8.3

1

一番目のデータはどのクラスタも同じ程度の分散を持っている一方で、二番目のデータはクラスタごとに分散の程度が異なっているという違いがある。二番目のクラスタでは、大きい1つのクラスタの左右に小さいクラスタが2つあるように色分けされることが期待されるが、うまくいっていないことがわかる。通常k-means法ではデータ同士の近さをユークリッド距離で評価する。そのため、クラスタごとの距離尺度を考慮していないため、二番目のクラスタリングがうまくいっていないと考えられる。距離尺度を考慮すれば、大きいクラスタほどそのクラスタに含まれるデータの距離が大きい。

全て一律のL2ノルムが 近さを測る尺度として用いられている
中央のクラスタの一部が右側のクラスタの一部として クラスタリング されている。

2

3番目のデータはどのクラスタリングも同程度の分散(共分散)をもっているが、

1番目のデータはクラスタ内において全ての方向に対して一様にデータが広がっている一方で、3番目のデータは方向によって広がり方が異なるという違いがある。これはユークリッド距離を用いてデータ同士の近さを評価しており、 3番目のデータにおいて、方向ごとの分散の違いが考慮していないためであると考えらえる。

全て一律のL2ノルムが 近さを測る尺度として利用されている

8.7

1. 第一主成分は、データの分散が最大になる方向、第二主成分は、第一主成分に直行する空間の中でそのデータの分散が最大になる方向、というように定義される。これら両者の組み合わせは、データを二次元空間に射影した際に、その広がりが最も大きくなるような方向の組み合わせであると言える。三種のアヤメはそれぞれ種類ごとに異なる形状を持つ場合、これらの組み合わせによる空間において、それぞれのアヤメはよく分離して分布すると考えられる。この時、第一主成分の寄与率が1に近い値をとることから、第一主成分の情報さえあれば三種のアヤメの分類するための十分な情報が得られると考えられる。
一方で、第三主成分、第四主成分の組み合わせは、その寄与率から分かるとおり、サンプルの情報をほとんど保持していないことがわかる。そのため、この二次元空間にアヤメのデータを射影しても、三種のアヤメを分類するための十分な情報を得ることができていないと考えられる。

データに含まれる特徴量が表示

2. メリット：
第一主成分のみの値を使えばより少ない特徴数で分類器を構築できるため、学習時間・予測時間を小さくすることができると考えらえる

主成分分析を用いることでデータの容量を減らすことができ結果としてより少ないメモリ量での学習が可能

デメリット：
第一主成分の値のみを使った場合、主成分の抽出において失われる情報があるため、すべての特徴を使った場合に比べて分類精度が低くなってしまう可能性があると考えられる。

学習前に主成分分析を計算する必要がある