教師なし学習（次元削減）

教師なし学習

出力に関するデータ（正解）が与えられていない
クラスタリングか次元圧縮を目的とする

- クラスタリング
  - 入力データに基づいてサンプルをいくつかのクラスタに分類する
- 次元圧縮
  - 多種類の入力を少数種類の入力にまとめる

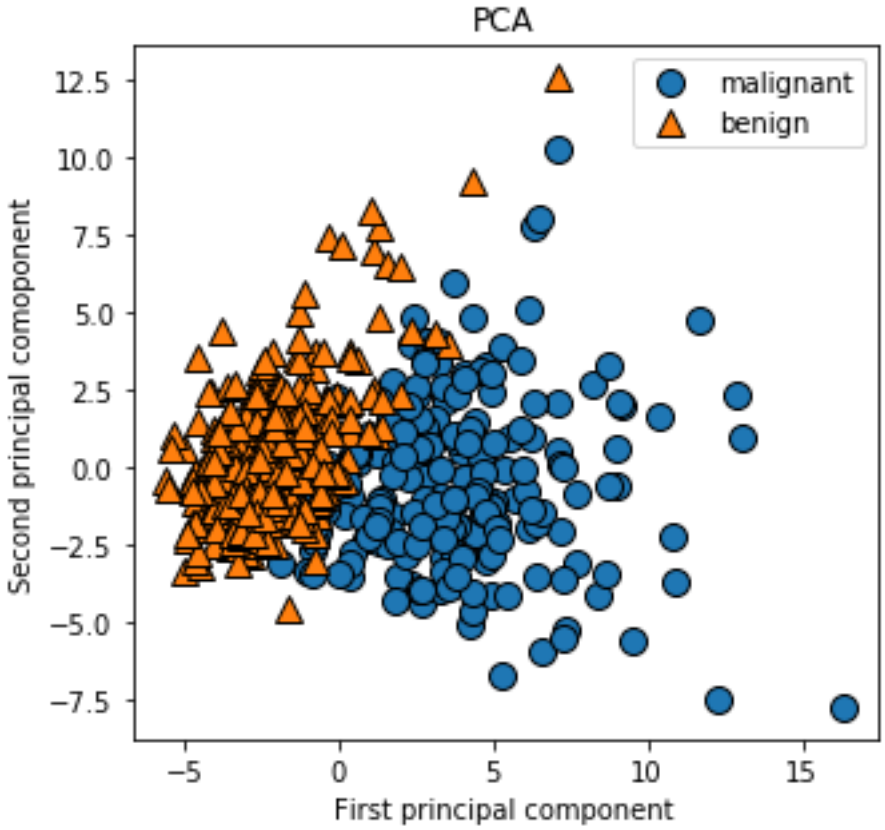
ex. 顧客の属性に基づいて顧客をいくつかのクラスタに分類する（クラスタリング）、学生の複数科目の試験成績データに基づいて学生の能力を説明できる少数の変数を作る（次元圧縮）

主成分分析（PCA）

教師なし学習（次元圧縮）

主成分分析の概要と事例

主成分分析：データの特徴を表現できる少数の合成変数を見出す手法



例：Breast cancer wisconsin (diagnostic) dataset

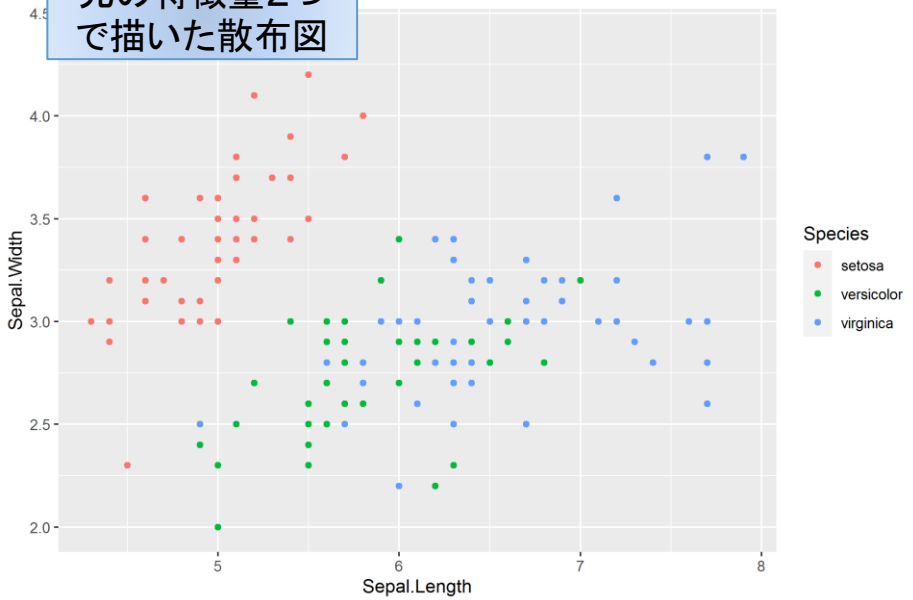
- 腫瘍細胞569例のそれぞれに関する直径や滑らかさ、対称性等30の特徴量の測定値からなるデータセット
- 良性（benign）357例、悪性（malignant）212例
- 30次元特徴空間は描画できない
- 主成分分析を行い30の特徴量を2つの主成分に縮約して散布図を描画

多次元特徴量を次元縮約して可視化が可能

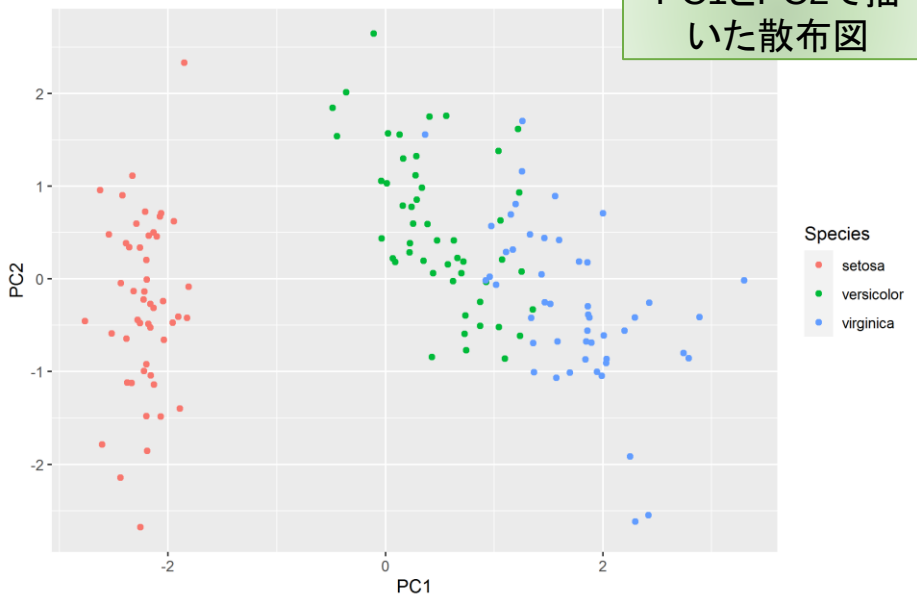
※データ出典：sklearn.datasets.load_breast_cancer()

