目次

K-Means 法によるクラスタリング

- 扱う問題
  - 多次元空間上のデータ点集合を考える
  - 各データが属するクラスタを決定する問題を考える

- 問題設定
  - D 次元ユークリッド空間における、確率変数 $\boldsymbol{x}$ を観測
  - $\boldsymbol{x}$ の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $(\boldsymbol{x}_i \in \mathbb{R}^D)$
  - データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を、K 個のクラスタに分割
  - K は、既知の定数であるとする

- クラスタとは
  - クラスタとは、簡単に言えば近接するデータの集合である
  - クラスタの内部のデータ点間の距離が、クラスタの外側のデータとの距離と比べて、小さいようなデータの集合

K-Means 法によるクラスタリング

- クラスタを代表するベクトルの表現
  - 各クラスタを代表する K 個の D 次元ベクトル $\mathcal{M} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ $(\boldsymbol{\mu}_k \in \mathbb{R}^D)$ を導入する
  - これらのベクトル $\mathcal{M} = \{\boldsymbol{\mu}_k\}$ を、プロトタイプという
  - k 番目のクラスタ ($\boldsymbol{\mu}_k$ が支配するクラスタ) を $\mathcal{M}(\boldsymbol{\mu}_k)$ と記す

  - ベクトル $\boldsymbol{\mu}_k$ は、k 番目のクラスタに対応するプロトタイプである
  - $\boldsymbol{\mu}_k$ は、$\mathcal{M}(\boldsymbol{\mu}_k)$ に属するデータ点の平均、即ち k 番目のクラスタの中心である

- 解くべき問題
  - N 個の全データ点を、うまくクラスタに割り振る
  - 各データ点から、対応する (そのデータ点が属するクラスタの) プロトタイプ $\boldsymbol{\mu}_k$ への、二乗距離の総和を最小化する

K-Means 法によるクラスタリング

- データ点のクラスタへの割り当てを表す変数
  - 各データ $\boldsymbol{x}_i$ $(1 \leq i \leq N)$ に対して、二値の指示変数 $r_{ik} \in \{0, 1\}$ $(k = 1, \ldots, K)$ を定める
  - r_{ik} は、データ $\boldsymbol{x}_i$ が、K 個あるクラスタのうちの、どれに割り当てられるのかを表す
  - データ点 $\boldsymbol{x}_i$ がクラスタ k に割り当てられるときに、$r_{ik} = 1$ となり、$j \neq k$ について $r_{ij} = 0$ である (1-of-K 符号化法という)

$$r_{ik} = \begin{cases} 1 & (\boldsymbol{x}_i \in \mathcal{M}(\boldsymbol{\mu}_k) \text{ の場合}) \\ 0 & \text{それ以外} \end{cases} \tag{1}$$

K-Means 法によるクラスタリング

- 目的関数の定義
  - 目的関数を以下のように定義する
  - 各データ点 $\boldsymbol{x}_i$ と、$\boldsymbol{x}_i$ に割り当てたベクトル $\boldsymbol{\mu}_k$($\boldsymbol{x}_i$ が属するクラスタのプロトタイプ) との、二乗距離の総和

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \tag{2}$$

  - 上式の J を最小化するような、$\{r_{ik}\}$ と $\{\boldsymbol{\mu}_k\}$ を求めるのが目標

K-Means 法によるクラスタリング

- 目的関数 J の $\{r_{ik}\}$ に関する最小化
  - 目的関数の式は、各 i について独立である

$$J = \sum_i \sum_k r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \tag{3}$$

  - $\{r_{ik}\}$ を決定する方法は簡単
  - 各 i について、$||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2$ が最小となるような k に対し、$r_{ik} = 1$ とする
  - それ以外のクラスタ $j \neq k$ については、$r_{ik} = 0$ とする

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases} \tag{4}$$

  - 各データ点 $\boldsymbol{x}_i$ を、$\boldsymbol{x}_i$ と最も近い $\boldsymbol{\mu}_k$ に割り当てることに相当
  - 各データ点 $\boldsymbol{x}_i$ を、クラスタを代表するベクトル (クラスタの中心) との二乗距離が最小になるような、クラスタに割り当てる

K-Means 法によるクラスタリング

- 目的関数 J の $\{\boldsymbol{\mu}_k\}$ に関する最小化
  - J を $\boldsymbol{\mu}_k$ について偏微分して 0 とおく

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} J &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_i r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_i r_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \boldsymbol{x}_i + \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{\mu}_k - \boldsymbol{x}_i^T \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} (2\boldsymbol{\mu}_k - 2\boldsymbol{x}_i) = 0 \qquad (5)
\end{aligned}
$$

K-Means 法によるクラスタリング

- ここで、次の関係を用いている

$$
\begin{aligned}
& \boldsymbol{\mu}_k^T \boldsymbol{x}_i \\
= \quad & \left(\boldsymbol{\mu}_k^T \boldsymbol{x}_i\right)^T \qquad (\because \text{スカラーであるため転置してもよい}) \\
= \quad & \boldsymbol{x}_i^T \left(\boldsymbol{\mu}_k^T\right)^T \qquad (\because (\boldsymbol{ab})^T = \boldsymbol{b}^T \boldsymbol{a}^T) \\
= \quad & \boldsymbol{x}_i^T \boldsymbol{\mu}_k
\end{aligned}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{a} = \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{a}^T \boldsymbol{x} = \boldsymbol{a} \tag{6}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} = \left(\boldsymbol{B} + \boldsymbol{B}^T\right) \boldsymbol{x} \quad (\text{2 次形式}) \tag{7}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{x} = 2\boldsymbol{x} \tag{8}
$$

$$
\because \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{x} = \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{I} \boldsymbol{x} = \left(\boldsymbol{I} + \boldsymbol{I}^T\right) \boldsymbol{x} = 2\boldsymbol{I}\boldsymbol{x} = 2\boldsymbol{x}
$$

K-Means 法によるクラスタリング

- これより、$\boldsymbol{\mu}_k$ について解くと次のようになる

$$
\begin{aligned}
\sum_i 2r_{ik}(\boldsymbol{\mu}_k - \boldsymbol{x}_i) &= 0 \\
\sum_i r_{ik}(\boldsymbol{\mu}_k - \boldsymbol{x}_i) &= 0 \\
\sum_i r_{ik}\boldsymbol{\mu}_k &= \sum_i r_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k \left(\sum_i r_{ik} \right) &= \sum_i r_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k &= \frac{1}{\sum_i r_{ik}} \sum_i r_{ik}\boldsymbol{x}_i
\end{aligned}
\tag{9}
$$

- $\sum_i r_{ik}$ は、クラスタ k に属するデータの個数である
- $\sum_i r_{ik} = N_k$ と表すことがある

K-Means 法によるクラスタリング

- r_{ik} は、i 番目のデータ $\boldsymbol{x}_i$ が、クラスタ k に割り当てられているときにのみ、1 となる
- $\sum_i r_{ik}\boldsymbol{x}_i$ は、クラスタ k に属しているデータのベクトル $\boldsymbol{x}_i$ の合計である

- 従って、$\boldsymbol{\mu}_k$ は、k 番目のクラスタに割り当てられた、全てのデータ点 $\boldsymbol{x}_i$ の平均値である

- この意味で、$\boldsymbol{\mu}_k$ のことを、クラスタ k の平均ベクトル、重心、セントロイドということもある

K-Means 法によるクラスタリング

- $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ についての最適化
  - 目的関数 J を、$\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ について最小化する式は次のようになる

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i \\ r_{ik} &= \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases} \end{aligned}$$

  - $\boldsymbol{\mu}_k$ を計算する式の中に r_{ik} が、r_{ik} を計算する式の中に $\boldsymbol{\mu}_k$ が入っている
  - $\boldsymbol{\mu}_k$ を求めるためには r_{ik} が、r_{ik} を求めるためには $\boldsymbol{\mu}_k$ が既知でなければならない
  - 従って、$\boldsymbol{\mu}_k$ と r_{ik} の両方を同時に最適化することはできない

  - どうすれば目的関数 J を、パラメータ $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ の両方について最小化できるか?

K-Means 法によるクラスタリング

- $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ についての最適化
  - $\boldsymbol{\mu}_k$ と r_{ik} の両方を同時に最適化することはできない
  - $\boldsymbol{\mu}_k$ と r_{ik} を交互に最適化すればよい
  - $\boldsymbol{\mu}_k$ と r_{ik} のそれぞれを最適化する、2 つのステップを交互に繰り返す
  - $\boldsymbol{\mu}_k$ と r_{ik} の最適化は、次のように行うことができる

1. $\mathcal{M} = \{\boldsymbol{\mu}_k\}$ の初期値を設定
   - N 個のデータ $\mathcal{D} = \{\boldsymbol{x}_i\}$ を、ランダムなクラスタに割り振って、各クラスタの平均ベクトル $\{\boldsymbol{\mu}_k\}$ を求める
   - データ $\mathcal{D}$ からランダムに選択した K 個のデータ点を、各クラスタの中心 $\boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$ とすることもできる

K-Means 法によるクラスタリング

2. 第 1 フェーズでは、$\boldsymbol{\mu}_k$ を固定しつつ、r_{ik} について J を最小化

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases}$$

3. 第 2 フェーズでは、r_{ik} を固定しつつ、$\boldsymbol{\mu}_k$ について J を最小化

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i$$

4. (2) と (3) を、$\boldsymbol{\mu}_k$ と r_{ik} が収束するまで繰り返す

- 上記の 2 つのステップが、後述する EM アルゴリズムの E(Expectation) ステップと M(Maximization) ステップに対応する

K-Means 法によるクラスタリング

- データ点へのクラスタの再割り当てと、クラスタの平均ベクトルの再計算

- この 2 段階の処理を、クラスタの再割り当てが起こらなくなるまで (2 段階の処理を行っても、データが属するクラスタが変化しなくなるまで) 繰り返す

- 各フェーズで J の値が減少するので、アルゴリズムの収束が保証される
- 大域的最適解ではなく、局所最適解に収束する可能性はある

K-Means 法によるクラスタリング

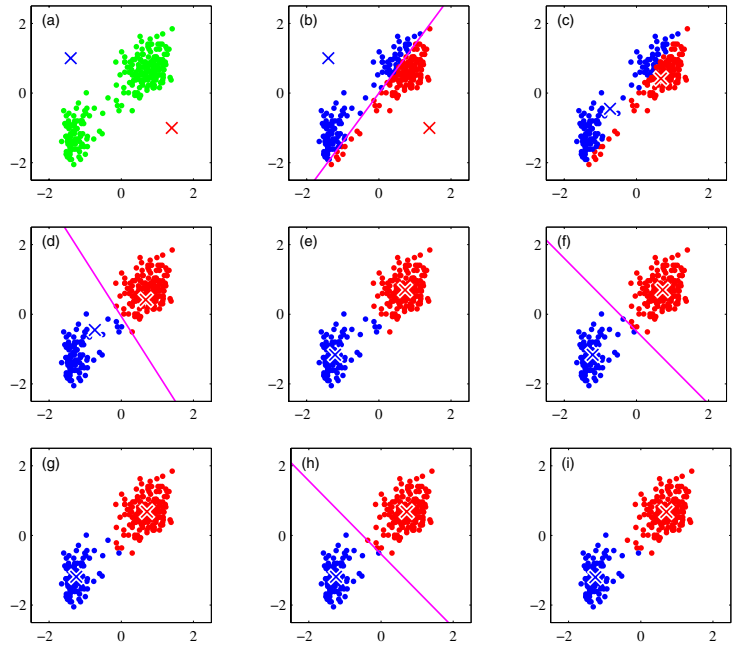

Figure 9.1 Illustration of the K-means algorithm using the re-scaled Old Faithful data set. (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on. (c) In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive E and M steps through to final convergence of the algorithm.

