2020･06･01　ゼミ資料

Pythonではじめる教師なし学習
5章3.4節～4.2節

1116 17 9036
山口真哉

やること

k-mean法の続きと

階層クラスタリングアルゴリズム

K-mean法

主成分数を変えてみる

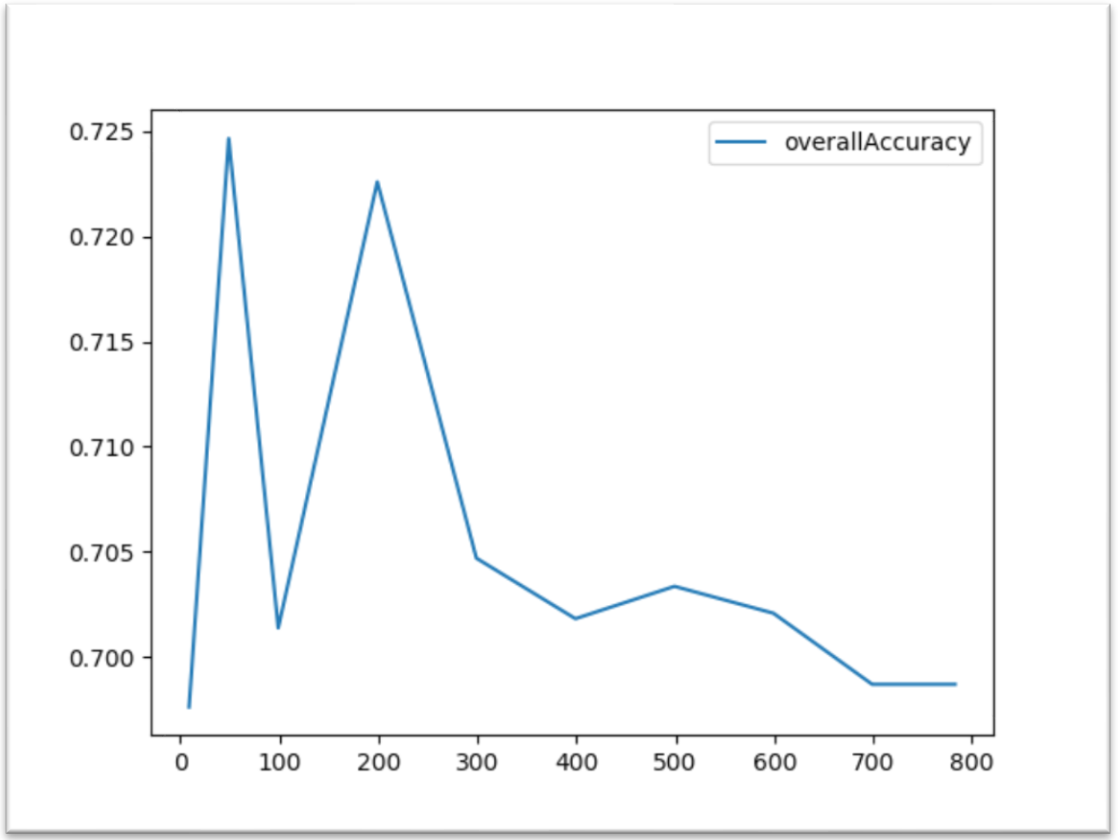
PCAがデータの背後にある構造をうまくとらえられているなら
主成分数を削っても主成分数100と大差ないかも？

主成分数10, 50, 100, 200, 300, 400, 500, 600, 700, 784
で実験してみた.

クラスタの数は20固定で他のパラメータは前と同じ

K-mean法

70%~72.5の間を
拡大したグラフ
であることに注意



X:主成分数
Y:クラスタリング精度

主成分数によらず大体70%付近を安定して推移していることがわかる.