主成分分析事例

元の特徴量2つで描いた散布図



PC1とPC2で描いた散布図

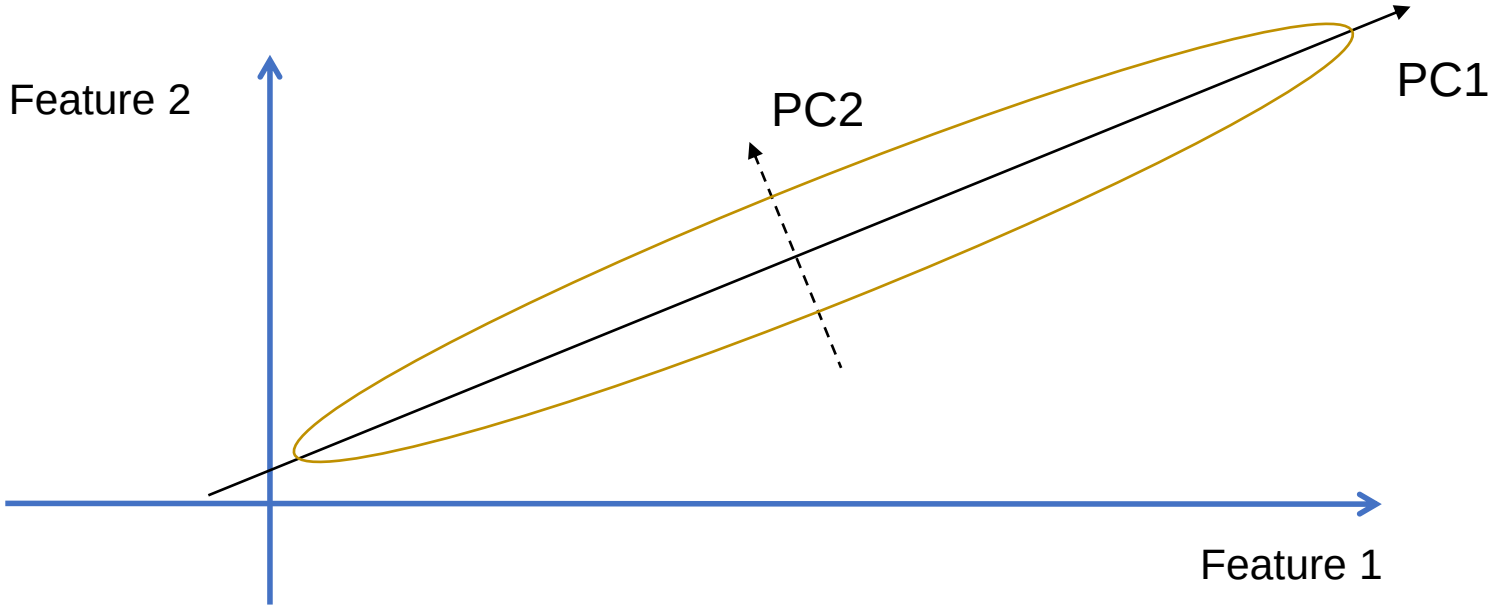


例：Iris Species Dataset

- 3種類のアヤメsetosa 50、versicolor 50、virginica 50の各個体に関するSepal.Width, Sepal.Length, Petal.Width, Petal.Lengthの4つの特徴量の測定値からなるデータセット
- 任意の2変数で散布図を描画するが（左）、どの変数も3種類のアヤメを分離できるような変数ではない
- 主成分分析を行い、第1主成分（PC1）と第2主成分（PC2）の2変数で散布図を描画（右）すると、PC1はそれだけでも3種類のアヤメを分離できるような変数になっている

識別しやすい特徴の抽出

主成分の見つけ方 -分散最大基準-



1. データの分散が最大となる方向を見つけ第1主成分とする
2. それに直交する方向で分散が最大となる方向を第2主成分とする
3. 次元数に至るまで繰り返す

- 特徴量が元々2つの場合、上図の黄色楕円のようにデータが分布している場合、長軸方向が第1主成分、それに直交する方向が第2主成分となり完了
- 特徴量がd個の場合、d次元特徴空間において同様のことを行う

主成分分析のアルゴリズムと実装

- データ行列に基づいて共分散行列Sを求める
- Sの固有値問題を解く
  - $S\boldsymbol{v} = \lambda\boldsymbol{v}$
  - λ: 固有値、$\boldsymbol{v}$: 固有ベクトル
  - 固有値λは固有ベクトル$\boldsymbol{v}$方向に射影したデータの分散となっている
- 対応する固有値の大きい固有ベクトルから順に第1主成分、第2主成分…となる

分散最大化問題は共分散行列の固有値問題に帰着するため、固有値問題を解けばよい

- 実装
  - Rのprcompパッケージなどで利用可能

共分散行列の固有値問題(1/2)

分散最大化問題は共分散行列の固有値問題に帰着する

- 簡単のため1次元への射影を考える。射影先の単位ベクトルを$\boldsymbol{w}_1$とすると$\boldsymbol{w}_1^T\boldsymbol{w}_1 = 1$である。データ$\{\boldsymbol{x}_i\}_{i=1}^n$の平均を$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$とする。
- データ点$\boldsymbol{x}_i$の$\boldsymbol{w}_1$方向への射影は内積$\boldsymbol{w}_1 \cdot \boldsymbol{x}_i = \boldsymbol{w}_1^T\boldsymbol{x}_i = \boldsymbol{x}_i^T\boldsymbol{w}_1$で表されることをふまえ、$\{\boldsymbol{w}_1^T\boldsymbol{x}_i\}_{i=1}^n$の平均を計算すると

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_1^T\boldsymbol{x}_i = \frac{1}{n}\boldsymbol{w}_1^T\sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{w}_1^T\bar{\boldsymbol{x}}$$

- したがって$\{\boldsymbol{w}_1^T\boldsymbol{x}_i\}_{i=1}^n$の分散は

$$\frac{1}{n}\sum_{i=1}^n \left|\boldsymbol{w}_1^T\boldsymbol{x}_i - \boldsymbol{w}_1^T\bar{\boldsymbol{x}}\right|^2 = \frac{1}{n}\sum_{i=1}^n \{\boldsymbol{w}_1^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{w}_1\}$$

$$= \boldsymbol{w}_1^T\left\{\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\right\}\boldsymbol{w}_1$$

共分散行列の固有値問題(2/2)

ここで、$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ は$\boldsymbol{x}_i$の共分散行列であり、Sとおく。Sは(d, d)行列で、解くべき問題は次の最適化問題になる

$\boldsymbol{w}_1^T\boldsymbol{w} = 1$の条件の下 $\boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1$を最大化する

ラグランジュ未定乗数法で解くため以下の関数ϕを考える

$$\phi(\boldsymbol{w}_1, \lambda_1) = \boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1 - \lambda_1(\boldsymbol{w}_1^T\boldsymbol{w}_1 - 1)$$

$\frac{\partial\phi}{\partial\boldsymbol{w}_1} = 2\mathrm{S}\boldsymbol{w}_1 - 2\lambda_1\boldsymbol{w}_1 = 0$ より

$$\mathrm{S}\boldsymbol{w}_1 = \lambda_1\boldsymbol{w}_1$$

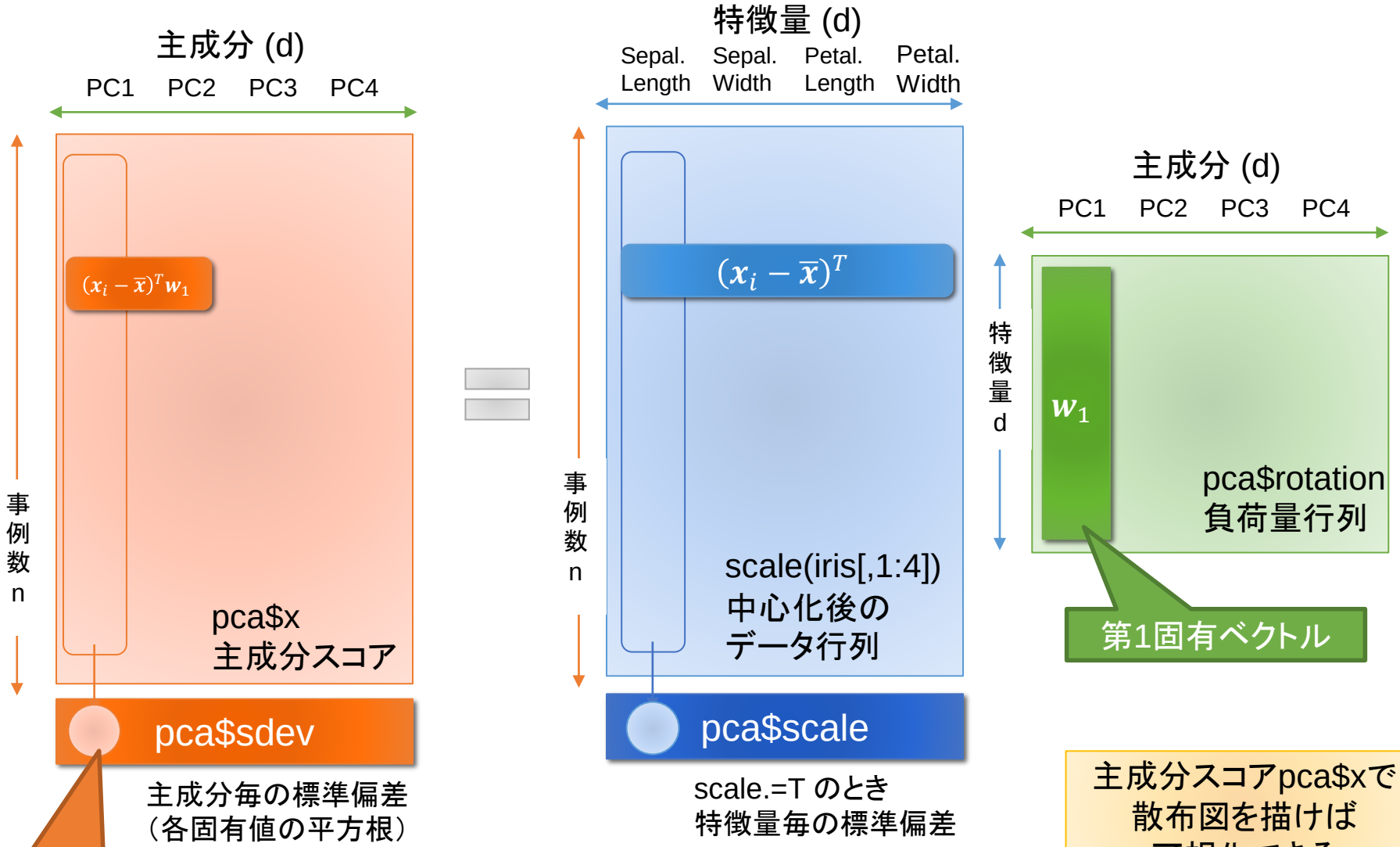
λ_1はSの固有値である。また、このとき

$$\boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1 = \boldsymbol{w}_1^T\lambda_1\boldsymbol{w}_1 = \lambda_1$$

であるため、分散は固有値に一致し、分散最大化問題は最大固有値を求める問題に帰着した

Rの主成分分析で得られるデータの構造

pca <- prcomp(iris[,1:4])
を例として



主成分スコアpca$xで
散布図を描けば
可視化できる

累積寄与率等

summary(pca)で以下の情報が得られる

- Standard deviation
  - 各主成分の標準偏差
- Proportion of Variance（寄与率）
  - 各主成分の固有値（分散）をその総和でわったもの
- Cumulative Proportion（累積寄与率）
  - 第1主成分から当該主成分までの寄与率の和

累積寄与率から、どの主成分まで採用すれば
全体の分散を十分に説明できるか判断する

教師なし学習（次元削減）

教師なし学習

出力に関するデータ（正解）が与えられていない
クラスタリングか次元圧縮を目的とする

- クラスタリング
  - 入力データに基づいてサンプルをいくつかのクラスタに分類する
- 次元圧縮
  - 多種類の入力を少数種類の入力にまとめる

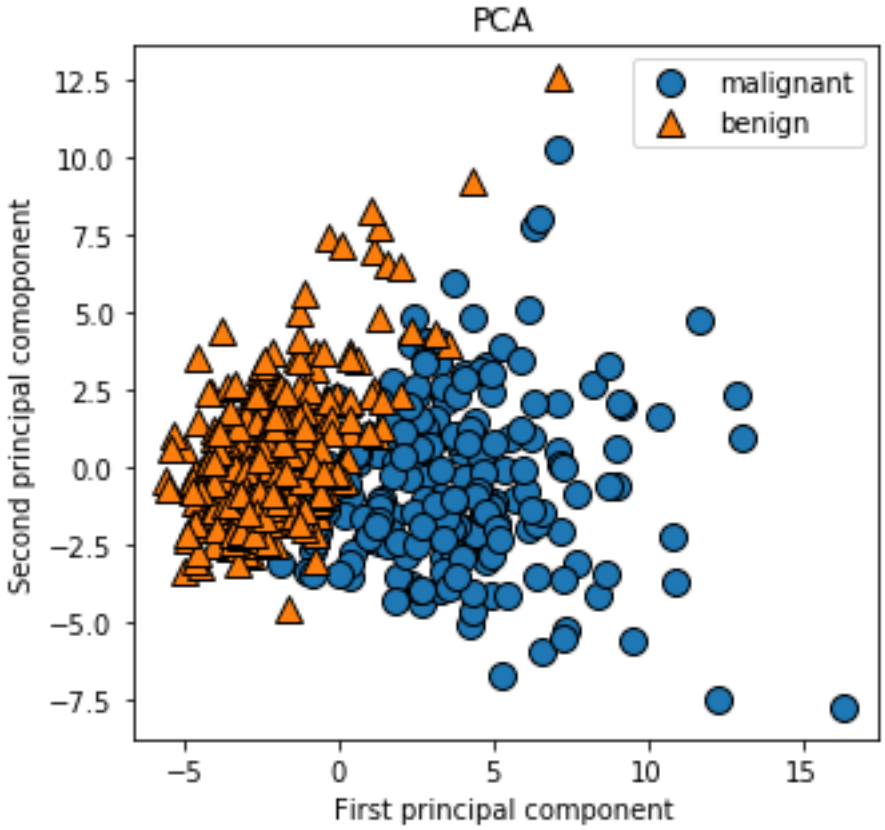
ex. 顧客の属性に基づいて顧客をいくつかのクラスタに分類する（クラスタリング）、学生の複数科目の試験成績データに基づいて学生の能力を説明できる少数の変数を作る（次元圧縮）

