目次

★14　パターン認識と機械学習 (5) — 教師なし学習

★14.1　教師なし学習とは

教師あり学習では，個々の学習データが「入力」と「それに対する出力の正解」のペアとして与えられていた．一方，**教師なし学習** (☆ 1) では，学習データとして「入力」のみが与えられる．教師なし学習の目的は，大量のデータが与えられたときに，そのデータのもつ規則性や構造を見い出し，有益な情報を抽出する処理を自動的に行うことである．**データ分析** (☆ 2) の手法の一種といえる．ここでは，様々な教師なし学習の問題設定のうち，**クラスタリング** (☆ 3) と**次元圧縮** (☆ 4) を取り上げる．

クラスタリングとは，大量のデータが与えられた時に，それらをいくつかの「塊」（クラスタ）に分類する手法である．データが D 次元実ベクトルの場合を考えると，N 個の D 次元ベクトルから成るデータ集合

$$\left\{\boldsymbol{x}_n \in \mathcal{R}^D | n = 1, 2, \ldots, N\right\} \tag{1}$$

が与えられた時に，これを $K \ll N$ 通りのクラスタに分類する．教師ありの識別問題と異なり，どのデータをどこに分類するかの正解は与えられないことに注意しよう．クラスタリングを行うと，大量のデータを要約したり，構造を見つけ出したりすることができる (☆ 5)．

一方，次元圧縮は，「データの数を減らす」かわりに「データの次元数を減らす」手法である．例えば，式 (1) のデータ集合の個々の要素を，その本質的な特徴をなるべく保ったまま，より低次元のベクトル $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N \in \mathcal{R}^H$ $(H \leq D)$ に変換する．高次元の（つまりたくさんの変数から成る）データを，その性質をよく表す低次元の（少数の変数から成る）データに変換することで，データを分析しやすくしたり (☆ 6)，記憶容量や後の処理の計算量を減らしたりすることができる (☆ 7)．

☆ 1) 教師なし学習: unsupervised learning.

☆ 2) データ分析: data analysis.

☆ 3) クラスタリング: clustering

☆ 4) 次元圧縮: dimensionality reduction または dimension reduction. 次元削減とも．

☆ 5) 例えば，遺伝子データをクラスタリングして解析することで，生物種の間の類縁関係を推定（種 A と種 B は遺伝子配列が似てるので同じ種から進化した可能性が高い，とか）したりできる．

☆ 6) 例えば，20 科目の点数データ（20 次元）を 3 つの変数（3 次元）に変換して，学生の成績傾向を把握・分析しやすくするとか．

☆ 7) 以前登場した離散コサイン変換 (DCT) も，次元圧縮の用途に使える．

★14.2 K 平均法によるクラスタリング

クラスタリングの手法には様々なものがあるが，ここでは代表的なアルゴリズムとして，K **平均法** (☆ 8) を紹介する．K 平均法は，2 つのデータ $\boldsymbol{x}$ と $\boldsymbol{y}$ の間の非類似度 (☆ 9) を両者間のユークリッド距離 (☆ 10) $\|\boldsymbol{x}-\boldsymbol{y}\|$ で表せるようなデータに対して，それらを K 個のクラスタに分ける分け方を見つけるための教師なし学習のアルゴリズムである．クラスタ数 K はあらかじめ適当な方法で決めておかねばならない．K 平均法の手順は次のようなものである．

1. 各学習データ $\boldsymbol{x}_n(n=1,2,\ldots,N)$ をランダムにクラスタ C_1 から C_K までの K 個のクラスタのいずれかに割り当てる (☆ 12).
2. クラスタ $C_k(k=1,2,\ldots,K)$ に割り当てられたデータの平均 $\boldsymbol{c}_k$ を求める．

$$\boldsymbol{c}_k = \frac{1}{|C_k|}\sum_{n:\boldsymbol{x}_n\in C_k}\boldsymbol{x}_n \tag{2}$$

これをクラスタ C_k の**セントロイド** (☆ 11) という．ここで，$|C_k|$ は集合 C_k の要素数，すなわちクラスタ C_k に割り当てられた学習データの数を表す．

3. 各学習データ $\boldsymbol{x}_n(n=1,2,\ldots,N)$ を，セントロイドとの距離が最小となるクラスタに割り当て直す．例えば $\boldsymbol{x}_n$ に対して

$$k^* = \underset{k=1,2,\ldots,K}{\operatorname{argmin}} \|\boldsymbol{x}_n-\boldsymbol{c}_k\|^2 \tag{3}$$

であれば (☆ 13) (☆ 14)，$\boldsymbol{x}_n$ はクラスタ C_{k^*} に割り当てる．

4. 上のステップの結果があらかじめ定めておいた条件を満たしている（後述）ならば終了，さもなくば 2. へ戻る．

この学習の終了条件としては，「クラスタ割り当てに変化がなくなった」，「ステップ 2,3 の実行回数が一定に達した」などが用いられる．また，K 平均法では次式の E の値 (これはクラスタリングの「誤差」を表している) が学習ステップ毎に単調減少することが知られているので，この値の減少幅が一定より小さくなったら終了する，という方法もよく用いられる．

$$E = \sum_{k=1}^{k}\sum_{n:\boldsymbol{x}_n\in C_k}\|\boldsymbol{x}_n-\boldsymbol{c}_k\|^2 \tag{4}$$

K 平均法の結果は，学習データに対するクラスタ割り当ての初期値に依存する．そのため，実際には初期値を変えて何度か K 平均法を実行し，上記の E の値が最も小さかった結果を採用する，というような方法がとられる．

上記の学習手続きによってクラスタセントロイドが推定できたら，式 (3) と同様の計算によって未知データの所属クラスタも決めることができる．例えば，ある未知データ $\boldsymbol{x}$ について，

$$i = \underset{k=1,2,\ldots,K}{\operatorname{argmin}} \|\boldsymbol{x}-\boldsymbol{c}_k\|^2 \tag{5}$$

であれば，このデータの所属はクラスタ C_i とすればよい．

図 1 に，2 次元のデータに対して K 平均法を適用した結果を示す．また，図 2 に，猫の顔画像 131 枚 (画素数は 64×64) に K 平均法を適用して得られたセントロイドすなわちクラスタ毎の平均画像を示す．

☆ 8) K 平均法: K-means 法とも．

☆ 9) 距離が小さい方が類似度が大きいので，「非」類似度が距離に対応すると考えている．

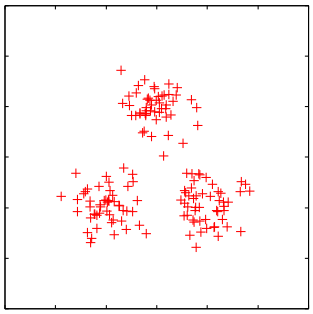
☆ 10) $\boldsymbol{x}$ が D 次元ベクトル $\boldsymbol{x}=(x_1,x_2,\ldots,x_D)$ であり，$\boldsymbol{y}$ も同様の場合，$\|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{\sum_{d=1}^{D}(x_d-y_d)^2}$

☆ 11) セントロイド: centroid.

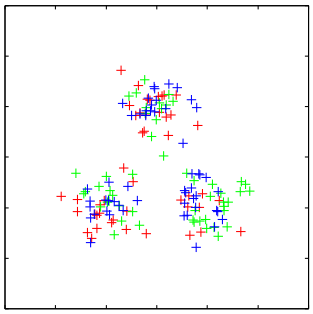
☆ 12) 後述のように K 平均法の結果は初期値に依存するので，実用的には初期値の選び方を工夫したアルゴリズムが用いられることが多い．

☆ 13) argmin の意味は次の例でわかるだろう．
$f(x)=(x-2)^2+3$ のとき，
$\min_x f(x)=3$,
$\operatorname{argmin}_x f(x)=2$.

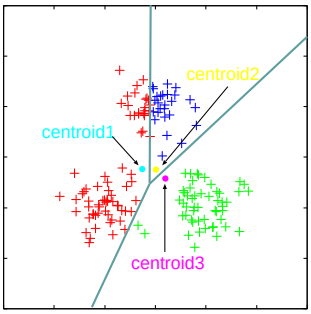
☆ 14) ここでは距離の大小関係のみが問題なので，ユークリッド距離そのものではなく二乗した値で考えている（実際の計算では平方根が出てこない分その方が簡単だから）．



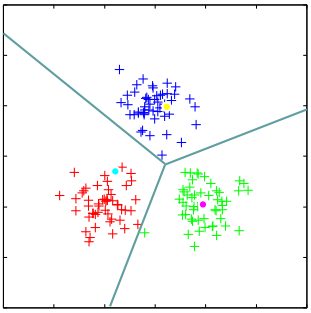
(1A) 学習データその 1



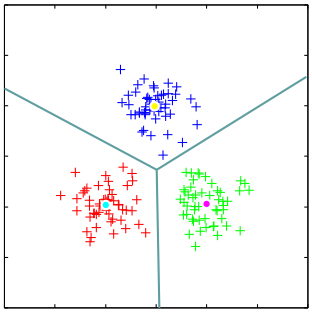
(1B) 初期値



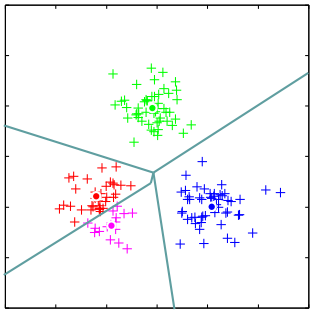
(1C) 学習 1 回



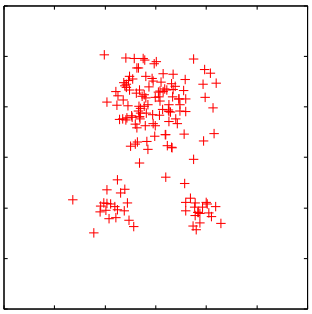
(1D) 学習 2 回



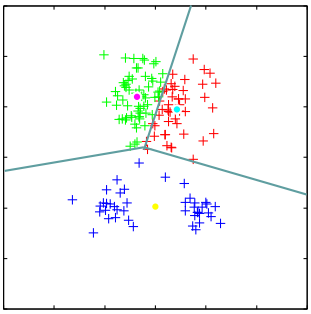
(1E) 学習 10 回



(2) $K = 4$ で学習 10 回



(3A) 学習データその 2



(3B) 学習 10 回

図 1: K 平均法による 2 次元データのクラスタリング．(1A) は学習データの例．(1B) から (1E) までは，(1A) のデータ対して $K = 3$ で K 平均法を適用した結果．(2) は，同じデータに $K = 4$ で K 平均法を適用した結果．(3A) と (3B) は，別の学習データとそのクラスタリング結果 ($K = 3$)．

図 2: 131 枚の猫画像に K 平均法を適用して得られたクラスタ平均画像 ($K = 5$)．

★14.3 主成分分析と次元圧縮

次元圧縮の手法にも様々なものがあるが，ここでは，統計的なデータ分析手法である**主成分分析** (☆ 15) に基づく手法を紹介する．主成分分析の目的は，式 (1) のデータから複数の互いに「無関係」な因子を取り出し，個々のデータをこれらの因子の線形結合で表すことである．

☆ 15) 主成分分析: Principal Component Analysis.

簡単のため，以下では因子を一つだけ取り出す場合について説明する．この場合，主成分分析の問題設定は次のようになる：「式 (1) のデータ $\boldsymbol{x}_n$ をベクトル $\boldsymbol{a}$ を用いて次式のようにスカラ y_n に変換するとき，$\{y_n\}$ の分散が最大となるように $\boldsymbol{a}$ を定めよ．ただし，$\bar{\boldsymbol{x}}$ はデータの平均すなわち $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$ である．」

$$y_n = \boldsymbol{a}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \tag{6}$$

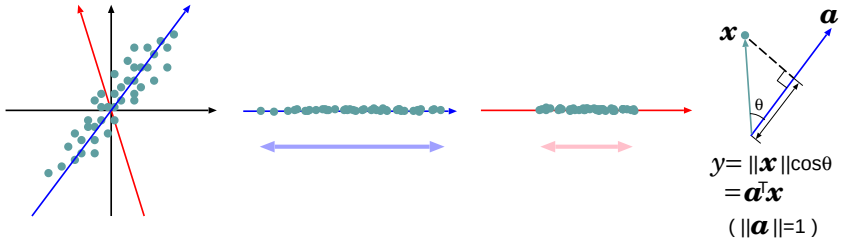


図 3: 左端の図に灰色で示された 2 次元データを青い軸方向の成分のみで説明しようとすると，その成分は左から 2 番目の図のように散らばる．赤い軸方向の成分のみだと，左から 3 番目の図のように散らばる．

この問題の解 $\boldsymbol{a}$ がどのようなベクトルになるか考えよう．まず，式 (6) より，y_n の平均は 0 である (☆ 16)．したがって，$\{y_n\}$ の分散は $\frac{1}{N}\sum_{n=1}^{N} y_n^2$ と書け，

☆ 16) $\sum_{n=1}^{N} y_n$
$= \sum_{n=1}^{N} \boldsymbol{a}^\top \boldsymbol{x}_n - \sum_{n=1}^{N} \boldsymbol{a}^\top \bar{\boldsymbol{x}}$
$= \boldsymbol{a}^\top \sum_n \boldsymbol{x}_n - \boldsymbol{a}^\top \sum_n \bar{\boldsymbol{x}}$
$= N\boldsymbol{a}^\top \bar{\boldsymbol{x}} - N\boldsymbol{a}^\top \bar{\boldsymbol{x}} = 0$

$$\frac{1}{N}\sum_{n=1}^{N} y_n^2 = \frac{1}{N}\sum_{n=1}^{N} \left(\boldsymbol{a}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\right)\left(\boldsymbol{a}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\right)^\top \tag{7}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{a}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top \boldsymbol{a} \tag{8}$$

$$= \boldsymbol{a}^\top \left(\frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top\right) \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a} \tag{9}$$

と変形できる（$\boldsymbol{V}$ はデータの**分散共分散行列**である）．すなわち，求めるべきは，$\boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}$ を最大にするベクトル $\boldsymbol{a}$ である．ただし，この値は $\boldsymbol{a}$ の長さを大きくすればいくらでも大きくできるので，$\|\boldsymbol{a}\|$ に条件を付けないと意味がない．そこで，$\|\boldsymbol{a}\| = 1$ という制約条件のもとで $\boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}$ を最大化する $\boldsymbol{a}$ を求めよう．

上記のような制約条件付きの最適化問題を解く定番の手法は，ラグランジュの未定乗数法である．この問題の場合，ラグランジュ乗数を λ として

$$L = \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a} - \lambda(\boldsymbol{a}^\top \boldsymbol{a} - 1) \tag{10}$$

とおけば，$\frac{\partial L}{\partial \boldsymbol{a}} = \boldsymbol{0}$ を満たす $\boldsymbol{a}$ が解の候補となる．ここで，

$$\frac{\partial L}{\partial \boldsymbol{a}} = 2\boldsymbol{V}\boldsymbol{a} - 2\lambda\boldsymbol{a} \tag{11}$$

なので (☆ 17)，解は

☆ 17) $\frac{\partial L}{\partial \boldsymbol{a}}$ は，「L を $\boldsymbol{a}$ の各要素で微分したものをならべたベクトル」である．要素ごとの微分を考えると，$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{x} = 2\boldsymbol{x}$ 等の公式を導出できる．

$$\boldsymbol{V}\boldsymbol{a} = \lambda \boldsymbol{a} \tag{12}$$

を満たさねばならない．すなわち，解候補は行列 $\boldsymbol{V}$ の単位固有ベクトルである．さらに，式 (12) を式 (10) に代入すると

$$L = \lambda \boldsymbol{a}^\top \boldsymbol{a} - \lambda \cdot 0 = \lambda \tag{13}$$

となることから，この問題の解は，$\boldsymbol{V}$ の最大固有値に対する単位固有ベクトルであることがわかる (☆ 18)．

ここではこれ以上深入りしないが，複数の因子を取り出す場合についても同様の議論ができる．その場合，$\boldsymbol{V}$ の固有ベクトルを対応する固有値の大きい方から順に選んで，それらを因子として採用すればよいことがわかる．

以下に示すのは，猫の顔画像に主成分分析を適用した例である（第 4 回講義資料と同じもの）．ここで「固有顔」と呼んでいるものは，猫の顔画像の学習データから求めた分散共分散行列の固有ベクトルである．

☆ 18) $\boldsymbol{V}$ は半正値対称行列なので固有値は必ず 0 以上の実数になる（詳細は省略）．

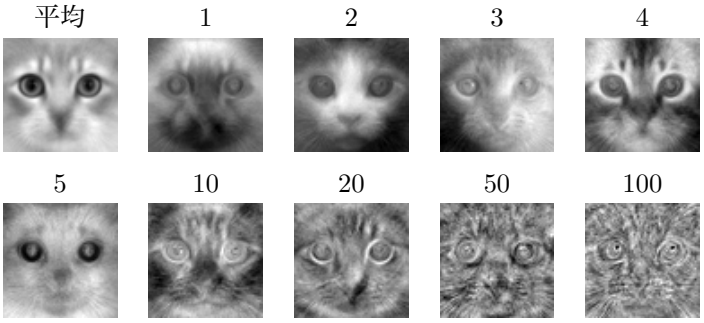


図 1: 猫の固有顔の例．左上の画像は多数の猫の顔画像の平均であり，$1, 2, \ldots$ と記された画像は固有顔を可視化したもの（番号の小さいものほど対応する固有値が大きく，「重要」な基底）．これらの固有顔は正規直交基底を成している．

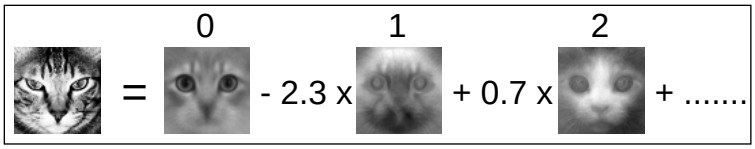


図 2: 固有顔を用いた猫の顔の展開の概念図．画像の上の番号は，固有顔の番号に対応している（0 と記された画像は平均）．

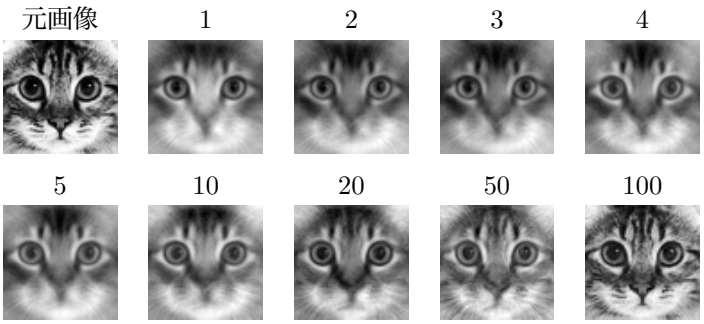


図 3: 固有顔を用いた近似．左上は 4096 個の画素値から成る元画像．それ以外は，画像の上に記された数の基底のみを用いて元画像を近似したもの．例えば，5 と記された画像は，上記の 1 番から 5 番までの基底のみを用いて元画像を近似しており，5 つの基底に対応する 5 つの展開係数の値のみでこの画像の情報を表現していることになる．

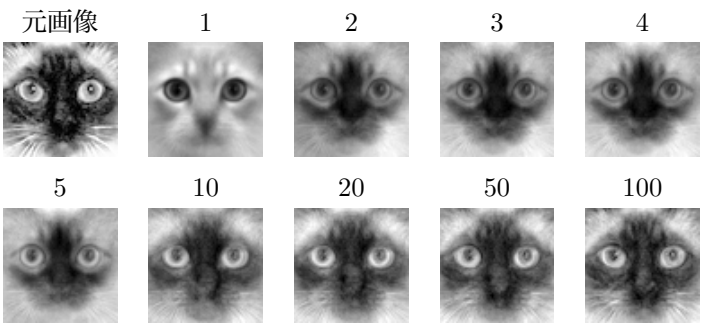


図 4: もうひとつの例．