パターン認識
Pattern Recognition

佐藤真一
Shin'ichi Satoh

国立情報学研究所
National Institute of Informatics

June 27, 2023

最終レポート

- いわゆるトップジャーナル・トップ国際会議の任意のパターン認識に関する論文についてまとめる
- e.g., IEEE TPAMI, IJCV, CVPR, ICCV, NeurIPS, ICML, ACMMM...
- 以下について記述せよ:
  - 論文の書誌事項
  - 論文の概要
  - どのような問題を扱っているのか、なぜその問題は重要なのか、どのように解決しているのか、どのように評価しているか
  - なぜその論文を選んだのか、何が特に興味深いのか
  - 講義への意見・コメント
- A4 2-4 ページ程度 PDF
- 締め切り: 07/31/2023
- ITC-LMS より提出

Final Report

- Geman and Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, TPAMI, 1984.
- Moghaddam and Pentland, Probabilistic Visual Learning for Object Detection, TPAMI, 1997.
- Belhumeur, Hespanha, and Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, TPAMI, 1997.
- Shi and Malik, Normalized Cuts and Image Segmentation, TPAMI, 2000.
- Belkin and Niyogi, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, NIPS, 2001.
- Zhang and Sim, When Fisher meets Fukunaga-Koontz: A New Look at Linear Discriminants, CVPR, 2006.
- Felzenszwalb, Girshick, McAllester, and Ramanan, Object Detection with Discriminatively Trained Part Based Models, TPAMI, 2009.
- Antonio Torralba and Alexei A. Efros, Unbiased look at dataset bias, CVPR 2011.

本日の内容

クラスタリング技術

- K 平均法 (k-Means)
- 凝集型・階層型クラスタリング (Agglomerative Hierarchical Clustering)
- デンドログラム (Dendrogram)
- クラスタリングの評価

クラスタリング

- ここまで、学習サンプルに対して正解のクラスラベルが与えられていることを想定していた
- しかし、クラスラベルを付与するのは時として困難
- クラスラベルが利用できない場合には何ができるか？
- 教師なし学習、あるいは「知的」前処理手法が考えられる
- クラスタリングは、データ分布に見られる明確な部分クラスを教師なしで発見する手法

クラスタリング

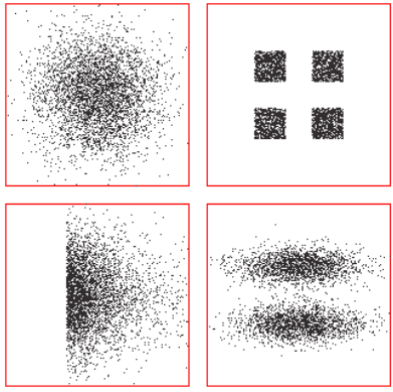


FIGURE 10.6. These four data sets have identical statistics up to second-order—that is, the same mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In such cases it is important to include in the model more parameters to represent the structure more completely. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

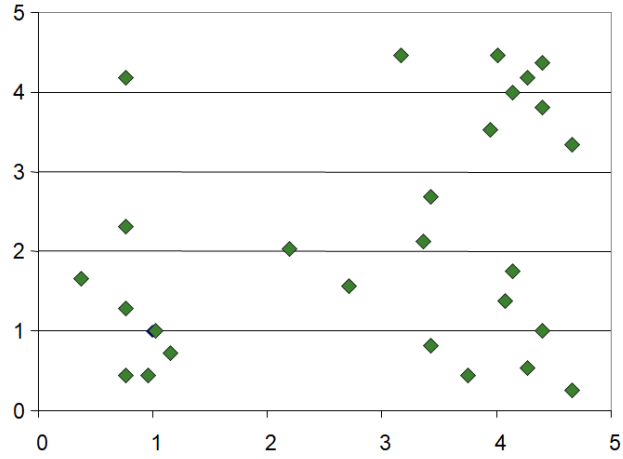
K-Means 法

入力: N 個のデータ
出力: k 個のクラスタ中心

1. クラスタ数 k を決定
2. k 個のクラスタ中心を決定 (ランダムでよい)
3. N 個のオブジェクトを最も近いクラスタ中心の所属に変更
4. 各クラスタごとにクラスタ中心を再計算
5. N 個のオブジェクトの所属に変化がなければ終了
6. さもなければ 3 から繰り返し

