多变量解析

2020年7月15日

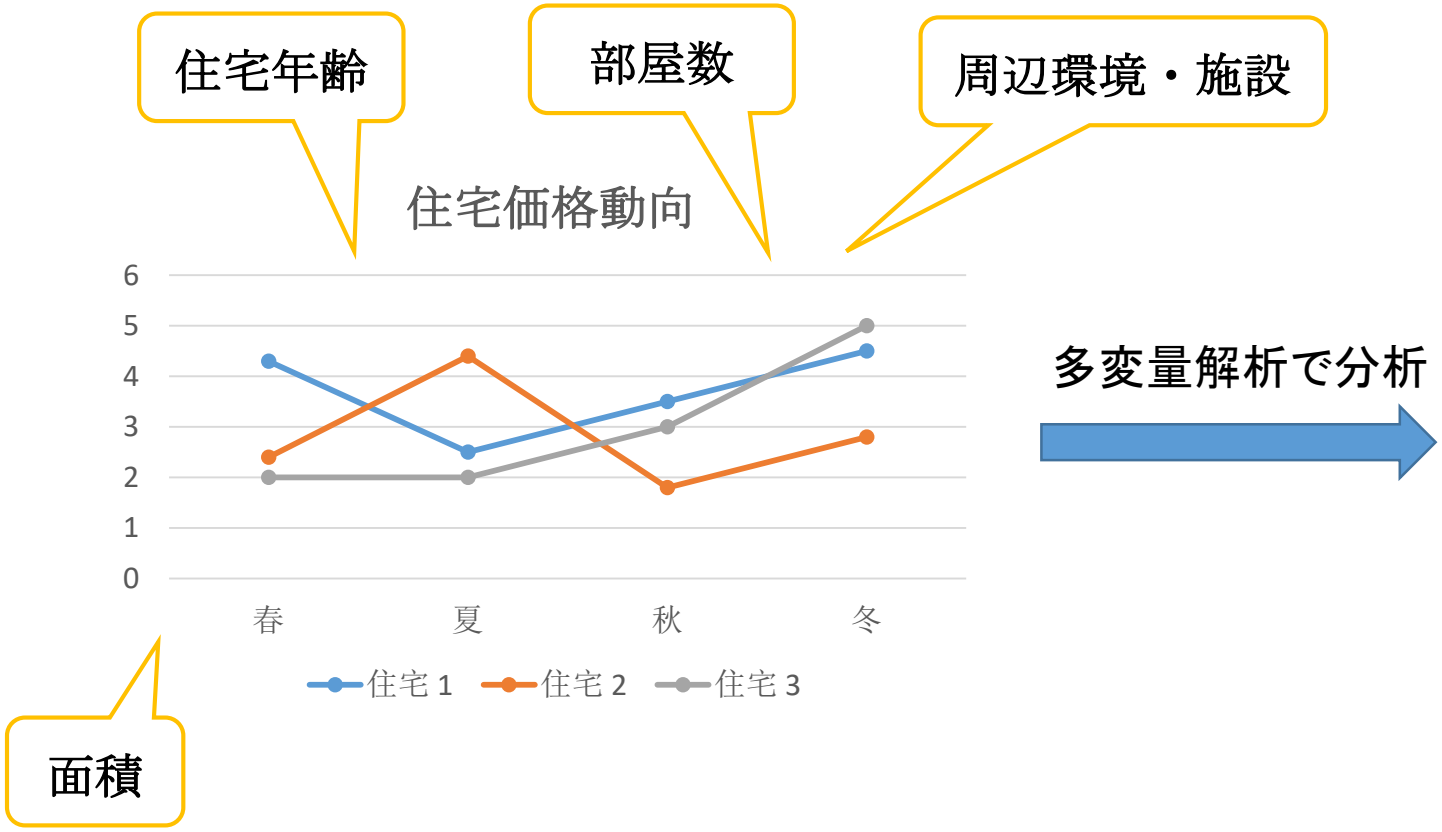
Liu Wenliang

目次

- 多変量解析の概要
  - 多変量解析の目的
  - 多変量解析の手順
- 重回帰分析
  - 重回帰分析の目標
  - クラシック方法 - 最小二乗法
  - 重回帰分析使用例
  - 評価方法
- 主成分分析
  - 主成分分析の使用目的と目標
  - 主成分分析式導出
  - 主成分分析の応用

