a - \lambda b = 0 \quad \cdots (2)$$

$(1) \times a + (2) \times b$は

$$S_x a^2 + S_y b^2 + 2S_{xy} ab - \lambda(a^2 + b^2) = 0$$
$$\Leftrightarrow S_x a^2 + S_y b^2 + 2S_{xy} ab = \lambda$$

となり、左辺は最大化を目指した主成分得点の分散var（A）に他ならず、varの値は固有値λそのものであることを示してる。このように主成分分析は固有値問題に帰結することがわかる。 各固有値の大きさが、その軸の主成分得点の分散の大きさを表す。それが大きいほど、全体のデータの特徴を、一方向からよく眺められることになり主成分分析の目的に合致する。大きい固有値の固有ベクトルを第１主成分、二番めの大きさの固有値の固有ベクトルを第２主成分と呼ぶ。 対称行列の固有ベクトルは互いに直交する。共分散行列は対称行列なので、その固有ベクトルの第１主成分と第２主成分は直交する。

In [ ]: