言語処理のための機械学習入門

3. クラスタリング
3.1. 準備
3.2. 凝集型クラスタリング
3.3. k-平均法

永田 純平

クラスタリング(Clustering)

- データの集合を共通の特徴を持つもつものごとにグループ化する

学習(Learning)

- あるデータが与えられたときそれに対応して何らかのモデルや処理を導くこと

凝集型クラスタリング

- 単純に最も類似しているデータをクラスタリングする手法
- 与えられたそれぞれの事例にひとつずつクラスタを割り当て、類似している事例同士を融合させて、ひとつのクラスタとする
- この作業の繰り返しを表した図を樹形図という。

クラスタ同士の類似度を測る方法

単連結法(single-link method)

二つのクラスタ内の最も近い事例の類似度をクラスタの類似度とする

$$sim(C_i, C_j) = \max_{x_k \in C_i, x_l \in C_j} sim(x_k, x_l)$$

- 完全連結法(complete-link method)

二つのクラスタ内の最も遠い事例の類似度をクラスタの類似度とする

$$sim(C_i, C_j) = \min_{x_k \in C_i, x_j \in C_j} sim(x_k, x_j)$$

クラスタ同士の類似度を測る方法

- 重心法(centroid method)

　各クラスタが含む事例の重心ベクトルを代表ベクトルとして、これらの代表ベクトル同士の類似度をクラスタの類似度とする

$$sim(C_i, C_j) = sim(\frac{1}{|C_i|}\sum_{x \in C_i} x\ , \frac{1}{|C_j|}\sum_{x \in C_j} x)$$

k-平均法

- k個にクラスタリングする方法

<手順>

- 適当にk個のクラスタの代表ベクトルを決定し、各事例に一番近いクラスタに帰属させる
- 各クラスタの重心ベクトルを新たに代表ベクトルとし、再び各事例をクラスタに帰属させる。
- この手順を収束するまで繰り返す。