主成分分析（PCA）

教師なし学習（次元圧縮）

主成分分析の概要と事例

主成分分析：データの特徴を表現できる少数の合成変数を見出す手法



例：Breast cancer wisconsin (diagnostic) dataset

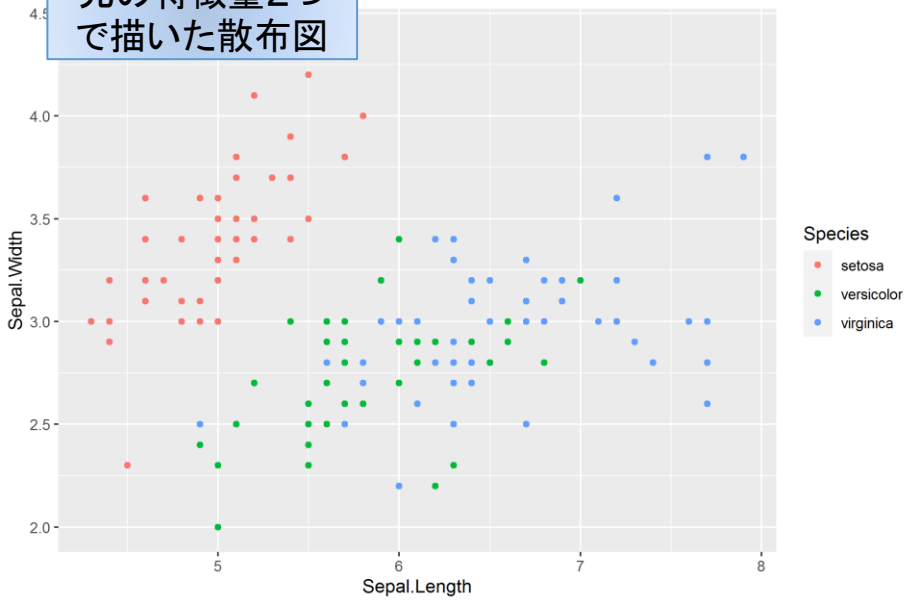
- 腫瘍細胞569例のそれぞれに関する直径や滑らかさ、対称性等30の特徴量の測定値からなるデータセット
- 良性（benign）357例、悪性（malignant）212例
- 30次元特徴空間は描画できない
- 主成分分析を行い30の特徴量を2つの主成分に縮約して散布図を描画

多次元特徴量を次元縮約して可視化が可能

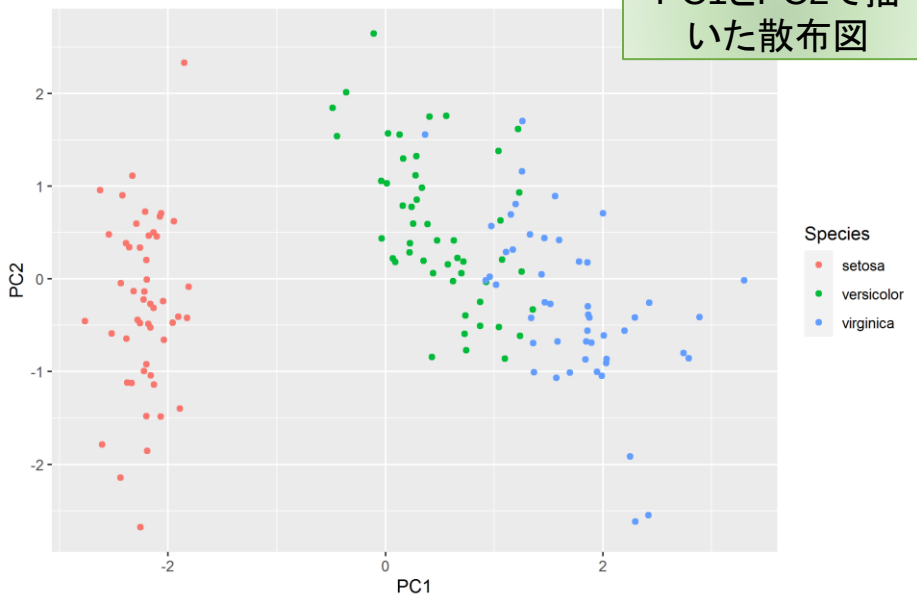
※データ出典：sklearn.datasets.load_breast_cancer()

主成分分析事例

元の特徴量2つで描いた散布図



PC1とPC2で描いた散布図

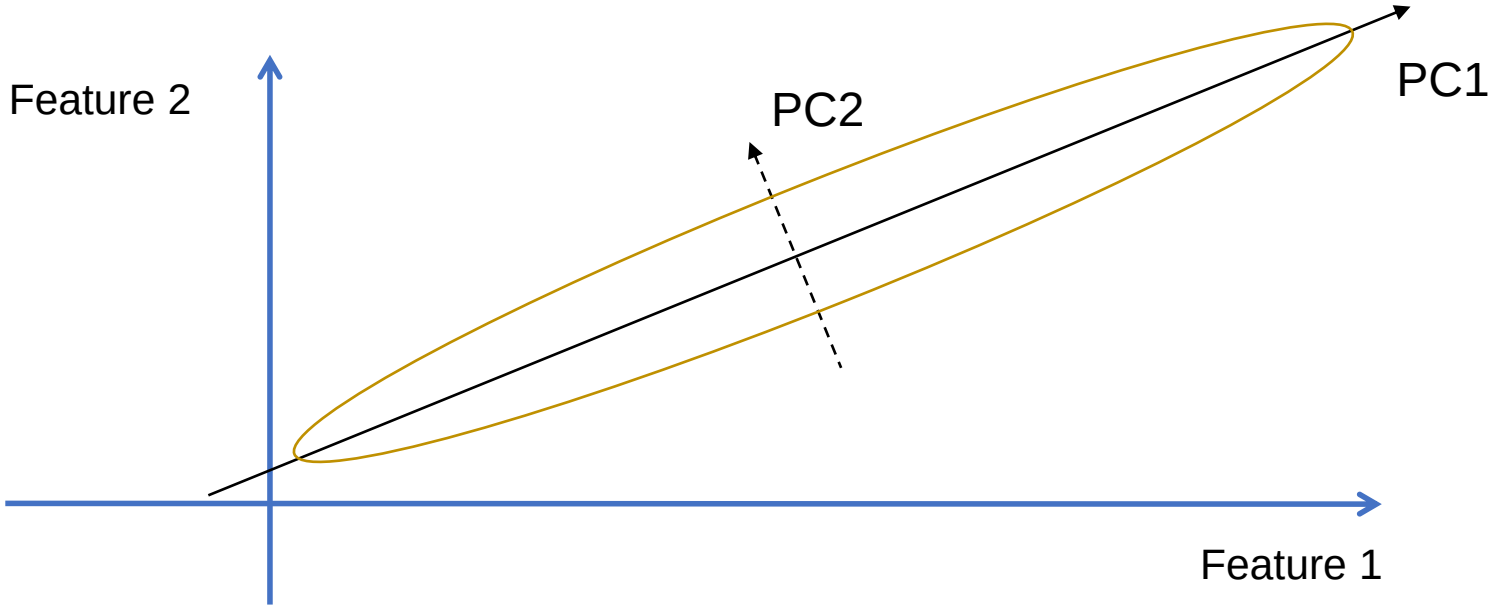


例：Iris Species Dataset

- 3種類のアヤメsetosa 50、versicolor 50、virginica 50の各個体に関するSepal.Width, Sepal.Length, Petal.Width, Petal.Lengthの4つの特徴量の測定値からなるデータセット
- 任意の2変数で散布図を描画するが（左）、どの変数も3種類のアヤメを分離できるような変数ではない
- 主成分分析を行い、第1主成分（PC1）と第2主成分（PC2）の2変数で散布図を描画（右）すると、PC1はそれだけでも3種類のアヤメを分離できるような変数になっている

