人工知能A

Topic 6: Introduction to clustering and PCA

クラスタリングと主成分分析（の入門）

6

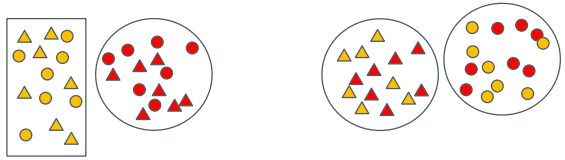
クラスタリング

- 講義の内容
  - 概念学習とクラスタリング
  - K-means法
  - 階層的クラスタリング
  - クラスタリングの応用例
  - 主成分分析
  - 混合（ガウス）分布（概要−＞最適化の講義）
- 目標：
  - クラスタリングの基本、主にk-means法と凝集法について学ぶ。主成分分析（次元圧縮）について学ぶ

3 概念学習とクラスタリング

クラスタリング

- クラスタリング(clustering) とは、データの集合を、その表現（通常は数値のベクトル）の類似性に従って分類すること。クラスタ分析 (cluster analysis) とも言う。
  - 分類した一つ一つのまとまりをクラスタ (cluster) という。
  - 特に正解データや報酬があるわけではなく、データだけから分類するので、教師なし学習の一つ。（どのように分類すべきかは未知）
  - 分類したものはまとめて概念として保存すると有効なこともあり、概念学習の一つでもある。
  - その分類の観点（あるいは意味）は複数有り、何が適切かは問題に応じて異なる



特徴の表現

- あるデータ（対象）を区別するには、その特徴を数値化して、表現する。ただし、特徴は多様であるので、複数の数値（ベクトル）として表現。
  - これを特徴ベクトル (feature vector) あるいは特徴量(feature value)と言う。
  - 各特徴を表す数値（ベクトルの要素）の範囲は、実数全体、整数、正数、[0, 1]区間などそれぞれによって決まる。表現可能な特徴量の空間を特徴空間 (feature space) と言う。
  - たとえば、みかんであれば、形、色、上に緑の点がある、などなど多くの特徴があり、これで（完全ではないが）表現できる。
  - クラスタリングは、ベクトル空間（の部分集合）の点を、適当な軸で分割・分類することとも言える。

6 基本的なクラスタリング法

クラスタリングの基本 (K-means法)

- K個の距離的に近いグループに分けること。
  - アルゴリズム：データの集合をΩとする。
  1. ここからランダムにK個の点を取り、それを$c_1, \ldots, c_K$とする。
  2. 全ての$d \in \Omega$についてc_iとの距離$\| d - c_i \|^2$を求め、一番近いc_iのインデックスiをdに付与する。これを$I(d) \in \{1, \ldots, K\}$と表す。
  3. すべてのインデックスiについて、そのデータの集合の重心（平均）を新たなc_iとする。
  4. 2.に戻る。この繰り返しを、インデックスの付け替えが無くなるまで行う。
  - 上記は以下の評価値 $J(c_1, \ldots, c_K)$ を最小化する$\forall c_j$と(Ω_i)を求める。

$$J(c_1, \cdots, c_K) = \sum_{i=1}^{K} \sum_{d \in \Omega_i} \| d - c_{I(d)} \|^2$$

ただし、 $\Omega_i = \{d \in \Omega : I(d) = i\}$である。

K-means法の特徴（利点と欠点）

- 利点
  - 単純で効率がよい。分かりやすい。
  - 多次元空間における外れ値（異常値）を容易に見つけることができる。
- 欠点
  - 複数の相関関係の影響を受ける。また高次元では固まりやすい（分割・分類しにくくなる）。
  - 基本的に距離に基づいているので、超球の形状のみに制限される。
  - 結果は初期値（$c_1, \ldots, c_K$の取り方）に依存し、安定的ではない。何回か繰返し良いものを選ぶ（Jを最小化）。

階層的クラスタリング（凝集型）

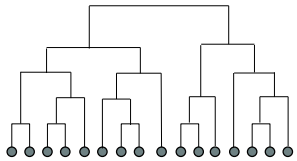
- K-meansとは逆に、個々の点から近いものを集めてクラスタリングする方法。
- 徐々に統合により大きくなるため階層性がある。
  - 凝集型クラスタリング (agglomerative clustering)あるいは凝集型階層クラスタリング (agglomerative hierarchical clustering) と言う
- アルゴリズム： $\Omega=\{d_1, \ldots, d_n\}$をデータの集合とする
  - 初期値として$\forall i$に対し、$c_i=\{d_i\}$とおく。 c_iはクラスタ。
  - 各クラスタ間で距離$dist(c_i, c_j)$を測り、最小となるペアを統合する。
    $$(c_{i_0}c_{j_0}) = arg \min_{(c_i,c_j)\in\Omega\times\Omega, i\neq j} dist(c_i, c_j)$$
  - 上記をクラスタの数がK（あるいは1）になるまで繰り返す。
- 議論： c_iは集合である。その距離の定義の仕方は？
  - Centroid法、 ward法などがあり、よく使われる。

凝集型階層クラスタリングの種類

- 最短距離法あるいは単連結法 (single-link method)
  - 二つの集合のうち、最も近い点同士の距離を集合の距離とする。
  - 同様に最長距離法 (=完全連結法 , complete-link method) もある。
- 群平均法 (group-average method)
  - 二つの集合の点のペアの距離の平均値を距離とする。
- 重心法（Centroid method）
  - それぞれの集合の重心を求めて、それらの距離とする。
- Ward 法 (Ward's method)
  - 集合Cの重心bとCの各点からの距離の2乗和を$L(C)$と置く。このとき、$dist(c_i, c_j) = L(c_i \cup c_j) - L(c_i) - L(c_j)$と定義する。
  - $L(C)$は、Cの全ての点ではなく、サンプルを利用する方法もある。

デンドログラム

- 集約すると右下のグラフのように階層的な木構造が生まれる。
  - このグラフをデンドログラム (dendrogram)という。



- 課題６－１（発展）
  この他に潜在的ディリクレ配分法
  LDA (Latent Dirichlet allocation)
  などもあるので、興味がある学生は調べてみるとよいでしょう
  （特に言語解析やトピック（話題）モデルに興味があれば）。

12 適用例

例：画像圧縮

- 画像は多数の点でできていて、各点は色 (RGBなど) で表されている
  - たとえば、RGBのそれぞれの色の強度を8ビット (0-255) の256段階で表現すると24ビット分の情報が必要。各点の色をこの3 次元ベクトルで表現。
  - たとえば、ある写真について、上記のベクトル空間にK=32としてK-means法を適用し、その重心を$c_1, \ldots, c_K$をとして、同一クラスタに属するものをすべて一色のc_iで描く(c_i にも色が対応する）。
  - 使われる色は、32色に。近似している色に変わる。これにより画像情報がかなり下がる。
    - 問題点：３次元空間での色の近さと人間の認知はかならずしも一致しない。
    - 問題点：人間はたとえば顔など自然に注目する箇所があり、感覚的に不自然に感じる。
    - 画像圧縮については、いろいろな研究があり、それらは専門の講義で。

例：言語処理

- 簡単な例としてbag-of-words (BOW, bagは多重集合のことでもある)
  - 単語 (全単語でもよいが普通は着目する単語。それでも数万以上はある) の出現数をベクトル表示。たとえば、
    “全単語でもよいが普通は着目する単語” ⇒
    全/単語/でも/よい/が/普通/は/着目/する/単語/
  - もし単語、普通、集合、出現、着目、数 だけを考えれば、
    [2, 1, 0, 0, 1 0]
    これを文や文書ごとに集計し、ベクトル表示。
  - 文章を数百から数万次元の空間に配置し、クラスタリングすると、文章を分類できる。
  - この他にも、関連する項目として、連続BOW (continuous bag-of-words, N-gram, word2vec等があるが、これらは自然言語処理の講義で述べられるはず)

マーケッティング（顧客分類）

- $\Omega=\{x_1, \ldots, x_n\}$を顧客の集合とする。ネット販売やポイント(たとえばTポイントなど) などで購入実績が分かる。
  - x_iの履歴（たとえば最近1年とか1月とか）を調査し、以下のベクトルで表現。
    - v_1: 購入総額、v_2: 雑貨の購入額、v_3: 食品の購入額、v_4: 電気製品の購入額
      v_5: 1月の購入額、v_6: 2月の購入額、v_7: 3月の購入額、v_8: 4月の購入額
      などなど。力まかせで、とにかくいろいろ調べる。
    - 重みを均一にしたいなら正規化（例えば0-10の範囲に縮小・拡大）する（たとえば、購入総額は大きいので、その強い影響を消し、バランスをとる。もちろん、解析者の意図もある）。
  - この表現を用いて顧客を分類し、クラスタ内の顧客は同じ傾向があると考える。
    - たとえば、購入の多い月の直前にメールを送る、
    - 類似した購入指向があれば、他の顧客の購入活動から購入推薦を行う、など

16 数学の復習

主成分分析 (PCA)の準備として

数学の復習(内積)