多変量解析の目的

- 多変量解析とは、多くの情報（変数に関するデータ）を、分析者の仮説に基づいて関連性を明確にする統計的方法のことです。



多変量解析で分析

- 住宅価格を推定する
- 住宅価格の変動を予測する

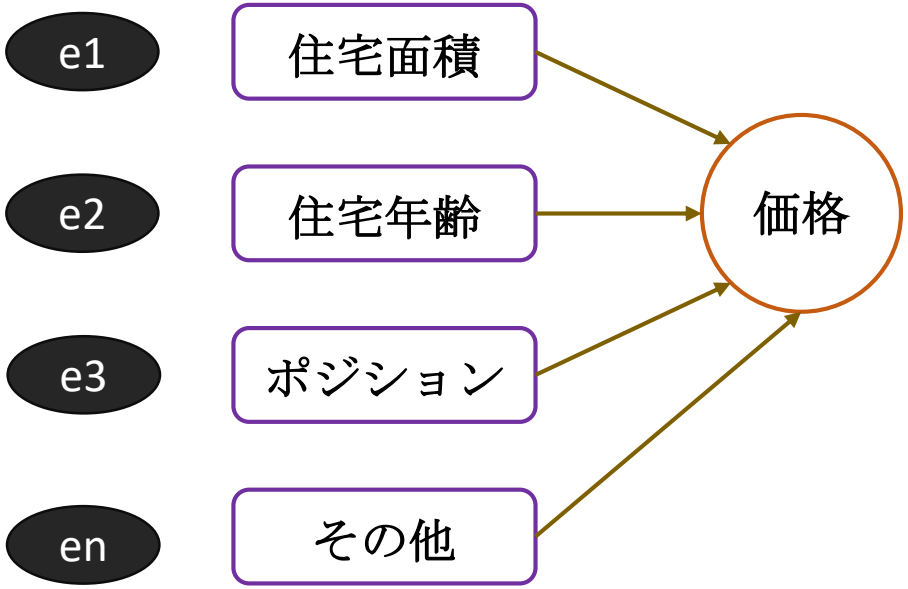
Datasets from UC Irvine Machine Learning Repository

多変量解析の目的

- 多変量解析の目的は、大きく分けて「**予測**」と「**要約**」の2つがあります。この2つの目的によって、手法が異なります。

予測

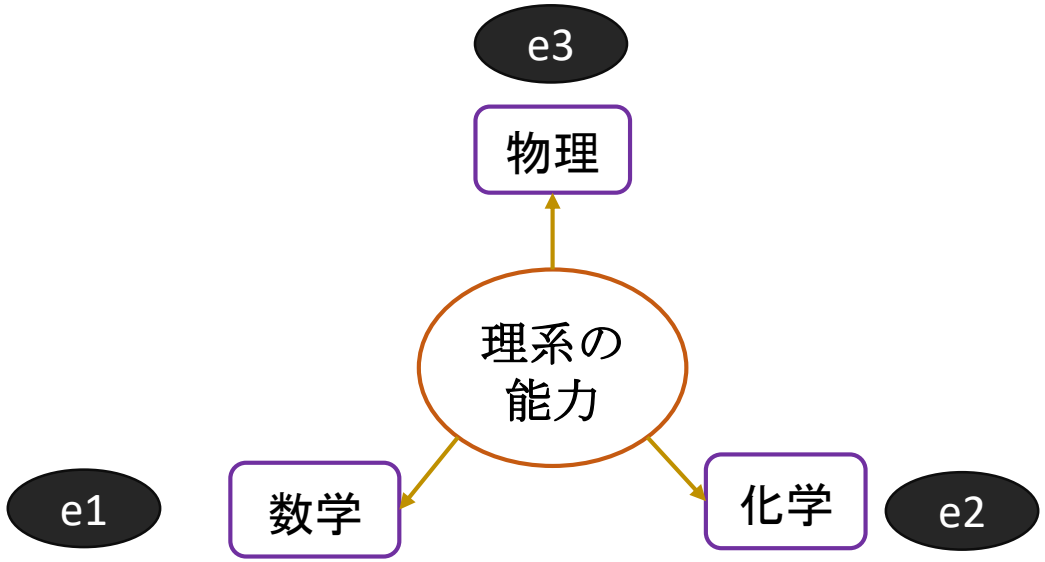
さまざまな倍率と複数の変数から結果を予測する



重回帰分析

要約

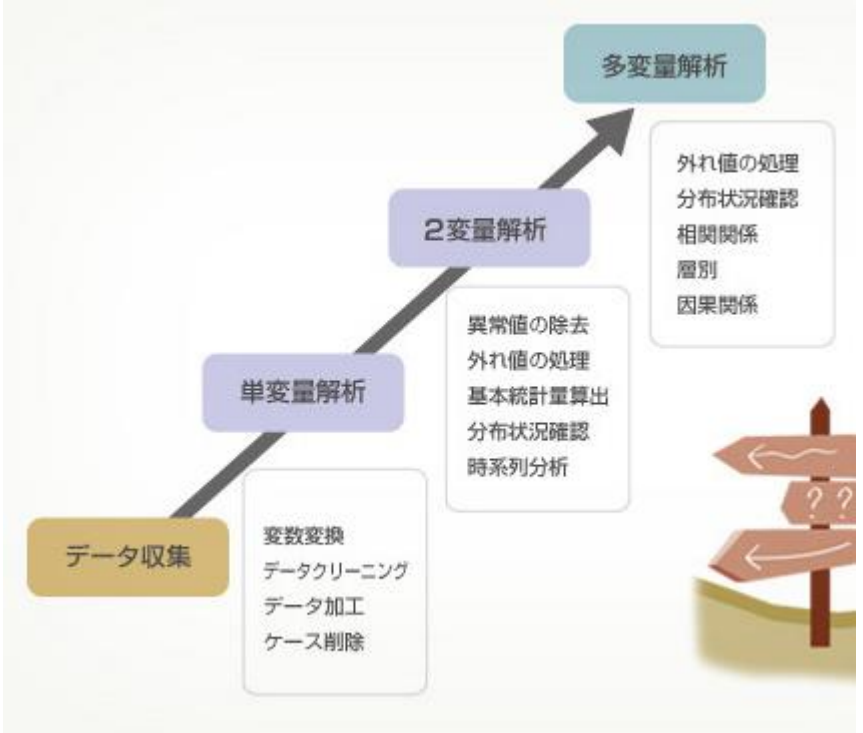
さまざまな倍率と複数の変数を新しい変数に要約する



主成分分析

enは重み係数

多変量解析の手順



多変量解析前の準備:

- データにクリーニングや加工などをして整える
- 可能であれば、データにマークを付ける
- トレーニングデータとテストデータに分割する
- データを正規化する

多変量解析後の評価基準:

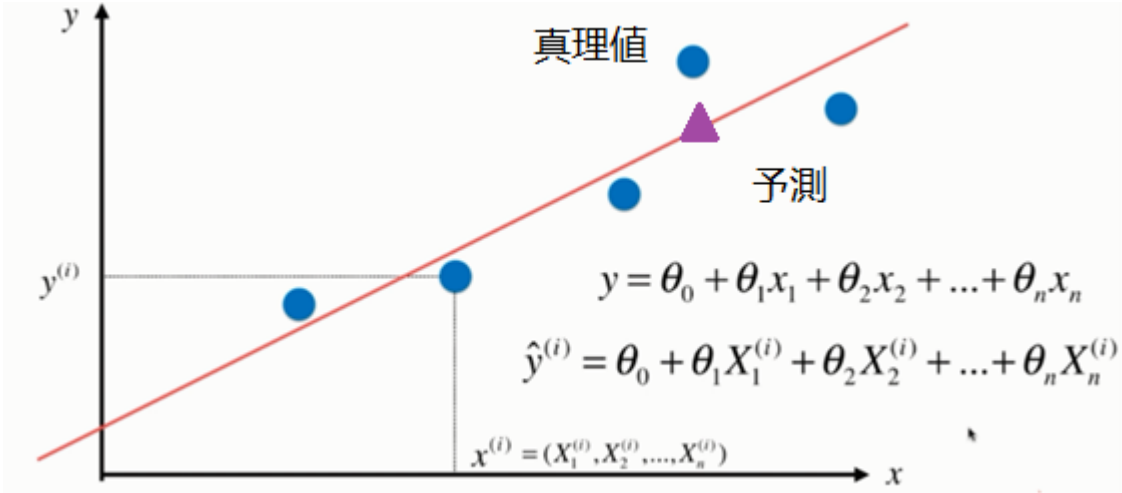
- エラー分析（MSE、RMSE、R Squared ）
- 時間の複雑さと時間の長さ
- 使用シナリオによって評価基準が異なる

重回帰分析の目標

サンプルフィーチャとサンプル出力ラベルの関係に最も適合する直線を見つける。

- 単回帰分析　　$Y = \mathrm{aX} + \mathrm{b}$
- 重回帰分析　　$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n$

目的変数→ Y　　説明変数→ $X_1, X_2, \ldots, X_n$
回帰定数→θ_0　　回帰係数→$\theta_1, \theta_2 \ldots \theta_n$
$X_n^{(i)}$→ i 番目の複数の変数
$\hat{y}^{(i)}$ → i 番目の予測値



最適なラインの効果

$$\sum (\boldsymbol{y}^{(i)} - \boldsymbol{a}\boldsymbol{x}^{(i)} - \boldsymbol{b})^2$$　　（損失関数、loss function）

損失関数を最小化するためにaとbの値を取得する

クラシック方法 - 最小二乗法

一次方程式の場合

aとbを解く:

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

重回帰分析

損失関数を最小化するために，$\theta_1, \theta_2 \ldots \theta_n$を解く:

$$X_b = \begin{bmatrix} 1 & X_1^{(1)} & \ldots & X_n^{(1)} \\ 1 & X_1^{(2)} & \ldots & X_n^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_1^{(m)} & \ldots & X_n^{(m)} \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \ldots \\ \theta_n \end{bmatrix} \quad => \hat{y} = X^b \cdot \theta$$

$$\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2 => (y - X_b \cdot \theta)^T (y - X_b \cdot \theta)$$ 最小化するために

$$\theta = (X_b^T X_b^T)^{-1} X_b^T y$$ これを正規方程式 (normal equation) と呼ぶ。

重回帰分析使用例

データセット：Boston house prices dataset（13個の特徴量，506個サンプル）

Pythonによって、トレニングデータを重回帰分析アルゴリズムによって得られた結果：

```
# 回帰係数マトリックス
reg.coef_
```

```
array([-1.20354261e-01,  3.64423279e-02, -3.61493155e-02,  5.12978140e-02,
       -1.15775825e+01,  3.42740062e+00, -2.32311760e-02, -1.19487594e+00,
        2.60101728e-01, -1.40219119e-02, -8.35430488e-01,  7.80472852e-03,
       -3.80923751e-01])
```

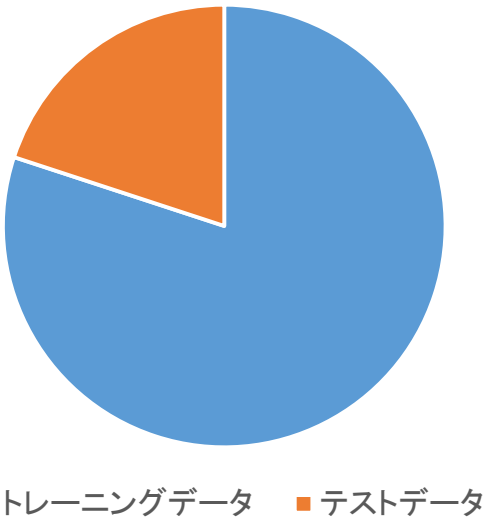
```
# 回帰定数
reg.intercept_
```

```
34.11739972320428
```

