2020･07･13　ゼミ資料

# Pythonではじめる教師なし学習
# 6章4節～6節

1116 17 9036
山口真哉

# やること

・LendingClubのデータに階層クラスタリングを適用してグループ分けをする.

・LendingClubのデータにHDBSCANを適用してグループ分けをする.

・まとめ

# 階層クラスタリング

## 階層クラスタリング（ward法)

- はじめすべて別々で距離が近い順に結合する.
- 事前にクラスタ数を指定する必要がない.
- 適当な位置で区切ってグループ分けをする.

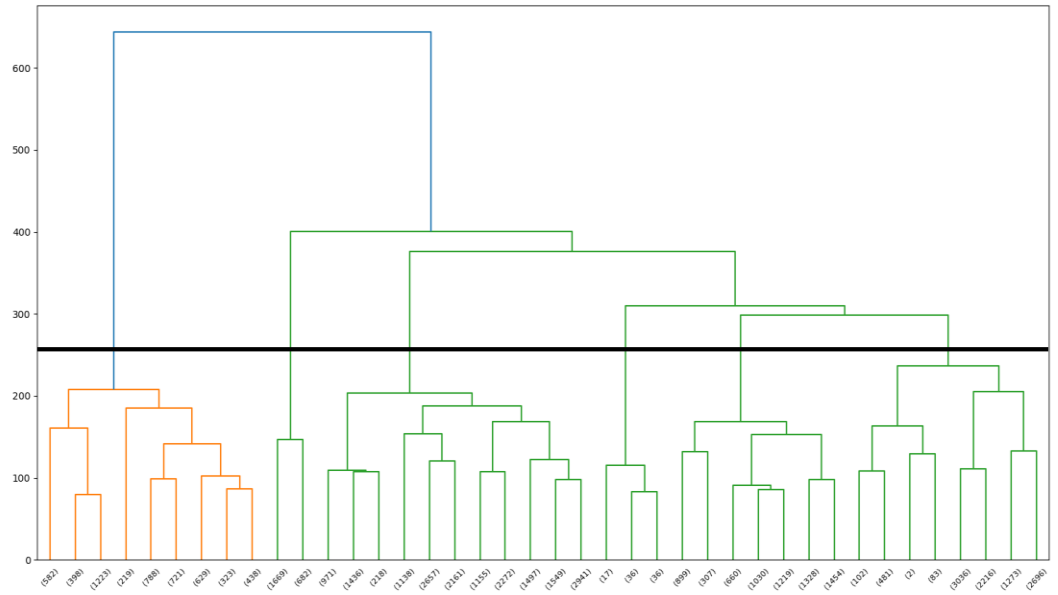


図1

# 階層クラスタリング

- ward法を使う(その他はlinkage_vectorのデフォルト)
- 右図のようにdistance=100で区切る.
  32個のグループが生成される.

