機械学習シリーズ：k-means

原文はこちら

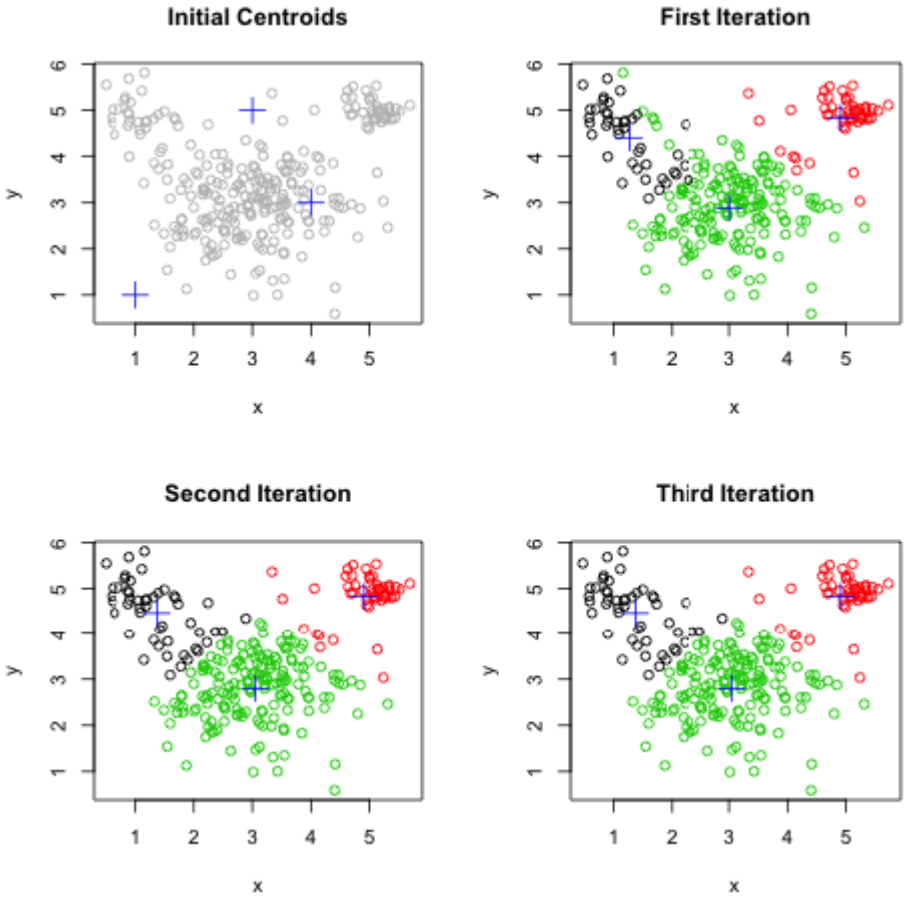
k-means クラスタリングとは？

k 平均クラスタリングは、教師なし学習アルゴリズムであり、類似性に基づいてデータをグループにクラスタ化します。k-means を使用すると、重心で表される k 個のデータクラスタを見つけることができます。ユーザーは、クラスタ数を選択します。

たとえば、購買履歴に基づいて顧客をグループに分けて、異なるグループにターゲットを絞った電子メールを送信したいとします。グループを作成する際には、店舗に何回訪れたか、店頭あるいはオンラインで購入したのか、クーポンを使用するかどうか、購入するアイテムの種類など、さまざまな要素が考慮されるでしょう。

これらの 3 つの異なるグループ、つまり頻繁な購入者、たまにしか買わない購入者、まれな購入者に対して異なる電子メールメッセージが必要だとします。この場合、k-means 関数では、クラスタ数 k を 3 と指定できます。

k-means アルゴリズムの中で、データポイントをクラスタに割り当てるには、データポイントからすべての重心までの距離が計算され、各データポイントに対して最も近い重心が選択されます。次に、各クラスタに割り当てられたデータポイントに基づいて重心が再計算されます。重心の位置を計算するこのサイクルは、重心がもはや大きく動かなくなるか、またはイテレーションの最大回数に達するまで続きます。次の図は、データポイントがどのようにイテレーション中に移動するかを示しています。十字は重心を表し、色付きの円はデータポイントを表します。


Initial Centroids
First Iteration
Second Iteration
Third Iteration
x
y

Vertica での k-means クラスタリング

Vertica での k-means 関数の実行は、2 つのステップからなるプロセスです。

1. k-mean 関数を実行して、クラスタの中心を特定し、その中心をモデルに格納する。
2. 得られたモデルを適用して、識別されたクラスタの中心にデータポイントを割り当てます。

