k-meansの派生手法の紹介
～x-means、k-modes、k-prototype

k-meansはいろいろな派生する手法があります

- x-means
  - クラスター数の自動決定をするアルゴリズムが組み込んである手法（Pelleg and Moore, 2000）
  - 補足：クラスター数Kの決定には、BIC（ベイズ情報量規準）を用いて、各クラスターがガウス分布していると想定した際に、BICの値が最大になるように2-meansを収束するまで繰り返している

- k-modes
  - 変数が全てカテゴリ変数の場合を想定した手法（Zhexue Huang, 1997, 1998）
  - 補足：k-meansはサンプル間の距離を用いてクラスタリングを行なっているため、各変数は、大きさ、長さ、購入回数といった連続値を仮定している。そのため、変数がカテゴリ変数の場合には距離を算出できないためk-meansを使うことはできない。そこで、連続値としての距離ではなく、単純にカテゴリ変数の一致度合い（ サンプルA[男性,30代,東京]とサンプルB[女性,20代,東京]では一致度が2/3など）を使うことでクラスタリングを実施する

- k-prototype
  - 変数が連続値とカテゴリ変数の混合データを想定した手法（Zhexue Huang, 1998）
  - 補足：サンプル間の類似性を、連続値の変数はクラスターの平均から、カテゴリ変数は一致度から算出することでクラスタリングを実施する

k-modesのアルゴリズム（１）代表点の選択と類似度の算出

データ

サンプル	性別	年齢	居住地
a	男性	20代	東京
b	男性	20代	名古屋
c	女性	20代	東京
d	女性	20代	大阪
e	女性	40代	大阪
f	男性	40代	東京
g	女性	20代	東京
h	女性	40代	大阪

①クラスタ数kを決め、代表となる点をランダムにサンプリングする（下はk=3）

クラスタ1	a	男性	20代	東京
クラスタ2	d	女性	20代	大阪
クラスタ3	g	女性	20代	東京

②各代表点（今回はa,d,g）からの類似度の算出とクラスターのアサイン（3つの変数のうち何個が一致しているかを算出している）

サンプル	クラスタ1	クラスタ2	クラスタ3	該当クラスタ
a	**3**	1	2	クラスター1
b	**2**	1	1	クラスター1
c	2	2	**3**	クラスター3
d	1	**3**	2	クラスター2
e	0	**2**	1	クラスター2
f	**2**	0	1	クラスター1
g	2	2	**3**	クラスター3
h	0	**2**	1	クラスター2

参考）https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

k-modesのアルゴリズム（２）最頻値による代表点の更新

③各クラスターの変数を最頻値に更新する

サンプル	性別	年齢	居住地	該当クラスタ
a	男性	20代	東京	クラスター1
b	男性	20代	名古屋	クラスター1
c	女性	20代	東京	クラスター3
d	女性	20代	大阪	クラスター2
e	女性	40代	大阪	クラスター2
f	男性	40代	東京	クラスター1
g	女性	20代	東京	クラスター3
h	女性	40代	大阪	クラスター2

	性別	年齢	居住地
クラスタ1	男性	20代	東京
クラスタ2	女性	40代	大阪
クラスタ3	女性	20代	東京

引用元）https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

k-prototype

- 手順
  - (1)プロトタイプをランダムに選ぶ
  - (2)それぞれのサンプルについて距離が一番小さいプロトタイプに割り当て、プロトタイプを更新する
  - (3)動かなくなるか反復数が最大に達するまで(2)を繰り返す

- プロトタイプ（クラスターの中心）
  - k-modesでは各変数について最頻値をプロトタイプの値として使用
  - k-prototypeでは、カテゴリ変数では最頻値、連続値では平均をプロトタイプの値として使用