識別しやすい特徴の抽出

主成分の見つけ方 -分散最大基準-



1. データの分散が最大となる方向を見つけ第1主成分とする
2. それに直交する方向で分散が最大となる方向を第2主成分とする
3. 次元数に至るまで繰り返す

- 特徴量が元々2つの場合、上図の黄色楕円のようにデータが分布している場合、長軸方向が第1主成分、それに直交する方向が第2主成分となり完了
- 特徴量がn個の場合、n次元特徴空間において同様のことを行う

主成分分析のアルゴリズムと実装

- データ行列に基づいて共分散行列Sを求める
- Sの固有値問題を解く
  - $S\boldsymbol{v} = \lambda\boldsymbol{v}$
  - λ: 固有値、$\boldsymbol{v}$: 固有ベクトル
  - 固有値λは固有ベクトル$\boldsymbol{v}$方向に射影したデータの分散となっている
- 対応する固有値の大きい固有ベクトルから順に第1主成分、第2主成分…となる

分散最大化問題は共分散行列の固有値問題に帰着するため、固有値問題を解けばよい

- 実装
  - Rのprcompパッケージなどで利用可能

共分散行列の固有値問題(1/2)

分散最大化問題は共分散行列の固有値問題に帰着する

- 簡単のため1次元への射影を考える。射影先の単位ベクトルを$\boldsymbol{w}_1$とすると$\boldsymbol{w}_1^T\boldsymbol{w}_1 = 1$である。データ$\{\boldsymbol{x}_i\}_{i=1}^n$の平均を$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$とする。
- データ点$\boldsymbol{x}_i$の$\boldsymbol{w}_1$方向への射影は内積$\boldsymbol{w}_1 \cdot \boldsymbol{x}_i = \boldsymbol{w}_1^T\boldsymbol{x}_i = \boldsymbol{x}_i^T\boldsymbol{w}_1$で表されることをふまえ、$\{\boldsymbol{w}_1^T\boldsymbol{x}_i\}_{i=1}^n$の平均を計算すると

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_1^T\boldsymbol{x}_i = \frac{1}{n}\boldsymbol{w}_1^T\sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{w}_1^T\bar{\boldsymbol{x}}$$

- したがって$\{\boldsymbol{w}_1^T\boldsymbol{x}_i\}_{i=1}^n$の分散は

$$\frac{1}{n}\sum_{i=1}^n \left|\boldsymbol{w}_1^T\boldsymbol{x}_i - \boldsymbol{w}_1^T\bar{\boldsymbol{x}}\right|^2 = \frac{1}{n}\sum_{i=1}^n \{\boldsymbol{w}_1^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{w}_1\}$$

$$= \boldsymbol{w}_1^T\left\{\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\right\}\boldsymbol{w}_1$$

共分散行列の固有値問題(2/2)

ここで、$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T$ は$\boldsymbol{x}_i$の共分散行列であり、Sとおく。Sは(d,d)行列で、解くべき問題は次の最適化問題になる

$$\boldsymbol{w}_1^T\boldsymbol{w}=1\text{の条件の下 }\boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1\text{を最大化する}$$

ラグランジュ未定乗数法で解くため以下の関数ϕを考える

$$\phi(\boldsymbol{w}_1,\lambda_1)=\frac{1}{2}\boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1-\lambda_1(\boldsymbol{w}_1^T\boldsymbol{w}_1-1)$$

$\frac{\partial\phi}{\partial\boldsymbol{w}_1}=\mathrm{S}\boldsymbol{w}_1-\lambda_1\boldsymbol{w}_1=0$ より