部分的な特徴量の説明

No. 5 NOX	一酸化窒素濃度
No. 6 RM	部屋の平均数
No. 7 AGE	住宅年齢
No. 10 TAX	固定資産税率
…	…

506個サンプルを分割される



コードリポジトリ：https://github.com/JamesLiuwenliang/Zemi_Respository

評価方法– R Squared

モデルをフィットした後、どのように評価しますか？

$$R^2 = 1 - \frac{\sum(\hat{y}^i - y^i)^2}{\sum(\bar{y} - y^i)^2}$$

計算されたモデルを使用して、発生した予測エラー

平均値を使用して、発生した予測エラー

結果分析：

- R^2<=1はいつでも成り立つ；
- R^2がもっと大きくなって、モデルは正しさは高くなる；
- もしR^2<0、結果のモデルは、ベンチマークモデルより良くない。データ自体を回帰分析に使用できない可能性がある。

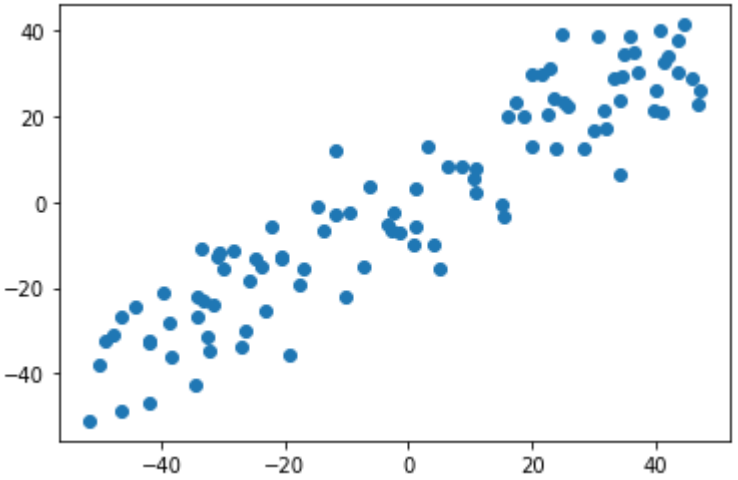
テストデータをモデルに代入しで、R Squaredを評価する

```
# R squard
%time reg.score(X_test, y_test)
```

```
Wall time: 1 ms
0.8129794056212832
```