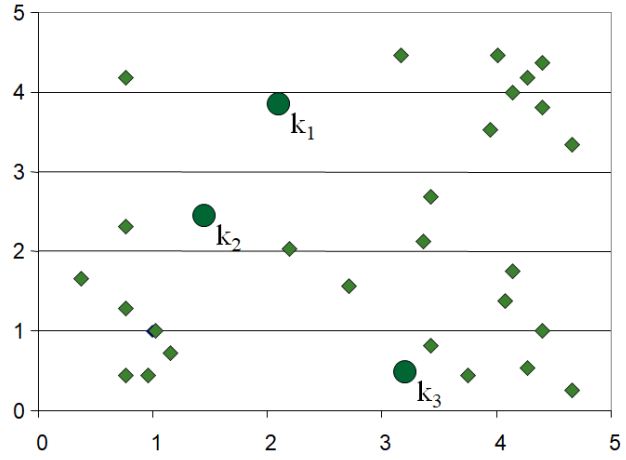
K-Means 法: step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



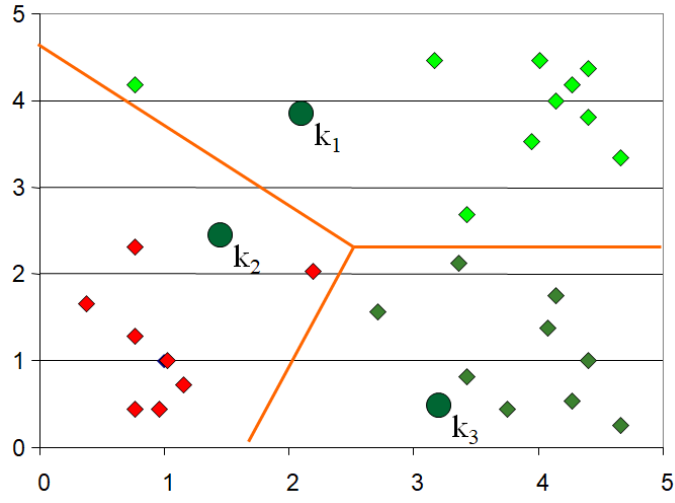
K-Means 法: step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



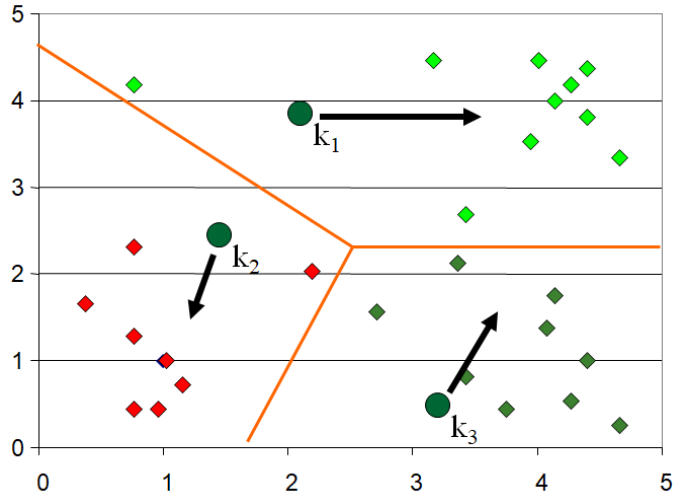
K-Means 法: step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



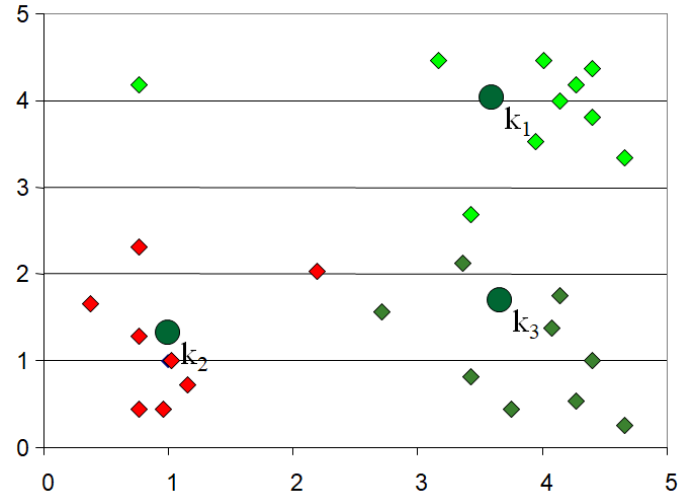
K-Means 法: step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



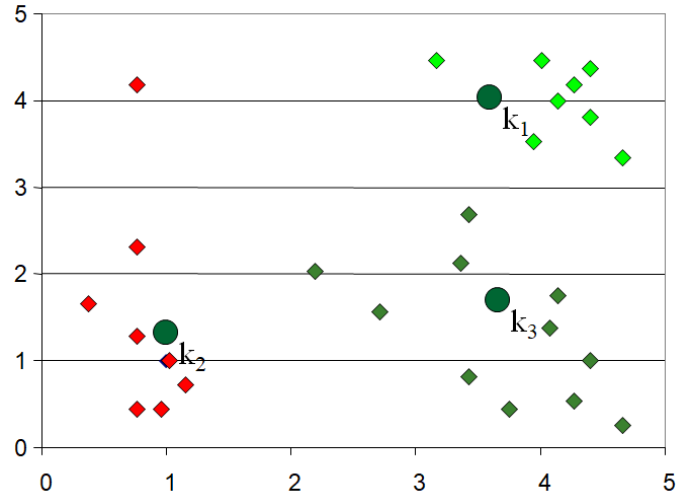
K-Means 法: step 3

Algorithm: k-means, Distance Metric: Euclidean Distance



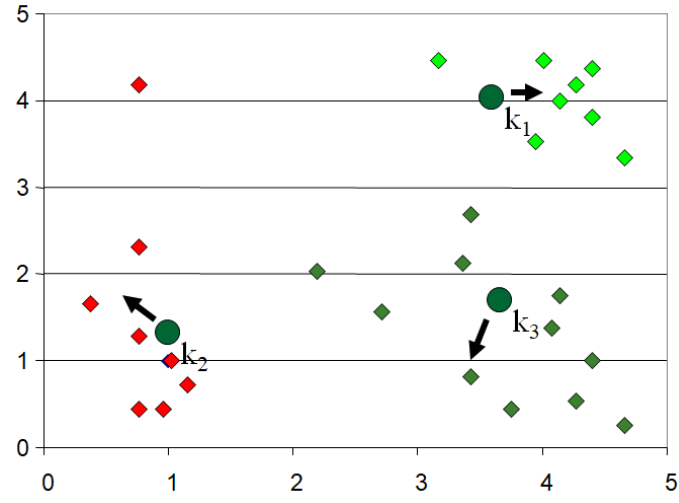
K-Means 法: step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



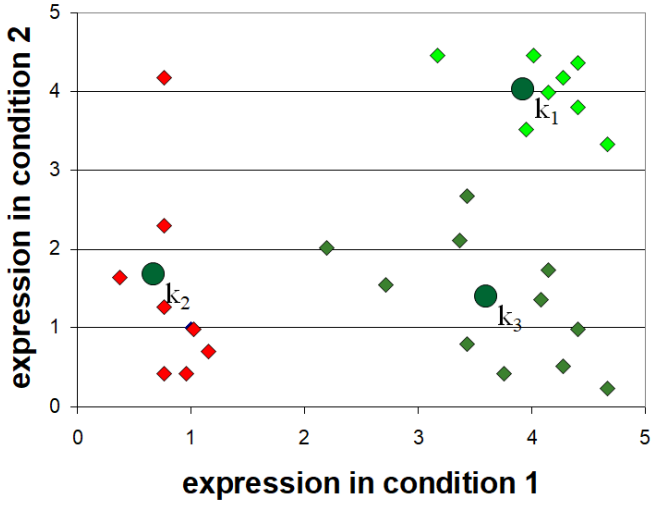
K-Means 法: step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



K-Means 法: step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



階層的クラスタリング

入力: N 個のデータ x_i, $i = 1, \cdots, N$
出力: c 個のクラスタ D_j, $j = 1, \cdots, c$

1. c: 望ましいクラス数に初期化 $c_1 = n$, $D_i = x_i$ for $i = 1, \cdots, n$
2. $c_1 = c_1 - 1$
3. 相互にもっとも近いクラスタの組: D_i と D_j を決定
4. D_i と D_j を併合
5. $c = c_1$ となるまで 2 から繰り返し
6. c 個のクラスタを出力