理解を深めるために、例を見ていきましょう。

小麦のデータセットについて

ここでは、UCI から公に入手可能な 210 行と 8 列の小さなデータセットを使用します。このデータセットには、3 つの異なる種類の小麦（Kama, Rosa, Canadian）の幾何学的特性が含まれています。

- Area（エリア）
- Perimeter（境界）
- Compactness（コンパクトさ）
- Length of kernel（穀粒の長さ）
- Width of kernel（穀粒の幅）
- Asymmetry coefficient（アシンメトリー係数）
- Length of kernel groove（穀粒の溝の長さ）
- Type: 1=Kama, 2=Rosa or 3=Canadian（タイプ： 1=Kama, 2=Rosa or 3=Canadian）

あなたが小麦愛好家、農家、あるいはワインの専門家である場合、小麦のクラスタリングはあなたに何かを意味するかもしれません。他の人にとっては、このデータセットを選んだのは、k-means アルゴリズムで比較的理解しやすいためです。

データセットのロード

小麦のデータデットを格納するために、次のテーブルを使用します。

```
=> CREATE TABLE PUBLIC.wheat (
       area FLOAT
       ,perimeter FLOAT
       ,compactness FLOAT
       ,klength FLOAT
       ,width FLOAT
       ,asymmetry FLOAT
       ,glength FLOAT
       ,type INTEGER
       ,
       );
```

テーブルには、それぞれのカラムですべてのプロパティが定義されています。

データベースにデータをロードする前に、データセットに下記の sed コマンドを実行することにより、いくつかのタブ文字を削除しました。

```
$ sed 's/\t\t*/\t/g' seeds_dataset.txt | vsql -c "
```

データをロードするために、次のコマンドを実行します。

```
=> COPY public.wheat (
area, perimeter, compactness, klength,
```

```
width, asymmetry, glength, type
)
FROM STDIN
DELIMITER E'\t'
ABORT ON ERROR
DIRECT
;
```

小麦のデータセットのクラスタリング

public.wheat にデータセットをロードした後、Vertica の k-means 関数を使用して小麦のデータセットをクラスタリングします。
k-means を実行する前に、データセットにクラスタの数を定義する必要があります。これは困難な場合もありますが、今回のデータセットでは、3 種類の小麦があるので、3 つのクラスタが必要であると想定できます。

public.wheat テーブルに対して、Vertica の k-means 関数を実行してみましょう。

```
=> SELECT KMEANS('wmod1', 'public.wheat', '*', 3,'--
exclude_columns=type');
```

この SQL 文は、3 つのクラスタを含む wmod1 という k-means モデルを作成します。type 以外の public.wheat テーブルのすべての列が使用されます。

モデルを作成したため、次のステートメントでモデルの内容にアクセスできます。

```
=> SELECT SUMMARIZE_MODEL('wmod1');
```

SUMMARIZE_MODEL()は、次の情報を返します。

Column	Cluster 0	Cluster 1	Cluster 2
Area	11.9768421	14.6788235	18.7307143
Perimeter	13.2563158	14.4514706	16.3067857
Compactness	0.8510079	0.8821382	0.8848821
Klength	5.2288684	5.5525294	6.2260357
Width	2.8636316	3.2895588	3.7175714
Asymmetry	4.8777895	2.4767118	3.3530000
glength	5.0730000	5.1879118	6.0807143

上記の表は、クラスタの中心を示しています。各中心は、7 つのプロパティの値の組み合わせによって表されます。

注意：このアルゴリズムは、中心の最初のセットをランダムに選択します。したがって、中心のベクトルの次数とその値の両方が、異なる実行でわずかに変化する可能性があります。

