類似度とクラスタリング

abstract : データ点の間の類似度を定義するさまざまな方法について説明します。また、階層クラスタリング・k-means法への応用を説明します。特にk-means法では、質的変数が含まれるデータに対してはアルゴリズムに変更が必要なことを説明し、k-prototypesというアルゴリズムを導入します。

1. Introduction

多変量解析には、データ点の間の「類似度」に注目するタスクがたくさんあります。類似したデータ点の集まりを作るタスクはクラスタリングとよばれていて、k-means法や階層クラスタリングといったアルゴリズムが代表的です。また、低次元埋め込みといって、2次元や3次元の散布図の上で、似ているデータ点どうしを近い位置に表現するタスクもあります。多次元尺度構成法やt-SNE、UMAPといったアルゴリズムが知られています。

データ点の間の「類似度」とはなにかについて考えてみましょう。D 種類の特徴量を持つ2つのデータ点 $x_1 = (x_{11}, x_{12}, \cdots, x_{1D}), x_2 = (x_{21}, x_{22}, \cdots, x_{2D})$ に対して

$$d_{Euclid}(x_1, x_2) := \sqrt{\sum_{d=1}^{D} (x_{1d} - x_{2d})^2}$$

と定めたとき、これを2点間のユークリッド距離といいます。データ点の間の「類似度」を定義する単純な方法は、ユークリッド距離が小さいほど2つのデータ点は似ていると定める方法です。

しかし、特徴量に質的変数が含まれている場合、ユークリッド距離を用いた「類似度」の定義には問題があります。例えば、3種類のカテゴリ `A`, `B`, `C` からなる質的変数が特徴量に含まれているなら、

- `A` - `B`
- `B` - `C`
- `C` - `A`

を定義しなければ、ユークリッド距離を計算することができない点です。

世の中のデータセットは、大抵の場合、質的変数を含んでいます。分析者がこのようなデータセットに対して、クラスタリングや低次元埋め込みを試みたいとしましょう。分析者がデータ点の間の「類似度」をどのように定義するかは、アルゴリズムから得られる結果に大きく影響することが知られています。このとき、質的変数の存在は、分析者が「類似度」を適切に定義するうえでの障害になりうるのです。

ID	身長	性別	趣味
1	168	"女性"	"数学"
2	151	"男性"	"英会話"
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	177	"男性"	"読書"
N	158	"女性"	"英会話"

Table 1. 質的変数を含むデータセットの例

そこで、このノートでは、データ点の間の「類似度」のさまざまな定義を紹介し、特に質的変数が含まれている場合を重点的に解説します。また応用の話として、質的変数が含まれるデータに対する階層クラスタリング・k-means法の適用と改善点を説明します。

2. 量的変数のみからなるデータセットの類似度

2.1 ミンコフスキー距離

D 種類の特徴量を持つ 2 つのデータ点 $x_1 = (x_{11}, x_{12}, \cdots, x_{1D}), x_2 = (x_{21}, x_{22}, \cdots, x_{2D})$ に対して

$$d_{Euclid}(x_1, x_2) := \sqrt{\sum_{d=1}^{D} (x_{1d} - x_{2d})^2}$$

と定めたとき、これを 2 点間のユークリッド距離といいます。また、p を正の整数とするとき、

$$d_p(x_1, x_2) := \left(\sum_{d=1}^{D} |x_{1d} - x_{2d}|^p \right)^{\frac{1}{p}}$$