階層的クラスタリング

デンドログラム

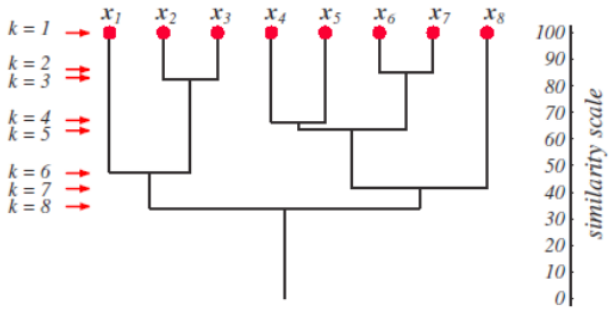


FIGURE 10.11. A dendrogram can represent the results of hierarchical clustering algorithms. The vertical axis shows a generalized measure of similarity among clusters. Here, at level 1 all eight points lie in singleton clusters; each point in a cluster is highly similar to itself, of course. Points $\mathbf{x}_6$ and $\mathbf{x}_7$ happen to be the most similar, and are merged at level 2, and so forth. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

最近傍アルゴリズム

最近傍クラスタリングアルゴリズム

- クラスタ間の距離として、両クラスタの要素間の最小距離を使うとすると、クラスタリング手法は最近傍クラスタリングアルゴリズムという
- クラスタ間距離がある閾値以上となったときにアルゴリズムを停止するようにした場合、これを単連結法 (the single-linkage algorithm) という

最近傍アルゴリズム

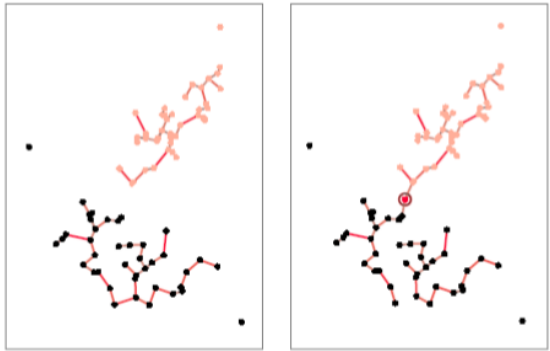


FIGURE 10.13. Two Gaussians were used to generate two-dimensional samples, shown in pink and black. The nearest-neighbor clustering algorithm gives two clusters that well approximate the generating Gaussians (left). If, however, another particular sample is generated (circled red point at the right) and the procedure is restarted, the clusters do not well approximate the Gaussians. This illustrates how the algorithm is sensitive to the details of the samples. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

最長距離アルゴリズム

最長距離クラスタリングアルゴリズム

- クラスタ間の距離として、両クラスタの要素間の最大距離を使うとすると、クラスタリング手法は最長距離クラスタリングアルゴリズムという
- クラスタ間距離がある閾値以上となったときにアルゴリズムを停止するようにした場合、これを完全連結法 (the complete-linkage algorithm) という

最長距離アルゴリズム

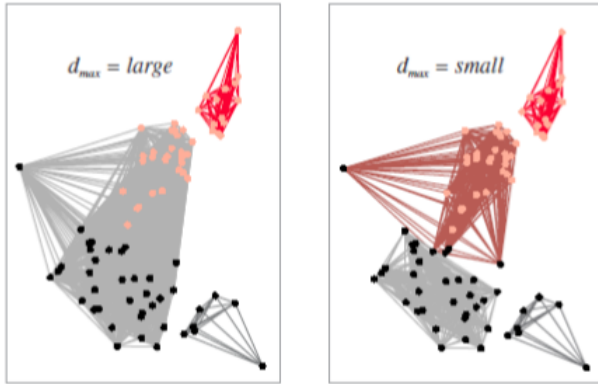


FIGURE 10.14. The farthest-neighbor clustering algorithm uses the separation between the most distant points as a criterion for cluster membership. If this distance is set very large, then all points lie in the same cluster. In the case shown at the left, a fairly large d_{max} leads to three clusters; a smaller d_{max} gives four clusters, as shown at the right. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