図 1: K-Means アルゴリズムの動作

K-Means 法によるクラスタリング

Figure 9.2 Plot of the cost function J given by (9.1) after each E step (blue points) and M step (red points) of the K-means algorithm for the example shown in Figure 9.1. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.

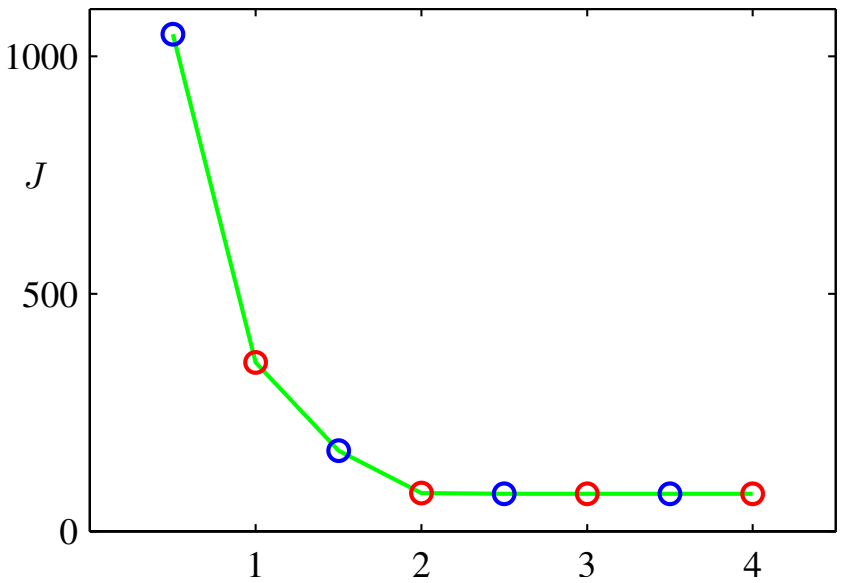

図 2: 目的関数 J の値の遷移

K-Means 法によるクラスタリング

- 素朴な実装では速度が遅いことがある
  - $\{r_{ik}\}$ の更新 (E ステップ) において、各データ点と、各平均ベクトルの全ての組み合わせ間の距離を計算する必要がある

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases}$$

  - K-Means の高速化について、これまでに様々な手法が提案されてきた
  - 近接するデータ点同士が、同一の部分木に属するような木構造を採用する方法
  - 距離の三角不等式を利用して、不必要な距離計算を避ける方法

逐次版の K-Means 法

- 逐次版の K-Means 法の導出
  - これまでに紹介した K-Means 法は、利用する全てのデータ $\mathcal{D} = \{\boldsymbol{x}_i\}$ が、最初から用意されていることが前提であった
  - ここでは、Robbins-Monro 法を使って、オンラインのアルゴリズムを導出する

Robbins-Monro 法

- Robbins-Monro 法
  - 逐次学習アルゴリズムを導出するための手法
  - 関数 $f(\theta^*) = 0$ を満たす根 θ^* を、逐次的に計算するための式を与える

  - 同時分布 $p(z, \theta)$ に従う確率変数のペア θ, z について、関数 $f(\theta)$ は、θ が与えられたときの z の条件付き期待値 $\mathbb{E}[z|\theta]$ として、定義される

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int z p(z|\theta) dz \tag{10}$$