K-mean法

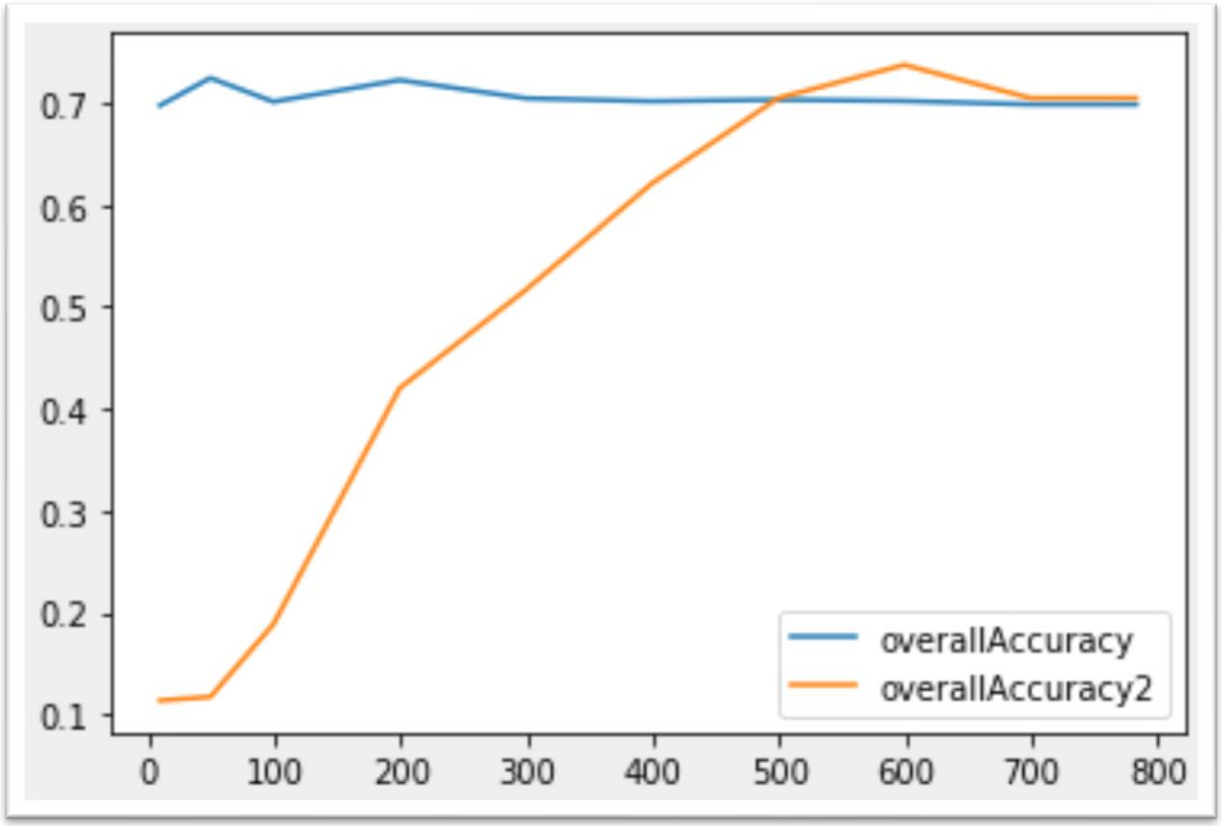
もとのデータセットでk-mean法をやってみる

PCAとの比較をすることでPCAがうまく作用しているか見てみる.

先程同様,
主成分数10, 50, 100, 200, 300, 400, 500, 600, 700, 784
で実験.

クラスタの数は20固定で他のパラメータは前と同じ

K-mean法



X:主成分数
Y:クラスタリング精度
青:PCA
橙:元のデータセット

元のデータセットでは600次元でようやく70%に到達するのに対して
PCAで前処理をしてあげると低次元でもうまくいく(次元削減が働いている)ことがわかる.