クラスタリングの評価

- クラスタリングの性能をどのように評価するか
- $X_1, \cdots, X_r$ を正解クラスタ (クラス), $Y_1, \cdots, Y_s$ を生成されたクラスタとする

純度 (Purity)

$$Purity(X, Y) = \frac{1}{N} \sum_{k} \max_{j} |X_j \cap Y_k|$$

- 純度 (Purity) は、全クラスタの純度の総和で算出される
- 問題点: クラスタ数を多くすれば高純度 (high purity) は簡単に達成可能
- 極端には、各データを独立したクラスタにすればよい

正規化相互情報量 (Normalized Mutual Information: NMI)

$$
\begin{aligned}
NMI(X, Y) &= \frac{2I(X; Y)}{H(X) + H(Y)} \\
I(X; Y) &= \sum_k \sum_j P(X_j \cap Y_k) \log \frac{P(X_j \cap Y_k)}{P(X_j)P(Y_k)} \\
&= \sum_k \sum_j \frac{|X_j \cap Y_k|}{N} \log \frac{N\,|X_j \cap Y_k|}{|X_j|\,|Y_k|} \\
H(X) &= -\sum_k P(X_k) \log P(X_k) \\
&= -\sum_k \frac{|X_k|}{N} \log \frac{|X_k|}{N}
\end{aligned}
$$

ランド指数 (Rand Index)

- 各アイテム対 (p, q) につき
- TP: X_i にも Y_j にも所属する対の数
- TN: X_i にも Y_j にも所属しない対の数
- FP: X_i には所属しないが Y_j には所属する対の数
- FN: X_i には所属するが Y_j には所属しない対の数

$$
\begin{aligned}
RI(X, Y) &= \frac{TP + TN}{TP + FP + FN + TN} \\
&= \frac{TP + TN}{N(N-1)/2}
\end{aligned}
$$

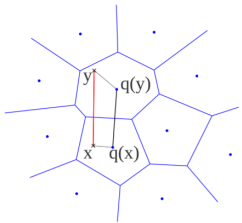
F 値 (F measure)

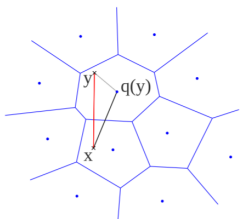
$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

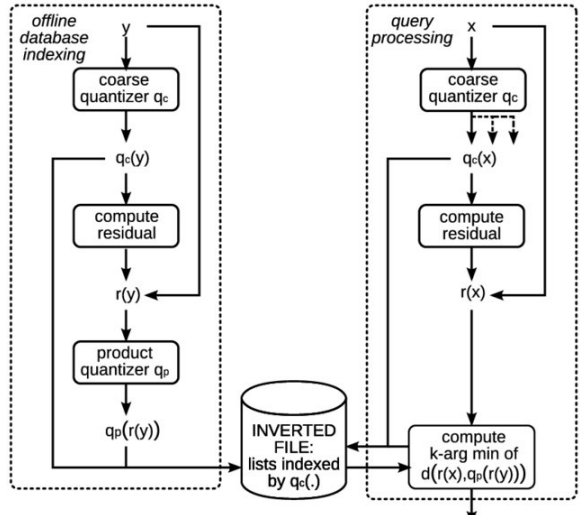
クラスタリング応用: 高速多次元データ検索

- 大量の高次元ベクトル (y) をデータベースに蓄えたい
- 任意の問い合わせベクトル x を受け付けて、データベース中のベクトルのうち問い合わせに近いものを高速に検索したい
- 基本アイデア
  - クラスタリングによりデータベース中のベクトルを大まかに分類しておく
  - 各クラスタ中のデータを、クラスタ中心などを見出し語として索引に登録しておく
  - (オプション) クラスタ中心からの差分を求めておいて、これもクラスタリングにより索引化しておく
  - 検索時には、与えられた x に最も近いクラスタ中心に対応する見出し語に対応するデータのみ調査する
  - (オプション) 調査の際、クラスタ中心からの差分でさらに索引を引いて相当するデータのみ調査する

クラスタリング応用: 高速多次元データ検索







Jégou, Douze, Schmid, Searching with quantization: approximate nearest neighbor search using short codes and distance estimators, INRIA, 2009