$$\mathrm{S}\boldsymbol{w}_1=\lambda_1\boldsymbol{w}_1$$

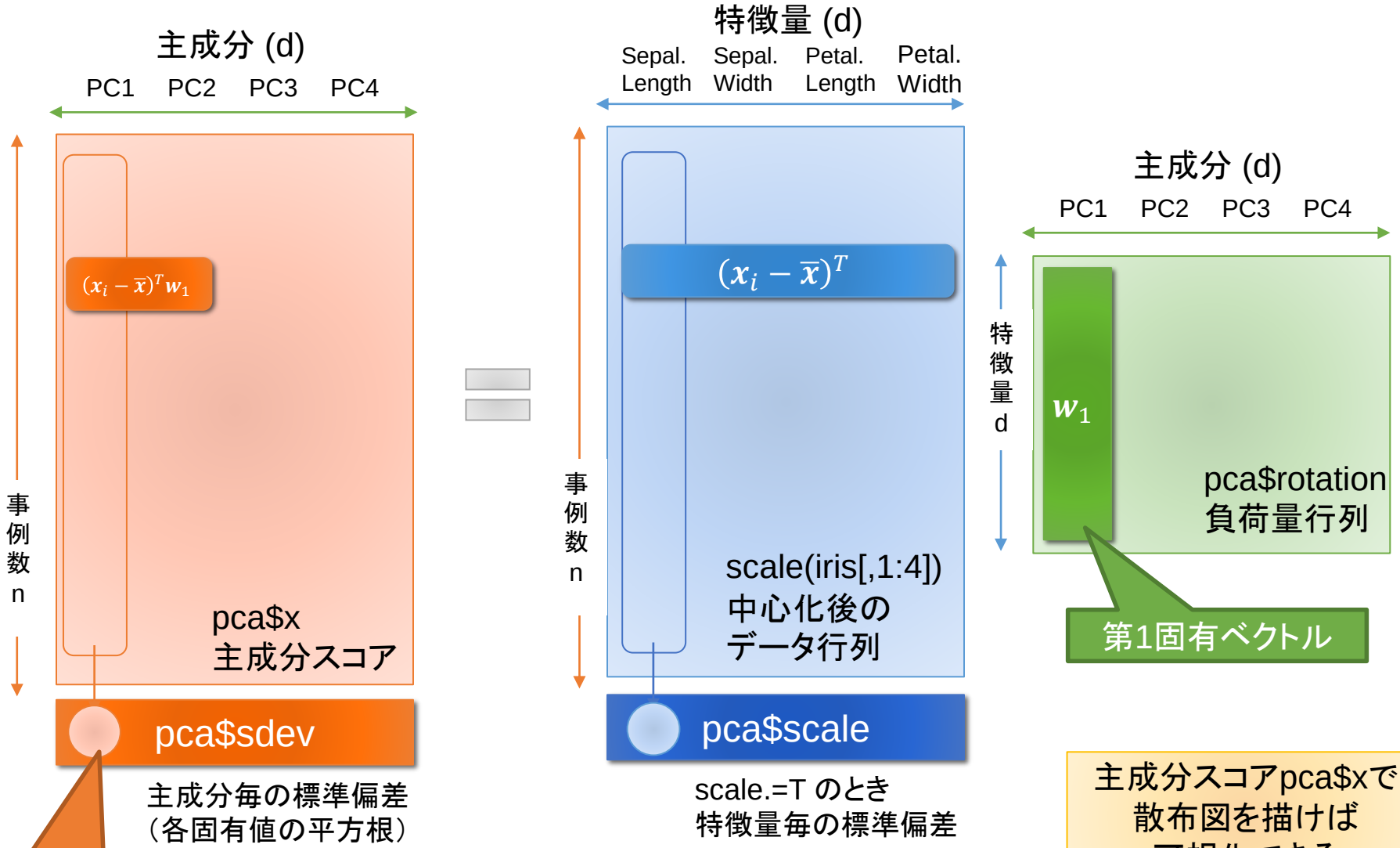
λ_1はSの固有値である。また、このとき

$$\boldsymbol{w}_1^T\mathrm{S}\boldsymbol{w}_1=\boldsymbol{w}_1^T\lambda_1\boldsymbol{w}_1=\lambda_1$$

であるため、分散は固有値に一致し、分散最大化問題は最大固有値を求める問題に帰着した

Rの主成分分析で得られるデータの構造

pca <- prcomp(iris[,1:4])
を例として



主成分スコアpca$xで
散布図を描けば
可視化できる

累積寄与率等

summary(pca)で以下の情報が得られる

- Standard deviation
  - 各主成分の標準偏差
- Proportion of Variance（寄与率）
  - 各主成分の固有値（分散）をその総和でわったもの
- Cumulative Proportion（累積寄与率）
  - 第1主成分から当該主成分までの寄与率の和

累積寄与率から、どの主成分まで採用すれば
全体の分散を十分に説明できるか判断する

教師なし学習
(クラスタリング)

クラスタリング（Clustering）

教師なし学習

データを類似性の高いグループに分ける

- ex. 顧客を収入や借入でグループに分ける，学生を成績や出席率によってグループに分ける等

クラスタリングのアルゴリズム

- k-平均法
- 自己組織化マップ

k-平均法の概要と事例

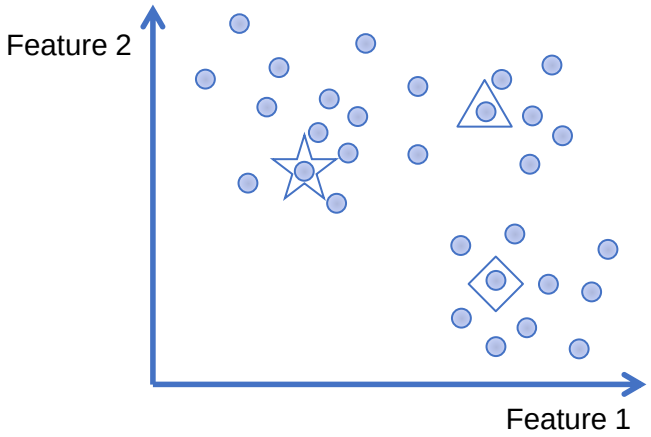
教師なし学習

特徴空間において距離が近いデータを同じクラスタとみなし、k個のクラスタを見出す手法

k-平均法のアルゴリズム(1/2)

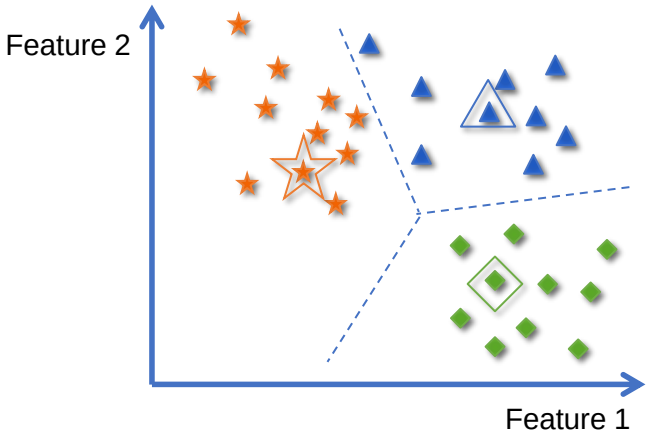
出典：Brett Lantz:「Rによる機械学習」翔泳社（2017）を基に鈴木作成

1



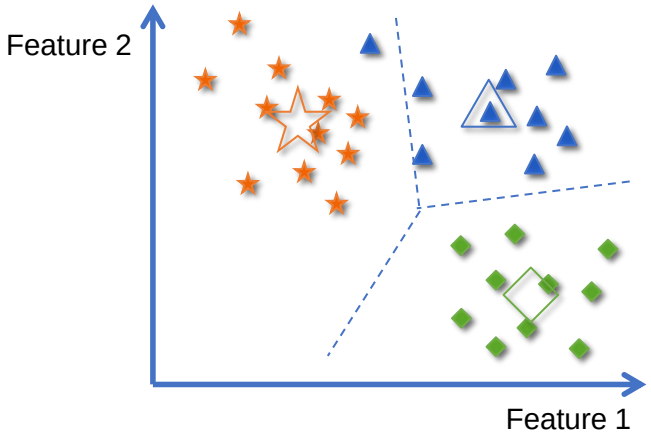
1. k個の点を訓練データセットから無作為に選ぶ（初期クラスタ中心）

2



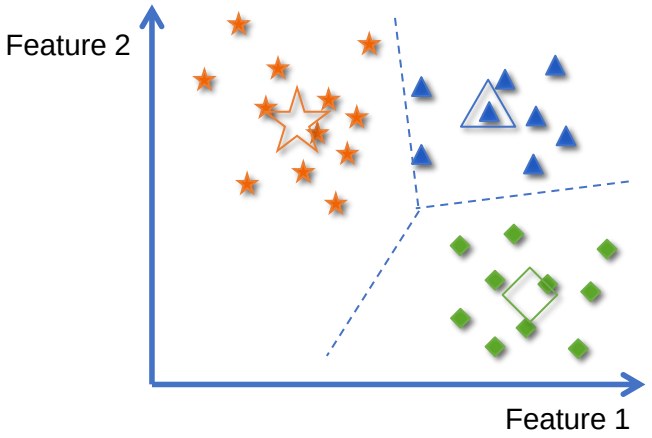
2. データ点毎に、クラスタ中心との距離を求め最も近いクラスタに割り振る

3



3. クラスタ毎に、その時点で含まれる点から重心を求め、新たなクラスタ中心とする

4

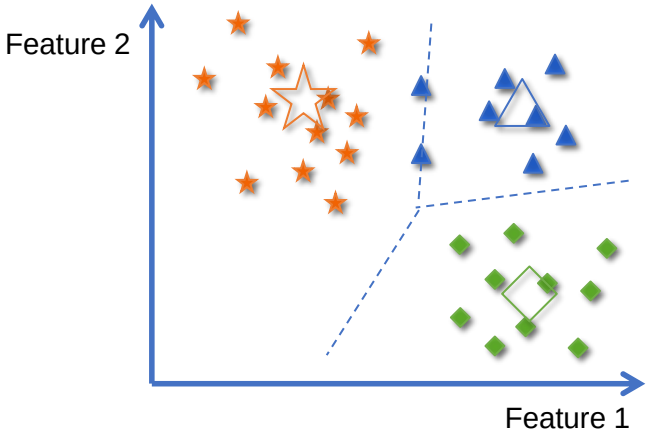


4. データ点毎に新たなクラスタ中心との距離を求め改めて割り振る

k-平均法のアルゴリズム(2/2)