主成分分析使用目的

- 主成分分析（Principal components analysis，PCA）使用目的は**次元圧縮**



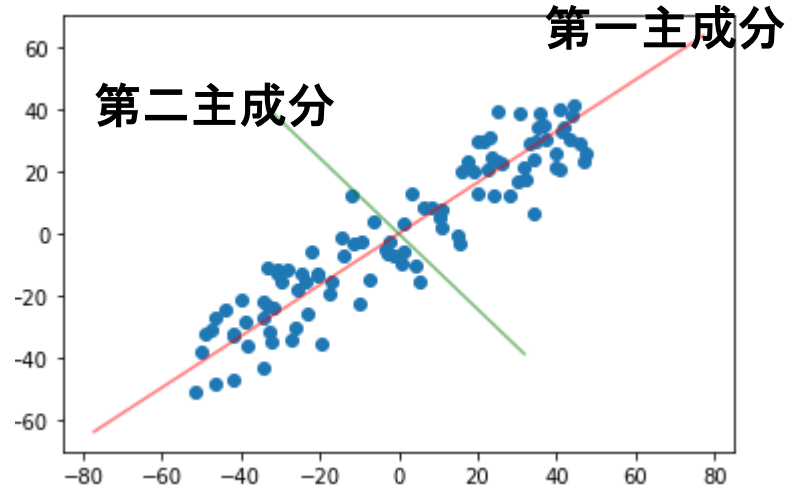
主成分分析使用目標

赤い線の物理的な定義：サンプルを直線上にマッピングし、マッピングされたサンプル間の距離が最大になるの軸

分散（Variance）：サンプル間の密度の程度を表す値

PCAの使用目標：分散最大化

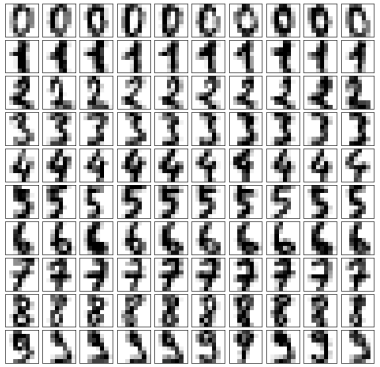
$$Var(x) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$$



主成分分析式導出

- リポジトリ：https://github.com/JamesLiuwenliang/Zemi_Respository/blob/master/PCA%E5%B0%8E%E5%87%BA%E3%83%97%E3%83%AD%E3%82%BB%E3%82%B9.md

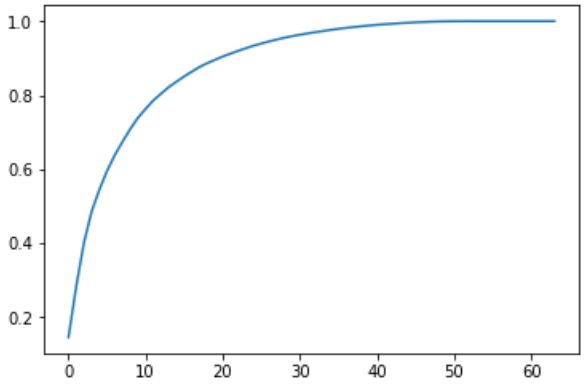
主成分分析の応用



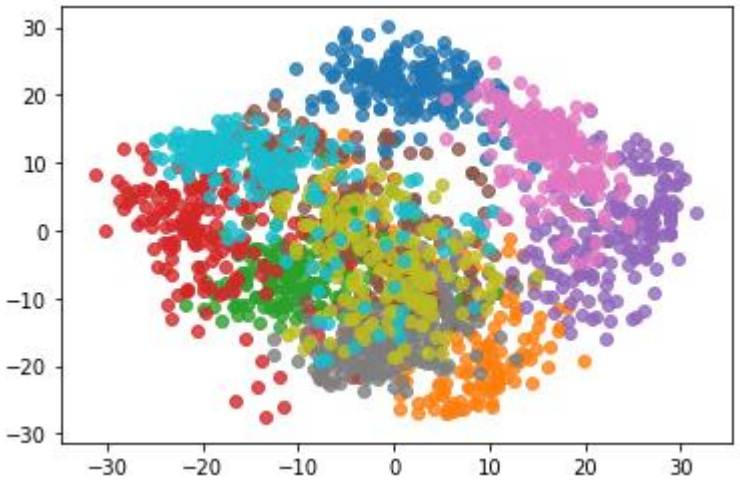
- データセット：Optical recognition of handwritten digits dataset（8*8個の特徴量，1797個サンプル、10個マーク）

(パターン認識アルゴリズムはk近傍法（kNN）)
Pythonによって、トレニングデータをPCAアルゴリズムによって得られた結果：

	テスト時間	テスト精度
kNNだけ	64.5 ms	0.978
PCAで2次元に圧縮	2.93 ms	0.731
PCAで28次元に圧縮	19.7 ms	0.978