SUMMARIZE_MODEL()は、次の情報を提供します。

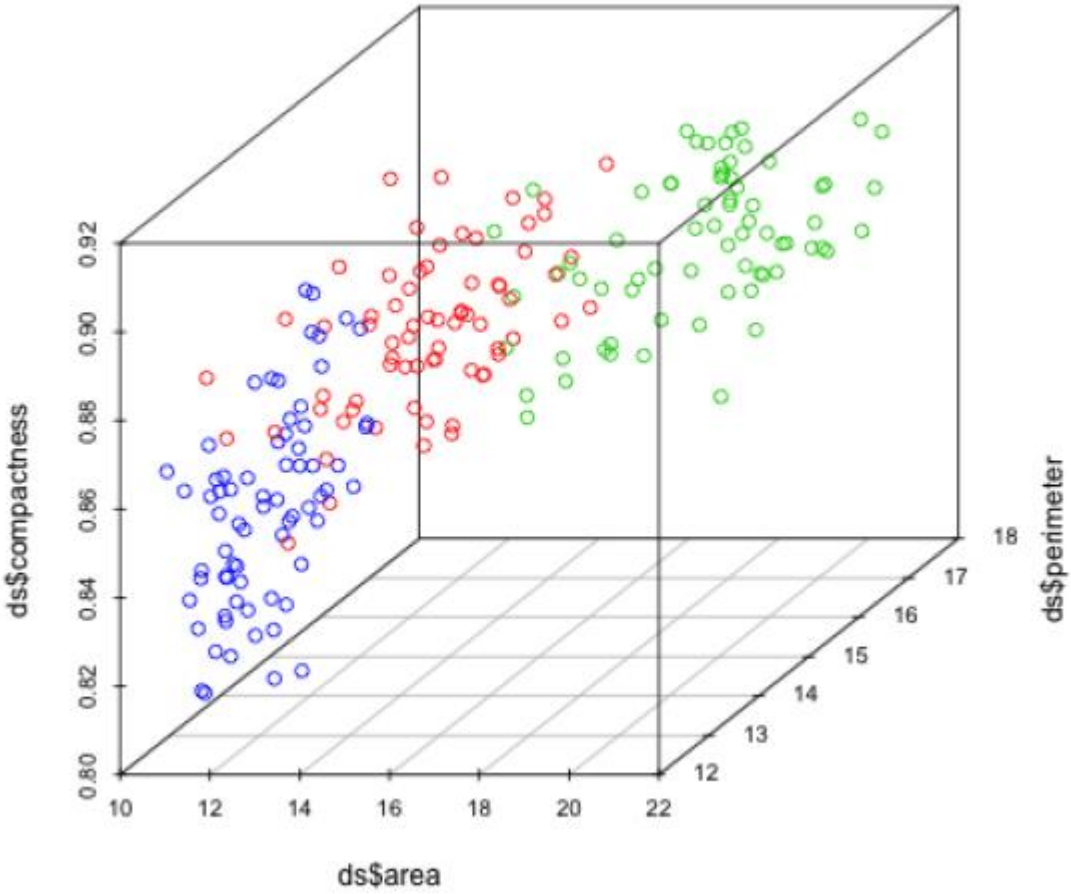
- クラスタの二乗和内（WCSS）－この値は、クラスタの結合を測定します。この値が小さいほど、クラスタがよりコンパクトになります。各データポイントが 1 つの重心と一致するようにデータポイントの数と同数のクラスタがある場合、WCSS はゼロです。
- クラスタの二乗和の間（BCSS）－この値は、クラスタ間の分離を測定します。クラスタがひとつだけの場合、この値はゼロです。
- 総二乗和（TSS）－BCSS とすべての WCSS の合計に等しい値です。
- BCSS/TSS 比－クラスター内の結合度とクラスター分離度が高いほど、この値は 1 に近くなります。

クラスタへのデータポイントの割り当て

今度は、Vertica の APPLY_KMEANS 関数を使用して、前の手順で開発したモデルを使用してテーブル内のデータポイントをマークすることができます。次の SQL 文は、シードタイプ上のクラスタ分散を提供します。

```
=> SELECT
SUM(CASE cid WHEN 0 THEN cnt ELSE 0 END) AS clust_0,
SUM(CASE cid WHEN 1 THEN cnt ELSE 0 END) AS clust_1,
SUM(CASE cid WHEN 2 THEN cnt ELSE 0 END) AS clust_2
FROM(
SELECT type, APPLY_KMEANS(area, perimeter,
compactness, klength, width, asymmetry, glength
USING PARAMETERS OWNER='dbadmin',
MODEL_NAME='wmod1') AS cid,
COUNT(*) as cnt
FROM public.wheat
GROUP BY 1, 2 ) x
GROUP BY 1 ORDER BY 1
;
type          | clust_0     | clust_1     | clust_2
--------------+-------------+-------------+---------
Canadian      | 34          | 0           | 0
Kama          | 5           | 34          | 0
Rosa          | 0           | 4           | 33
```

クラスターに割り当てられていない少数の種がそのタイプに対応しています。これは、データポイントをクラスタリングするために使用される幾何学的特性のいくつかが互いに重なるためです。領域、周囲、およびコンパクトさをプロットすることで、このオーバーラップをより視覚化することができます。



k-means についてのこの紹介が役立つことを願っています。Vertica の機械学習シリーズの次のブログ記事に注目してください。k-means の詳細については、Vertica のドキュメントの k-means を参照してください。

Vertica のすべての機械学習の関数については、Vertica ドキュメントの Machine Learning for Predictive Analytics にも記載されています。