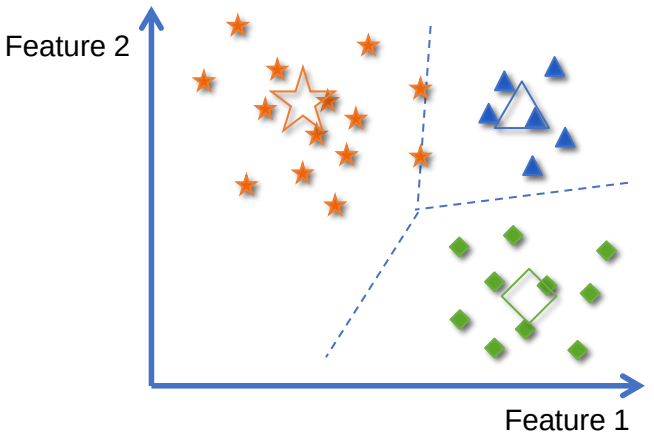
出典：Brett Lantz:「Rによる機械学習」翔泳社（2017）を基に鈴木作成

5



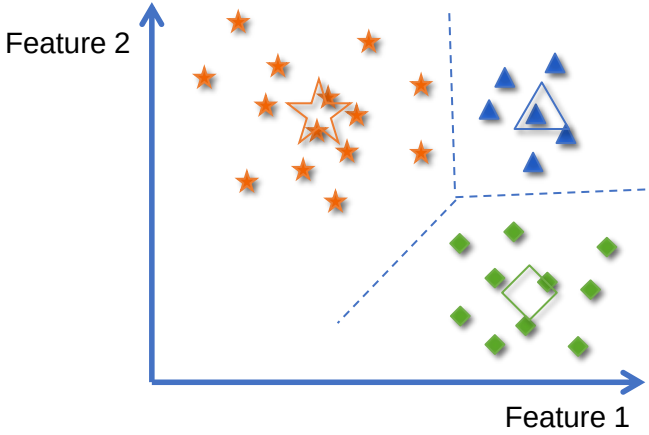
5. クラスタ毎に、その時点で含まれる点から重心を求め、新たなクラスタ中心とする

6



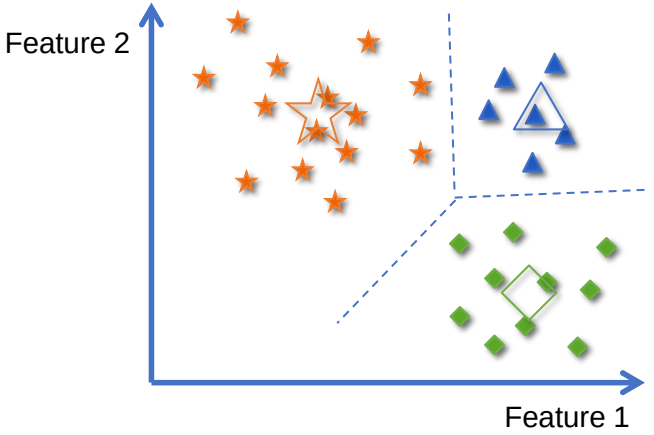
6. データ点毎に、クラスタ中心との距離を求め最も近いクラスタに割り振る

7



7. クラスタ毎に、その時点で含まれる点から重心を求め、新たなクラスタ中心とする

8



8.新たな割り振りが生じないため終了

k-平均法のアルゴリズム

1. クラスタ中心の初期設定
   - クラスタの中心となるk個の点を訓練データセットから無作為に選んで決める

2. データ点を最も近いクラスタへ割り振り
   - 各データ点について、クラスタ中心との距離を計算し、最も近いクラスタ中心を持つクラスタに割り振る

3. クラスタ中心の更新
   - クラスタ毎に、その時点で当該クラスタに含まれるすべての点から重心を求め、新たなクラスタ中心とする

4. Step.2とStep.3の繰り返し
   - クラスタ中心の更新にともない、所属クラスタが変わるデータ点が現れうる。Step.2 で各データ点について、新たなクラスタ中心との距離を計算し、最も近いクラスタへ改めて割り振る。割り振りに変化がない場合終了する

k-平均法適用上の注意(1/2)

初期クラスタ中心を無作為に選ぶため、結果が異なることがある

- 特徴空間全体から無作為な点を選ぶ方法、初期中心を選ばずにいきなり各インスタンスにクラスタを割りあてる方法などもあるが、各方法に依存したバイアスがある

距離関数を用いるため、事前にデータの正規化や標準化が必要である

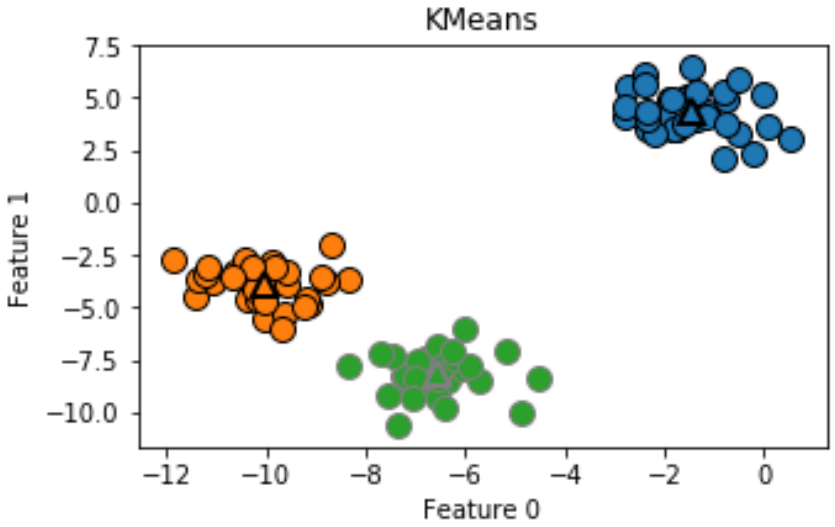
- ユークリッド距離を用いることが多いが、マンハッタン距離やミンコフスキー距離を用いることもある

