**第11回「最尤推定とクラスタリング」**

工学部 37021404 中村裕大

1. **ポスタリゼーションがどのようなアルゴリズムで実装されているのか論ぜよ。**

今回は授業で触れられたk-means法を用いたN色ポスタリゼーションについて、アルゴリズムがどのように構築されているか論ずる。

|  |
| --- |
| **ポスタリゼーションとは**   * 画像処理の一つで、色や階調を減らして**画像を単純化する**方法 * 元の画像を一定の階調レベルに分割し、各領域内で最も代表的な色や値を使用して再構築   **k-means法とは**   * **データをクラスタリング**するための単純かつ効果的な以下のようなアルゴリズム   1. k個の初期中心点をランダムに選択   2. 各データポイントを、最も近い中心点に割り当てる   3. 各クラスタの中心点を、クラスタに所属するデータの平均位置に更新 |

上記よりk-means法を用いてN色ポスタリゼーションのアルゴリズムを実装してみる。

**N-meansポスタリゼーションアルゴリズム**

1. 処理する画像からランダムにN個の色を選択
2. 画像中の全てのピクセルを最も近い色(RGB値)に割り当てる

**Pythonで実装してみた結果**

Git: <https://github.com/KameKingdom/-------11-/blob/main/N_means_postalization.py>

男性の顔の絵

低い精度で自動的に生成された説明 マップ

自動的に生成された説明 テキスト, 本 が含まれている画像

自動的に生成された説明

**元画像 　　　N=2 N=5**

1. **クラスタが存在しそうな二次元データを用意し、条件を変えて、クラスタリング処理 を実行して 収束状況や性能について議論せよ。**

今回評価するクラスタリング処理はプログラムより「**EMアルゴリズム**」と「**SGDアルゴリズム**」の２つである。評価方法は「**収束速度**」「**分離度**」「**外れ値への頑健性**」を主軸とする。

|  |
| --- |
| [データ](https://github.com/KameKingdom/-------11-/blob/main/3_countries_climate.csv)：気象庁の[サイト](https://www.data.jma.go.jp/gmd/cpd/monitor/dailyview/alldata_dl.php?&y=2023&m=6&d=29&e=0&t=0.000000&l=0.000000&k=0&r=1)から昨日(2023 / 6 / 29)の世界の気候データを取得し、３か国のデータをcsvで保存し、**平均気温と降水量**(二次元データ)でクラスタリングを試す。  [結果](https://github.com/KameKingdom/-------11-/blob/main/3_countries_climate_output.png)：**悲惨な分布**でクラスタリングに適しておらず[プログラム](https://github.com/KameKingdom/-------11-/blob/main/cluster_generator.py)で作成することを決意 |

**グラフ, 散布図

自動的に生成された説明** *グラフィカル ユーザー インターフェイス, テキスト, アプリケーション

中程度の精度で自動的に生成された説明*

**(人工的に)準備した二次元データ　　　　　　　　　　 　　　　評価方法**

|  |
| --- |
| **評価方法について**   * **収束速度**：変化量の絶対値の平均の総和、変化量が少ないほど速いと判断 * **分離度**：クラスタの重心のユークリッド距離を計算 * **外れ値に対する頑健性**：各データとクラスタの重心の距離の標準偏差の平均 |

**テーブル

自動的に生成された説明** **テーブル

自動的に生成された説明**

収束速度、分離度はEMアルゴリズムの方が優れており、外れ値への頑健性は僅かな差でSGDアルゴリズムが優れていた。実行速度もEMアルゴリズムの方が速かったので、総合的に**EMアルゴリズムの方が高い評価値を得た**。

**クラスタリングの実験結果**

グラフ, 散布図

自動的に生成された説明　**グラフ, 散布図

自動的に生成された説明**

**EMアルゴリズム　　　　　　　　 　　　　　　　SGDアルゴリズム**

**条件の変更**