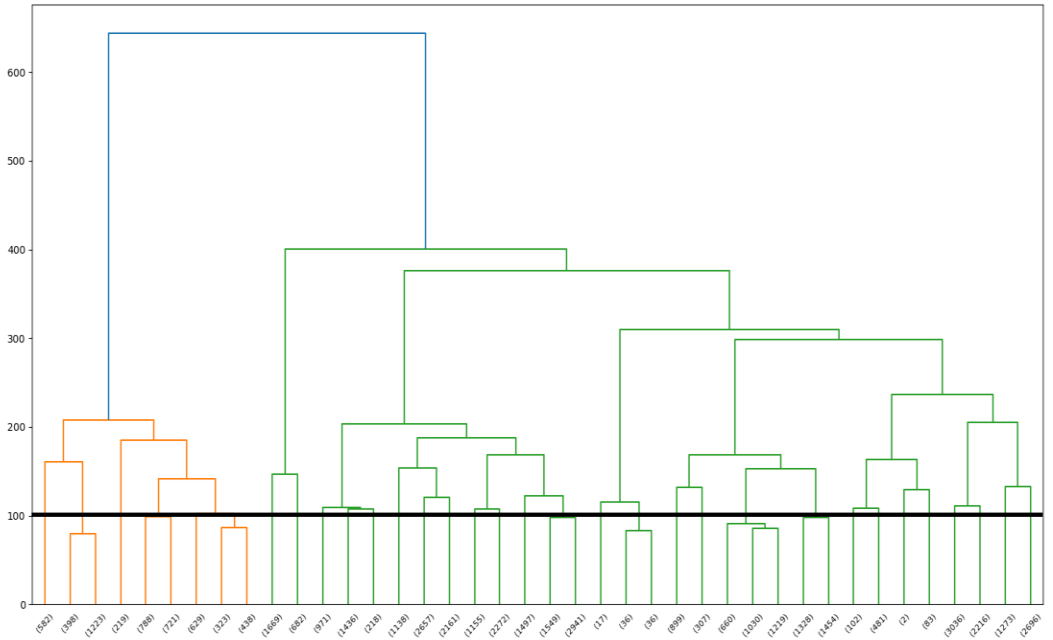


図2

# 階層クラスタリング

## 結果

- 前述の評価関数を適用した結果，全体の精度は36.5%となった．k平均法より精度が劣っている．

- 各クラスタの精度は右の通り．k平均法と同じように精度はクラスタによって大きく異なる．

| | |
|---|---|
| 0 | 0.304124 |
| 1 | 0.219001 |
| 2 | 0.228311 |
| 3 | 0.379722 |
| 4 | 0.240064 |
| 5 | 0.272011 |
| 6 | 0.31456 |
| 7 | 0.26393 |
| 8 | 0.246138 |
| 9 | 0.318942 |
| 10 | 0.302752 |
| 11 | 0.269772 |
| 12 | 0.335717 |
| 13 | 0.330403 |
| 14 | 0.34632 |
| 15 | 0.440141 |

| | |
|---|---|
| 16 | 0.744155 |
| 17 | 0.502227 |
| 18 | 0.294118 |
| 19 | 0.236111 |
| 20 | 0.254727 |
| 21 | 0.241042 |
| 22 | 0.317979 |
| 23 | 0.308771 |
| 24 | 0.284314 |
| 25 | 0.243243 |
| 26 | 0.5 |
| 27 | 0.289157 |
| 28 | 0.365283 |
| 29 | 0.479693 |
| 30 | 0.393559 |
| 31 | 0.340875 |

表3

# HDBSCAN

## HDBSCAN

- 密に集まった点を1つのグループに入れる.
- 事前にクラスタ数を指定する必要がない.
- 外れ値を外れ値として扱うことができる.
- 前回は極端に結果が悪かった.

# HDBSCAN

・ハイパーパラメータは以下の通り

min_cluster_size = 20
min_samples = 20
cluster_selection_method = 'leaf'

・前述の評価関数を適用した結果，全体の精度は32.5%となった．
前の2つより精度が劣っている．

・表4より77%近くが外れ値として扱われていることがわかる．

・表5は各クラスタの精度で比較的安定している．

| cluster | clusterCount |
|---|---|
| -1 | 32708 |
| 7 | 4070 |
| 2 | 3668 |
| 1 | 1096 |
| 4 | 773 |
| 0 | 120 |
| 6 | 49 |
| 3 | 38 |
| 5 | 20 |

表4(クラスタリング結果）

| | |
|---|---|
| 0 | 0.284487 |
| 1 | 0.341667 |
| 2 | 0.414234 |
| 3 | 0.332061 |
| 4 | 0.552632 |
| 5 | 0.438551 |
| 6 | 0.4 |
| 7 | 0.408163 |
| 8 | 0.590663 |

表5（各クラスタの精度)

# まとめ

## まとめ

- LendingClubの2007年から2011年にかけて無担保個人ローンを申し込んだ借入者のデータを対象に教師なしクラスタリング応用システムを構築した.

- k平均法, 階層クラスタリング, HDBSCANを試して精度はそれぞれ39%, 36%, 32%でk平均法が最も優れていた.

- k平均法と階層クラスタリングは全く別物なのでアンサンブルを試す価値はある