Python で始める機械学習

3.5 クラスタリング

15T4032G 芝山直希

2018 年 5 月 25 日

クラスタリング？

clustering。データセットを複数のグループ (クラスタ) に分割するタスク。

- 各クラスタに似たデータが集まり、異なるクラスタのデータとは類似しないように分けるのが目的
- どのクラスタに属するかを表すラベルが各データにつけられる
- 各クラスタの特徴は解析するまでわからない

アルゴリズム：

1. k-means クラスタリング
2. 凝集型クラスタリング
3. DBSCAN

k-means クラスタリング

各クラスタの重心を求めることで分類するアルゴリズム
学習法：

1. 各データを最寄りのクラスタ重心 (クラスタセンタ) に割り当てる
2. 各クラスタ重心を、その重心に割り当てられた全データの平均値に更新する
3. クラスタ重心が動かなくなるまで 1、2 を繰り返す

長所：

- 理解しやすく実装しやすい
- 実行が高速
- 大規模データセットにも適用可能
- 成分分解手法として利用できる (ベクトル量子化)

短所：

- クラスタ重心の初期位置が乱数依存
- 円形でないクラスタを識別できない
- 探すクラスタ数をユーザーが指定する

凝集型クラスタリング

各データポイント一つをクラスタとみなして開始し、類似度の高い 2 クラスタを結合していくアルゴリズム

- scikit-learn では指定したクラスタ数になるまで結合を繰り返していく
- 階層型の手法であり、SciPy を使えばデンドログラムという形式で分割過程を可視化できる
- 連結度が類似度の尺度となり、小さいほど類似度が高いと判断される

連結度の種類

- ward…結合時の分散の増加量で判断する 多くの場合、比較的同サイズのクラスタに分けられる
- average…クラスタ間の全データポイント間の距離の平均値で判断
- complete…クラスタ間の点間距離の最大値で判断

ほとんどの場合 ward(デフォルト) でうまくいく クラスタのデータポイント数に極端な差が出た場合他の連結度の方がうまくできるかも

長所：

- デンドログラムから階層的な分割の候補を得られる
- k-means クラスタリングでは不可能な円形でない単純な形状を扱える

短所：

- 複雑な形状のクラスタに対してうまくいかない
- scikit-learn から利用する場合、クラスタ数を指定する必要がある
- 新規テストデータに対する予測ができない

scikit-learn はデンドログラムの描画をサポートしていない

DBSCAN

density-based spatial clustering of applications with noise。「クラスタは特徴空間内の比較的低密度な領域で区切られた高密度領域」という考えに基づくアルゴリズム。

1. 適当な分類されていないデータポイントを１つ選択し、距離 eps 以内に min_samples 以上のデータがない場合、ノイズとする
2. 距離 eps 以内に min_samples 以上データがあった場合、クラスタのコアサンプルとなる
3. あるクラスタに属するコアサンプルから距離 eps 以内にあるデータポイントはそのクラスタに属する
4. 2、3 を繰り返していき、距離 eps 以内にコアサンプルがなくなった場合、1 から繰り返す
5. すべての点が調べられたら学習終了

主要パラメータ：eps(重要パラメータ、近さの意味・クラスタ数に影響)、min_sample(クラスタの最小サイズに影響)
長所：

- クラスタ数をユーザーが指定する必要がない
- 複雑な形状のクラスタを認識できる
- ノイズを検出できる

短所：

- 遅いアルゴリズムである
- 新規データに対する予測ができない

クラスタリングアルゴリズムの評価

正解データを使う評価指標：

- ARI(調整ランド指数)、NMI(正規化相互情報量) など
- 理想に近いクラスタリングであるほど 1 に近づくため、直感的にわかりやすい
- クラスタリング用のデータには正解データがない場合が多い

正解データを使わない評価指標：

- シルエット係数など
- あまりうまくいかない
  - シルエット係数ではコンパクトさを評価するが、複雑な形状のクラスタはコンパクトにならない

頑強性を用いた評価指標もあるが、人間にとって興味深いクラスタリングであるかどうかは、クラスタの中身を確認するしかない

顔画像データセットを用いた比較

Labeled Faces in the Wild データセットを元に PCA(whiten=True) で 100 成分の固有顔表現を生成し、これをクラスタリングさせる

- DBSCAN
  - eps を変化させると結果は変化したが、大きなクラスタを複数得ることはなかった
  - ノイズを検出できるため、外れ値検出に利用できる

Figure: ノイズとして検出された画像 (eps=15)

- k-means
  - クラスタ数を 10 にしてクラスタリングしたところ、顔の向き・笑っている顔を抽出できた
  - クラスタセンタとクラスタセンタから最も遠い画像はあまり似てない

