第五回授業課題

端本知史

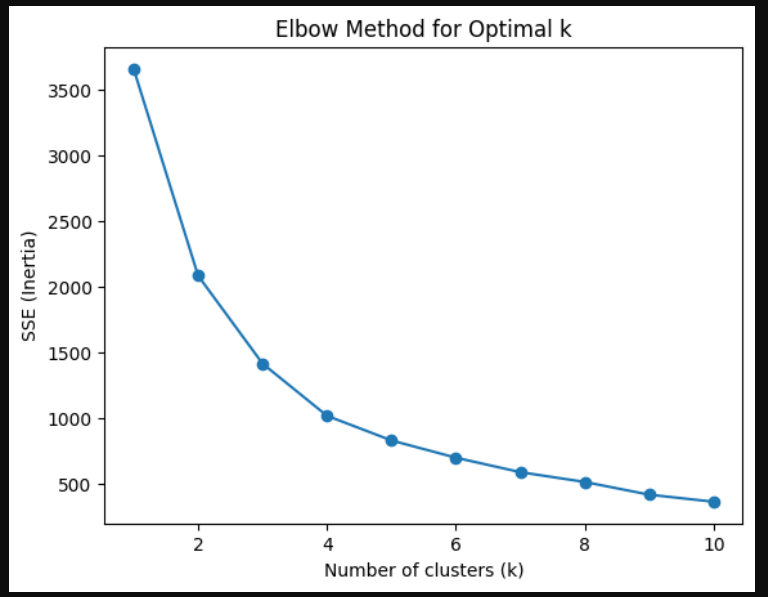
演習1

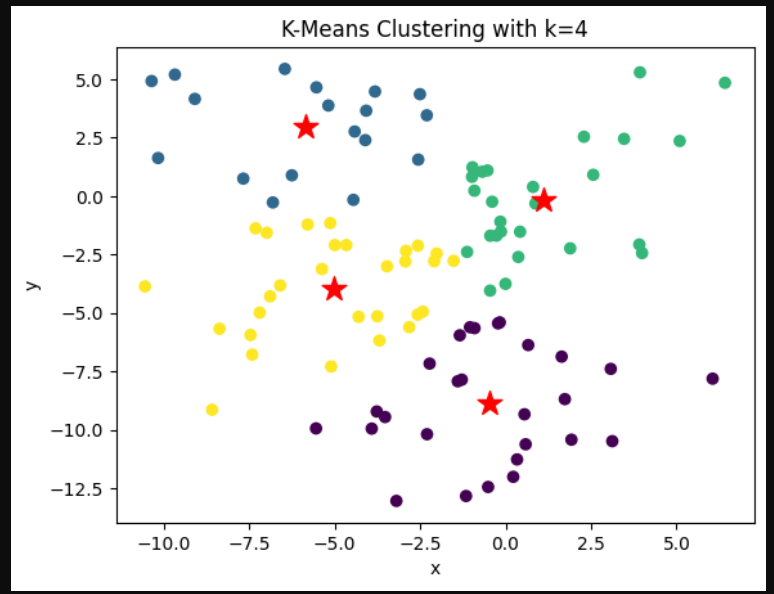
for loopでk=1~10でのSSEを計算し、結果をまとめた折れ線グラフを出力する。

→k = 4辺りから損失関数の減少が緩やかになっていて、過剰適合し始めていると考えられるため、最適なクラスタ数は4と推測する。この手法はエルボー法と呼ばれるらしい(グラフの肘を特定するから)。

他に最適なクラスタ数を特定するための指標としてシルエットスコアという、各インスタンスがどれだけ適切なクラスに属しているかを測るものがあるらしい。

→損失関数にそのような項を加えて正規化したものが階層型クラスタリングだったりするのだろうか?