階層クラスタリング

階層クラスタリング（のward法)

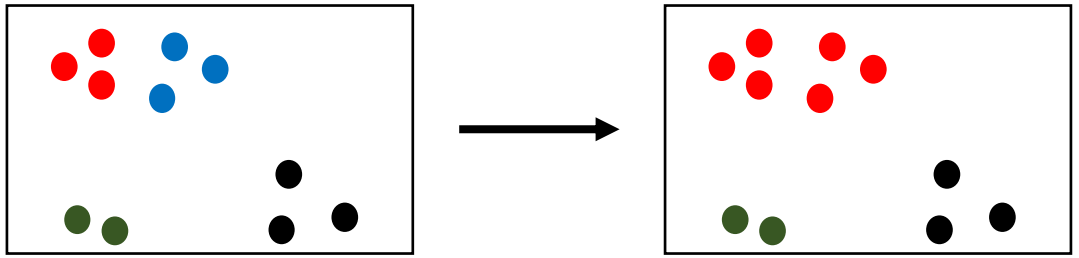
・すべての点が別々のクラスタである状態から始める

While (クラスタ数>1){

今あるクラスタの中で最も距離に近い2つのクラスタを選んで結合.

その操作を記録

}



わかりやすかった

https://mathwords.net/wardmethod

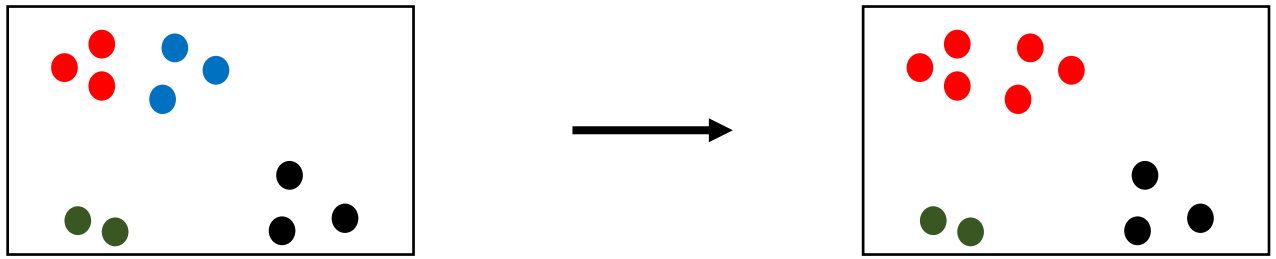
階層クラスタリング

ward法のクラスタ間の距離

$$d(C_1, C_2) = L(C_1 \cup C_2) - (L(C_1) + L(C_2))$$

ただし, $L(C)$ は C の分散

何故この距離かはさっきのリンクを見るとよい



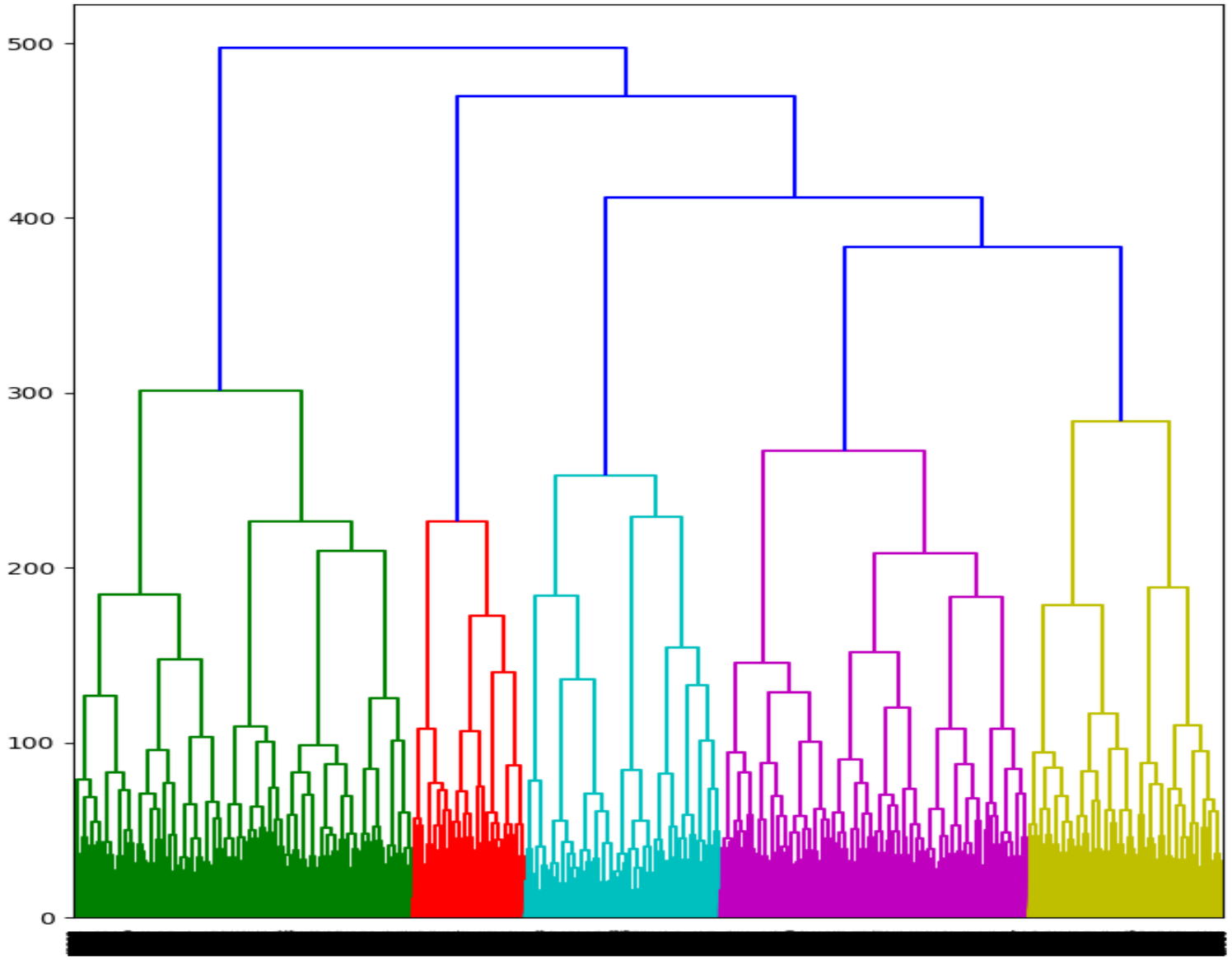
階層クラスタリング

Cutoff=100でward法を実行したときに生成される行列の一部

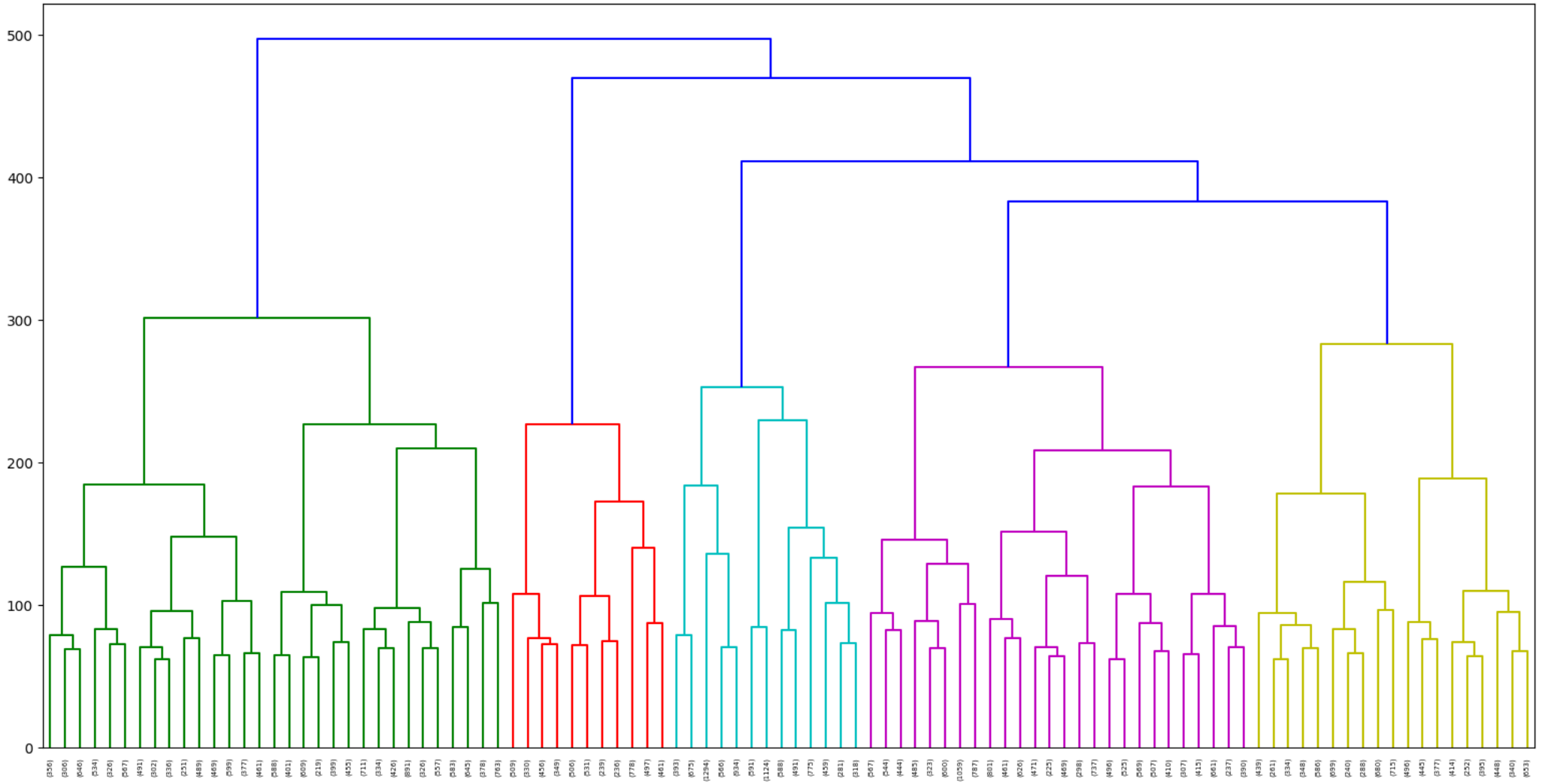
	clusterOne	clusterTwo	distance	newClusterSize
0	42194.0	43025.0	0.562832	2.0
1	28350.0	37674.0	0.590933	2.0
2	26696.0	44705.0	0.621501	2.0
3	12634.0	32823.0	0.627761	2.0