  - このとき、根 θ^* の逐次計算式は、次のように記述される

$$\theta^{(N)} = \theta^{(N-1)} - \eta_{N-1} z(\theta^{(N-1)}) \tag{11}$$

  - η は学習率
  - $z(\theta^{(N)})$ は、確率変数 θ が値 $\theta^{(N)}$ をとるときに観測される z の値

Robbins-Monro 法

- Robbins-Monro 法を適用するための条件
  - Robbins-Monro 法を使用するためには、満たすべき条件が幾つか存在する

1. z の条件付き分散 $\mathbb{E}\left[(z-f)^2|\theta\right]$ が、有限でなければならない

$$\mathbb{E}\left[(z-f)^2|\theta\right] = \int (z-f(\theta))^2 p(z|\theta)dz < \infty \tag{12}$$

2. $\theta > \theta^*$ では $f(\theta) > 0$、$\theta < \theta^*$ では $f(\theta) < 0$ を仮定 (このように仮定しても一般性は失われない)

Robbins-Monro 法

3 学習率の系列 $\{\eta_N\}$ は次の条件を満たす

$$\lim_{N\to\infty} \eta_N = 0 \tag{13}$$

$$\sum_{N=1}^{\infty} \eta_N = \infty \tag{14}$$

$$\sum_{N=1}^{\infty} \eta_N^2 < \infty \tag{15}$$

- 最初の条件は、推定系列 $\theta^{(N)}$ が、目標の根 θ^* に収束していくように、θ の修正量を減らしていくことを保証
- 次の条件は、根 θ^* 以外の値に収束しないことを保証
- 最後の条件は、分散を有限に抑えることで、いつまで経っても収束しないことを防止

Robbins-Monro 法

Figure 2.10 A schematic illustration of two correlated random variables z and θ, together with the regression function $f(\theta)$ given by the conditional expectation $\mathbb{E}[z|\theta]$. The Robbins-Monro algorithm provides a general sequential procedure for finding the root $\theta^\star$ of such functions.

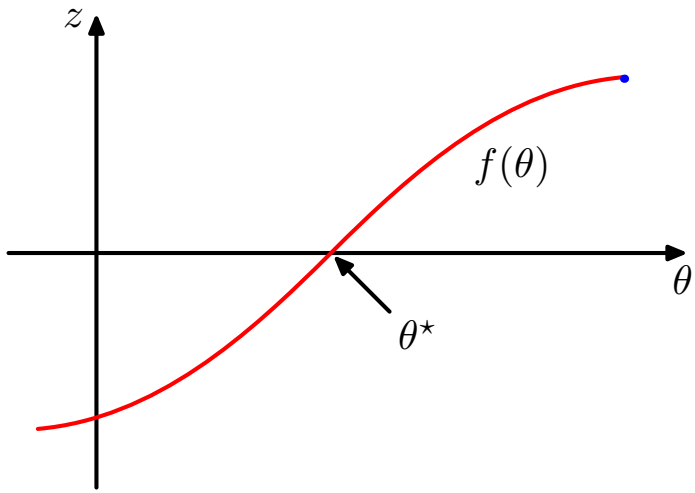

図 3: $f(\theta)$ のグラフ

逐次版の K-Means 法

- 逐次版の K-Means 法の導出
  - 先程のパラメータ θ は、$\{r_{ik}\}$ と $\{\boldsymbol{\mu}_k\}$ である
  - パラメータの最適解 $\{\boldsymbol{\mu}_k^*\}$ は、以下の式を満たす

$$\left.\frac{\partial}{\partial \boldsymbol{\mu}_k} J\right|_{\boldsymbol{\mu}_k^*} = \left.\frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{i=1}^{N} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2\right|_{\boldsymbol{\mu}_k^*} = 0 \tag{16}$$

これは以下の式と等価である $(N_k = \sum_i r_{ik})$

$$\left.\frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2\right|_{\boldsymbol{\mu}_k^*} = 0 \tag{17}$$

これ以降、クラスタ k に属するデータのみを考えることにして、これを $\boldsymbol{x}_j$ と書く $(j = 1, \ldots, N_k)$

逐次版の K-Means 法

- このとき、上式から r_{ik} を省略でき、次のようになる

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_{j=1}^{N_k} ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \Bigg|_{\boldsymbol{\mu}_k^*} = 0 \tag{18}$$

