統計数学と主成分分析による
ビッグ・データの可視化とパターン認識

Part 3: 原理解説編

リニア・テック　別府 伸耕

linear tec : Nobuyasu Beppu

セミナ全体の流れ

Sample

イントロダクション

↓

2020年アメリカ大統領選挙を
主成分分析で理解する

↓

Rによる主成分分析の実演

↓

主成分分析のアルゴリズム

↓

Pythonによる主成分分析の実演

↓

主成分分析に必要な数学の知識
(微分・積分，線形代数，統計学)

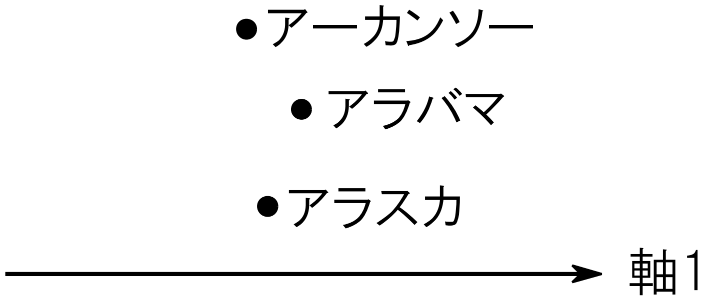
「Part 3 原理解説編」では
この部分を扱います.

本編の流れ

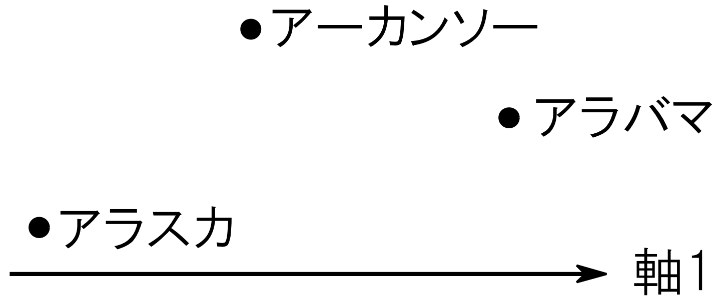
- 多変数関数の微分（偏微分と全微分）
- ラグランジュの未定乗数法
- 線形代数の基礎
- ベクトルの微分
- 主成分分析のアルゴリズムの導出
- Pythonによる実装例の紹介

データの分散

×悪い軸の取り方（分散が小さい）



〇良い軸の取り方（分散が大きい）



「分散が最も大きく見える状態」が理想的であると言える.

「射影した後のデータ」の分散を求める

ベクトル“$\boldsymbol{u}$”とデータ・ベクトル“$\boldsymbol{x}_i$”との内積をとり，スカラ“y_i”をつくる.

$$y_i = \boldsymbol{u} \cdot \boldsymbol{x}_i = \boldsymbol{u}^T \boldsymbol{x}_i$$

スカラ“y_i”（$i = 1, 2, \cdots, N$）の分散 “σ_y^2” を求める.

スカラ

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{u}^T\boldsymbol{x}_i - \boldsymbol{u}^T\bar{\boldsymbol{x}})^2$$

両方ともスカラ.
スカラは転置してもスカラ.

$$= \frac{1}{N}\sum_{i=1}^{N}\{\boldsymbol{u}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}\{\boldsymbol{u}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}^T$$

$$= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{u}$$

横ベクトル　縦ベクトル　横ベクトル　縦ベクトル