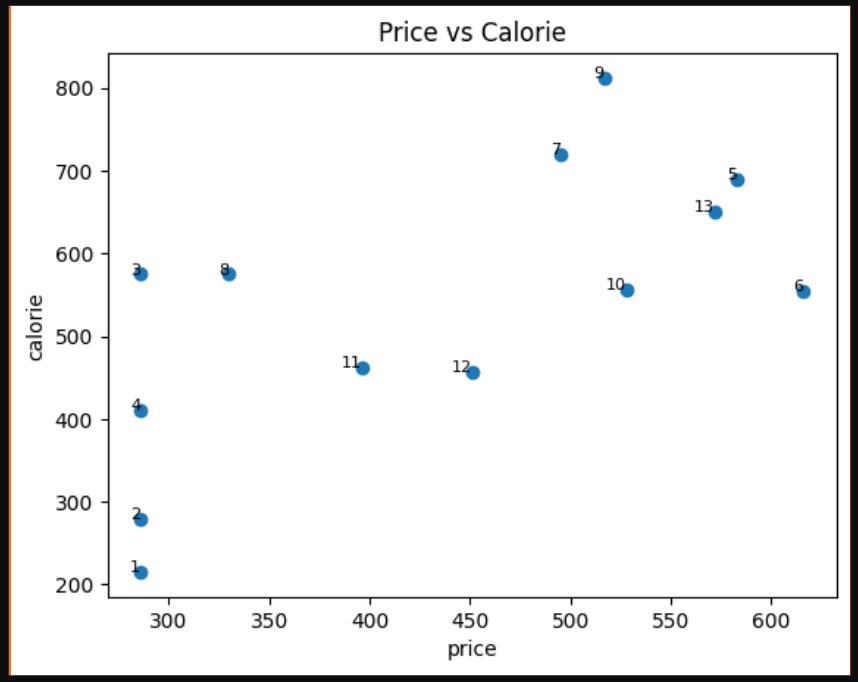


演習2

身近な二次元データセットとして食堂のメニューの価格とカロリーを選択。Matplotlibで日本語が表記できなかったためidで代用。見づらいが逐一元のデータを見てメニュー名を確認する。

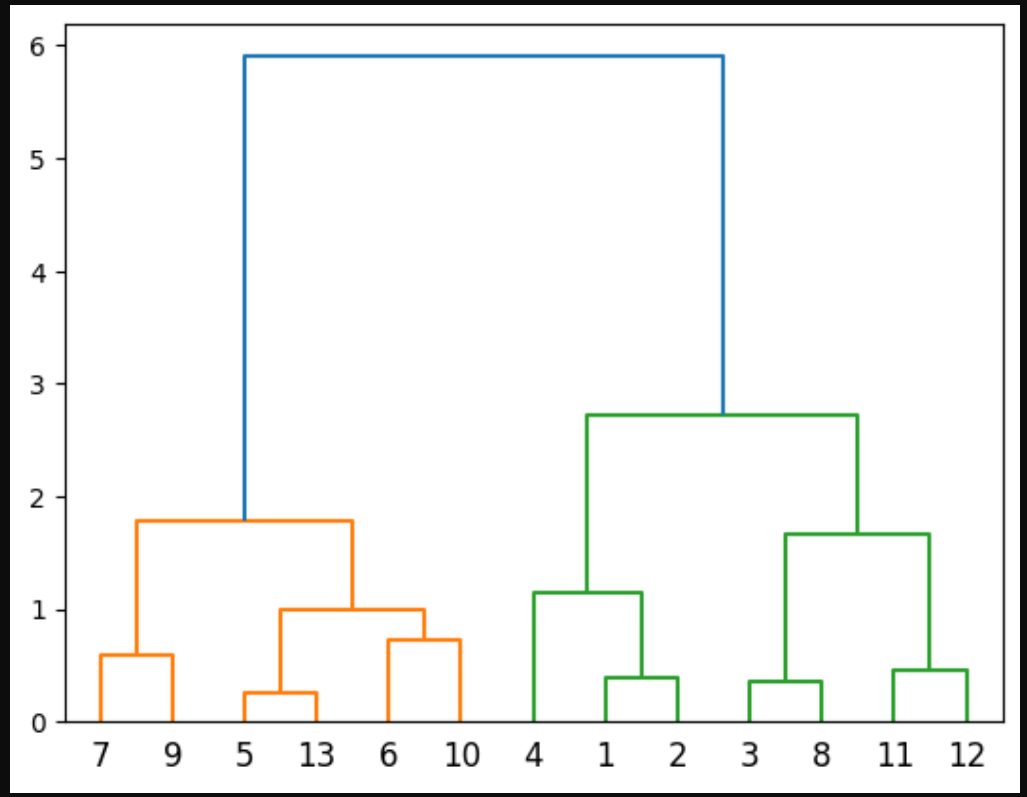
データの概要を確認するために散布図をプロット。比較的線形なデータが得られた？

id	name	price	calorie
1	グリルチキン	286	215
2	照りタルハンバーグ	286	279
3	白身フライタルタル	286	575
4	ゴーヤチャンプル	286	410
5	タコライス	583	689
6	鮭丼	616	555
7	辛みそ豚丼	495	720
8	カレーライス	330	575
9	ヒレカツカレー	517	812
10	熊本ラーメン	528	556
11	かき揚げうどん	396	462
12	醤油ラーメン	451	456
13	大豆ミートのキーマカレー	572	650



階層型クラスタリング

サンプルコードを少し改良したもの(値の標準化を追加)にデータを入れて、Dendrogramを出力。



考察

- 階層型クラスタリングによって作られたクラスの内中くらいのサイズの4つを散布図上で可視化してみる。3番の白身フライは価格に対するカロリーが以上に高い外れ値のように見えるのに11,12という一般的な値と同じクラスになっていることに驚いた。今後の発展としてMethod argumentを色々変えてみてどうなるかも試してみようと思う。