4	24707.0	43151.0	0.637644	2.0
5	20465.0	24483.0	0.662482	2.0
6	466.0	42098.0	0.664156	2.0
7	46542.0	49961.0	0.665527	2.0
8	2301.0	5732.0	0.671106	2.0
9	37564.0	47668.0	0.675121	2.0
10	3375.0	26243.0	0.685907	2.0
11	15722.0	30368.0	0.686356	2.0
12	21247.0	21575.0	0.694361	2.0
13	14900.0	42486.0	0.696768	2.0
14	30100.0	41908.0	0.699282	2.0
15	12040.0	13254.0	0.701136	2.0
16	10508.0	25434.0	0.708637	2.0
17	30695.0	30757.0	0.710037	2.0
18	31019.0	31033.0	0.712047	2.0
19	36264.0	37285.0	0.713131	2.0

	clusterOne	clusterTwo	distance	newClusterSize
49980	99962.0	99975.0	172.296831	3248.0
49981	99951.0	99968.0	178.616930	4590.0
49982	99963.0	99964.0	183.158605	4517.0
49983	99934.0	99974.0	183.939471	3862.0
49984	99971.0	99977.0	184.909538	6510.0
49985	99948.0	99967.0	189.035924	3820.0
49986	99978.0	99982.0	208.571744	8605.0
49987	99956.0	99970.0	209.718978	5614.0
49988	99966.0	99987.0	226.613573	8285.0
49989	99965.0	99980.0	226.698038	4892.0
49990	99941.0	99979.0	229.430787	4627.0
49991	99983.0	99990.0	252.604735	8489.0
49992	99976.0	99986.0	267.070652	13414.0
49993	99981.0	99985.0	283.527956	8410.0
49994	99984.0	99988.0	301.224914	14795.0
49995	99992.0	99993.0	383.194595	21824.0
49996	99991.0	99995.0	411.671391	30313.0
49997	99989.0	99996.0	469.653834	35205.0
49998	99994.0	99997.0	497.468107	50000.0

階層化されてることがわかるグラフ



上位100個の様子



お　わ　り