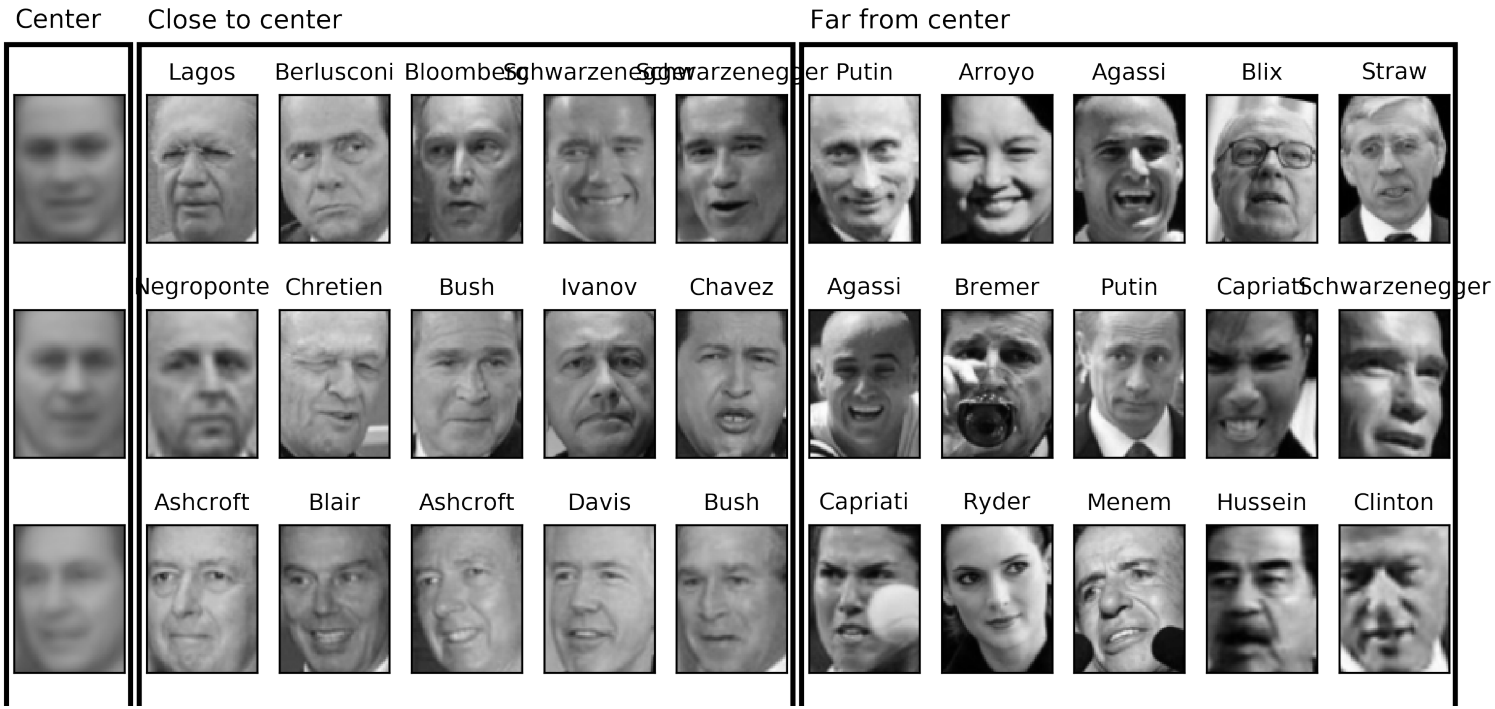

Figure: k-means で 10 クラスタに分割したときのクラスタセンタと、クラスタセンタに近い (遠い) 画像の一例

- 凝集型クラスタリング
  - 10 クラスタに分割したところ、k-means とは異なる分け方をしていた
  - デンドログラムからは最適なクラスタ数はわからなかった
  - 40 クラスタに分割したところ、興味深い特徴をいくつか拾っていた

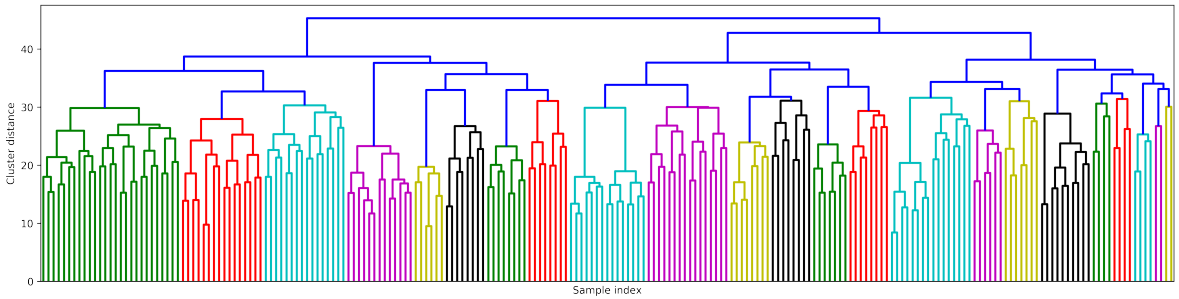

Figure: デンドログラム

Figure: 40 クラスタに分割した時のクラスタの一例