- $N_k \to \infty$ の極限を取って次のように変形する

$$\begin{aligned}
& \lim_{N_k \to \infty} \frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_j ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \\
= \quad & \lim_{N_k \to \infty} \frac{1}{N_k} \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_j ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \\
\simeq \quad & \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\mu}_k} ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \middle| \boldsymbol{\mu}_k \right] \quad (\boldsymbol{x} \text{ はクラスタ } k \text{ に属する}) \\
= \quad & \mathbb{E}\left[2(\boldsymbol{x} - \boldsymbol{\mu}_k) | \boldsymbol{\mu}_k \right] = f(\boldsymbol{\mu}_k)
\end{aligned} \tag{19}$$

逐次版の K-Means 法

- これより、K-Means 法は、関数 $f(\boldsymbol{\mu}_k^*) = 0$ をみたす根 $\boldsymbol{\mu}_k^*$ を求める問題に帰着させられる
- 従って、Robbins-Monro 法を適用すると、$\boldsymbol{\mu}_k$ の逐次更新式は、以下のようになる

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} - \eta_j(\boldsymbol{x}_j - \boldsymbol{\mu}_k^{\text{old}}) \quad (\boldsymbol{x}_j \text{はクラスタ } k \text{ に属する}) \tag{20}$$

- η_j は学習率パラメータであり、一般に、j の増加に伴って単調減少するように設定される
- クラスタ k に属するデータ $\boldsymbol{x}_j$ が 1 つずつ到着するときの、$\boldsymbol{\mu}_k$ のオンライン更新式が得られた

K-Means アルゴリズムの特徴

- K-Means アルゴリズムの特徴
  - 各データを、たった 1 つのクラスタにのみ割り当てる (ハード割り当て)
  - あるクラスタの中心ベクトル $\boldsymbol{\mu}_k$ に非常に近いデータ点もあれば、複数のクラスタの中間領域にあるようなデータ点も存在する
  - データ $\boldsymbol{x}_i$ が、クラスタ k に属するという結果を得たときに、そのクラスタに属することがほぼ確実なのか、それとも他のクラスタに割り振っても大差はないのかが、区別できない
  - 後者のようなデータ点の場合、単一のクラスタへのハード割り当ては最適でない (不正確) かもしれない
  - データ点を単一のクラスタに割り当てるのではなく、各クラスタに属する確率を、計算できれば良さそう
  - 各クラスタへの割り当ての不明瞭さを反映できる
  - このように曖昧さを含んだ割り当てを、ソフト割り当てという

K-Means アルゴリズムのまとめ

- K-Means アルゴリズムの目的
  - 各データが属するクラスタを決定する (ハード割り当て)

- K-Means アルゴリズムで行っていること
  - 各クラスタに対して、中心となるベクトル $\boldsymbol{\mu}_k$ を考えた
  - $\boldsymbol{\mu}_k$ は、対応するクラスタに属する、全てのデータベクトルの平均として得られた
  - 各データ点は、それと最も距離が近い $\boldsymbol{\mu}_k$ に対応するクラスタ k に割り当てた
  - 各データ点が属するクラスタを、r_{ik} という変数 (1-of-K 符号化法) で表した

K-Means アルゴリズムのまとめ

- ここまでの話の流れ
  - クラスタリングの問題を、目的関数 J の最小化として表現できた
  - 目的関数 J を、$\boldsymbol{\mu}_k$ と r_{ik} の両方について一度に最適化することはできなかった
  - そこで J を、$\boldsymbol{\mu}_k$ と r_{ik} について交互に最適化することを考えた

  - 交互に行う最適化は、後ほど説明する EM アルゴリズムの、E ステップと M ステップに対応していた

  - 逐次版の K-Means 法を、Robbins-Monro 法を使って導出した