クラスタ数kは分析者が決める必要がある

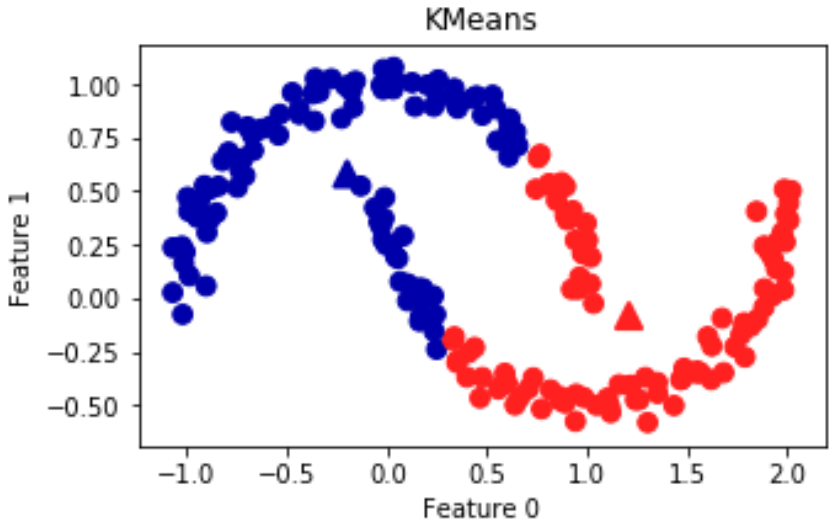
- いくつに分けるのが自然か、いくつに分けたいかで決める
- データに関する事前知識が無く目安が欲しい場合は$k = \sqrt{\frac{n}{2}}$とする方法もある
- kを増やしながらクラスタリングを試行し、グループ内の同質性の増加が鈍くなるkを採用する方法（elbow法）もあるが計算量過多

k-平均法適用上の注意(2/2)

適用可能なデータ
（丸い形状）



適用失敗するデータ
（複雑な形状）



- 各データ点は各クラスタ中心との距離に基づき、最も近いクラスタに割り当てられるため、複雑な形状のデータではクラスタリングに失敗する

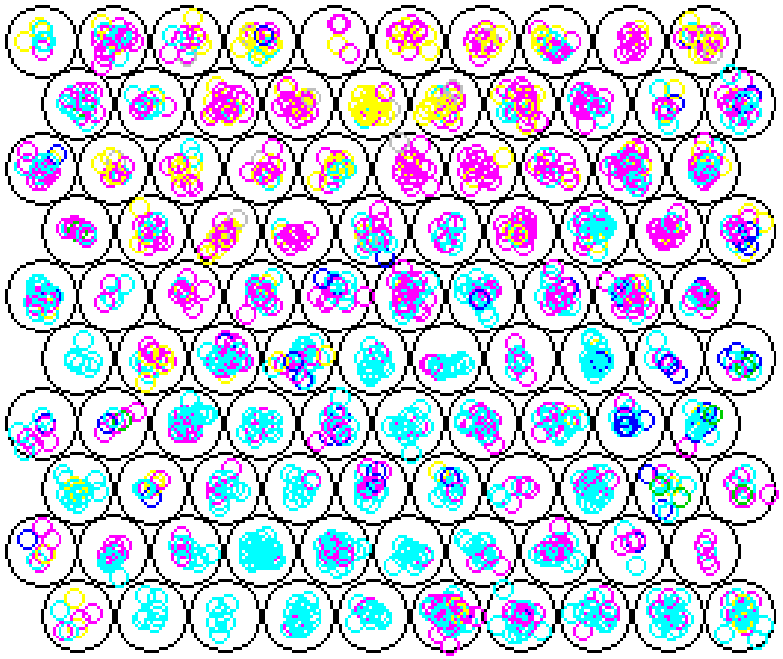
※データ出典：A. C. Muller et al. mglearn.datasets

自己組織化マップ（SOM）

教師なし学習（次元削減）

自己組織化マップ（SOM）の概要と例

高次元の特徴空間にあるインスタンスを2次元平面上に表現する手法

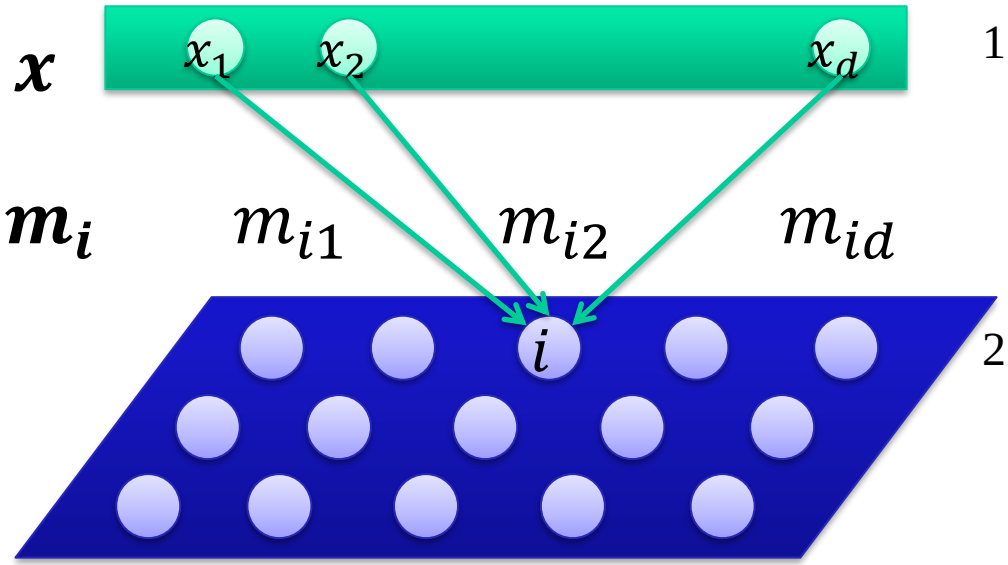


https://archive.ics.uci.edu/ml/datasets/Wine+Quality
赤ワインデータを正規化したものに対してKohonenを利用

- タスク事例
  - 11種類の科学的特性データであらわされる赤ワイン1,599品種を二次元平面上に可視化したい
  - 品質スコア（0から10の11段階）は学習に用いない（教師なし学習）
- 手法：SOM
  - 11次元特徴空間の1点となるひとつの品種は、2次元に並んだ10×10の出力ユニットのいずれかに対応付けられる
  - 黒丸は出力ユニット、中の小さい点ひとつがひとつの品種を表している
  - 科学的特性が互いに似ている品種は同じ出力ユニットか近くの出力ユニットに描画されることになるため、品種間の類似性の検討等に活用できる
- 結果の妥当性確認
  - 各品種の品質スコアを色で表現したところ、同じ色が同一か近傍のユニットに配置されている

自己組織化マップのアルゴリズム

入力層（d個のユニットからなる）



出力層（ 多数のユニットが二次元状に配置される）

入力層と出力層の間は全結合

1. 入力ベクトル$\boldsymbol{x}$と最も近い重みベクトル$\boldsymbol{m}_i$を持つ出力ユニットcを勝者とする

$$c = \arg\min_i |\boldsymbol{x} - \boldsymbol{m}_i|$$

2. 勝者ユニットやその近傍ユニットの重みベクトルを入力ベクトルに近付ける

$$\boldsymbol{\Delta m_i} = \boldsymbol{m}_i + \alpha\, h_{ci}(\boldsymbol{x} - \boldsymbol{m}_i)$$

$$h_{ci} = \exp\left(-\frac{|\boldsymbol{r}_c - \boldsymbol{r}_i|^2}{2\sigma^2}\right)$$

αは学習定数、h_{ci}は近傍関数（近傍ほど大きく、離れるほど小さい）

3. 入力ベクトルを更新して繰り返す

Rのsomパッケージ、kohonenパッケージなどで実装されている