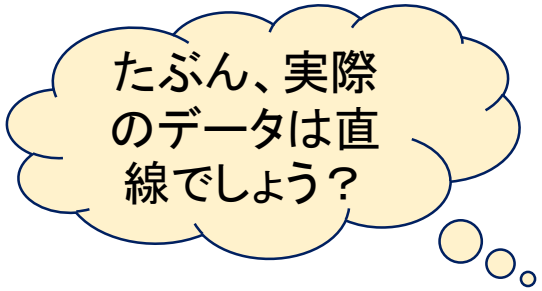


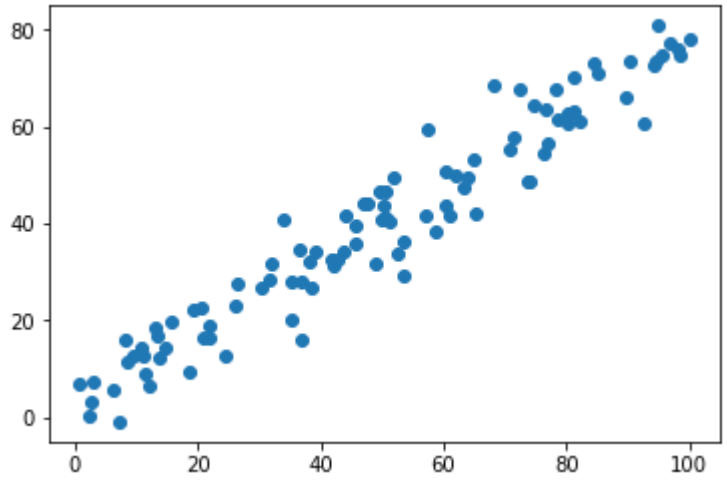
1797個サンプル
を分割される



2次元への次元縮小後の可視化

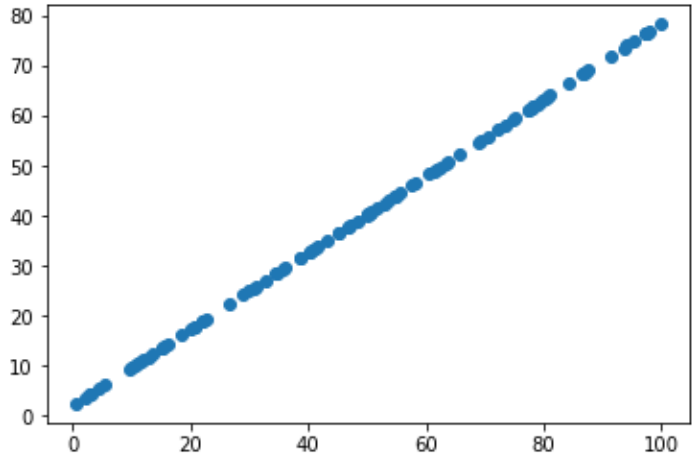
主成分分析の応用



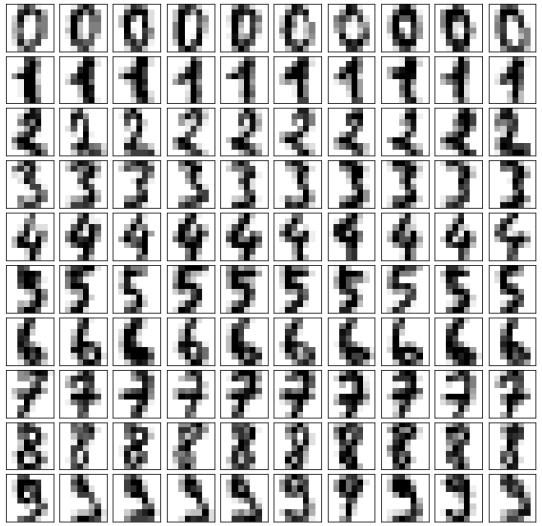


y = 0.75 * x + 3 + ノイズ

PCAアルゴリズムで、次元を削減しで、さらに次元を上昇する



復旧後、一部の情報が失われた



最初のデータ

ノイズを加える

PCAアルゴリズムで、次元を削減しで、さらに次元を上昇する

参考文献・出典

- 「図解雑学ー多変量解析」 丹慶勝市.ナツメ社.2005
- 「楽しく学べる多変量解析法」 藤沢栄作.現代数学社.1985
- 重回帰分析とは

https://www.albert2005.co.jp/knowledge/statistics_analysis/multivariate_analysis/multiple_regression

- 主成分分析の考え方

https://logics-of-blue.com/principal-components-analysis/

- 多変量解析2

https://www.macromill.com/service/data_analysis/d001.html