- n次元ベクトル空間のベクトル$\boldsymbol{a} = (a_1, \ldots, a_n), \boldsymbol{b} = (b_1, \ldots, b_n)$の内積は、

$$a \cdot b = a_1 b_1 + \cdots + a_n b_n$$

と定義。これは内積の公式から、

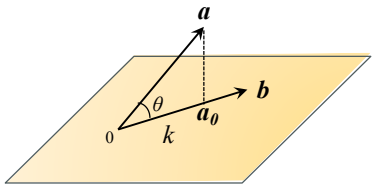
$$a \cdot b = |a||b|\cos\theta$$

aのbへの射影（垂線を下ろしたときの）の長さ（あるいはaを点と考えて、そのbへ垂線を下ろした点a_0の原点からの長さ）をkとすると、$k = |a|\cos\theta$。したがって、

$$k = \frac{a \cdot b}{|b|}$$

である。したがって、

$$a_0 = k\frac{b}{|b|} = \frac{a \cdot b}{|b|^2} b$$



- 特にbが単位ベクトルなら、 $k = a \cdot b$、 $a_0 = kb$である。

数学の復習 (不偏推定量)

- n個の観測データ$X_1, \ldots, X_n$から、母集団の平均と分散を推定する不偏推定量はそれぞれ、

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

と、

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

である。初めの係数を$1/n$とすると、$1/n$だけ期待値が小さくなり、ややずれる（勿論nが大きければ差は小さい）。

- n個の観測データ$X_1, \ldots, X_n$と$Y_1, \ldots, Y_n$から共分散を推定する不偏推定量は、

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

である。

数学の復習（ラグランジュの未定乗数法）

- ラグランジュの未定乗数法 (Lagrange multiplier method)
  - $f(x_1, \ldots, x_n)$, $g(x_1, \ldots, x_n)$は関数。 $\boldsymbol{x}=(x_1, \ldots, x_n)$ とおく。条件 $g(x_1, \ldots, x_n)=0$のもと、$f(x_1, \ldots, x_n)$を最大化（最小化）する解 $\boldsymbol{c}=(c_1, \ldots, c_n)$ は、
    $L(x_1, \ldots, x_n, \lambda)= f(x_1, \ldots, x_n) -\lambda g(x_1, \ldots, x_n)$としたとき、

$$\frac{\partial L(\boldsymbol{x},\lambda)}{\partial x_1} = \cdots = \frac{\partial L(\boldsymbol{x},\lambda)}{\partial x_n} = \frac{\partial L(\boldsymbol{x},\lambda)}{\partial \lambda} = 0 \quad (1)$$

の解である（つまりL の停留点）。 λをラグランジュ乗数という

  - Maximize (minimize) $f(x_1, \ldots, x_n)$
    subject to $g(x_1, \ldots, x_n)=0$
  - $f(\boldsymbol{x})$を (あるいは$g(\boldsymbol{x})=0$なので$L(\boldsymbol{x}, \lambda)$ を) を最大化する$\boldsymbol{c}$を求めればよい。

数学の復習（ラグランジュの未定乗数法）

- 証明（幾何学的な概要。「最適化理論」等の講義で詳しくやると思う）

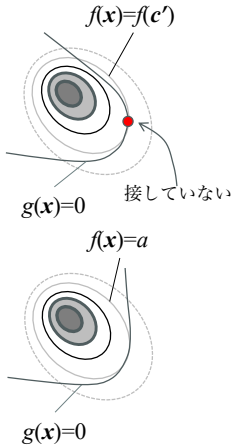
1. $\frac{\partial L(\boldsymbol{x},\lambda)}{\partial \lambda} = 0$は、 $g(x_1, \ldots, x_n)=0$を意味する。
2. $\lambda \neq 0$として、残りの式は$\frac{\partial f(x)}{\partial x_i} = \lambda \frac{\partial g(x)}{\partial x_i}$ (i=1, …, n)となる。
3. これをまとめて書くと

$$\begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = \lambda \begin{pmatrix} \frac{\partial g\ (x)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x)}{\partial x_n} \end{pmatrix}$$

$\lambda \neq 0$なので、これらのベクトルは平行。つまり$f(\boldsymbol{x})$, $g(\boldsymbol{x})$を偏微分したベクトルは勾配（つまり曲線の法線ベクトル（前出: Topic2, スライド113で）であるが、これが同じ向きであることを示している。

数学の復習（ラグランジュの未定乗数法）

4. これは、$f(x_1, \ldots, x_n)=\exists a$ と$g(x_1, \ldots, x_n)=0$ が接していることを意味する。
5. 他方、勾配が同じではない点$\boldsymbol{c}'=(c'_1, \ldots, c'_n)$があり、それが元の制約を満たす解（最大値）とする。
6. その点の近傍では、$f(\boldsymbol{x})=f(\boldsymbol{c}')$と$g_j(\boldsymbol{x})=0$は（接していないので）交差し、実数の連続性から、$f(\boldsymbol{x})>f(\boldsymbol{c}')$かつ$g(\boldsymbol{x})=0$となる点が存在する。これは、$\boldsymbol{c}'$が制約を満たす解であることと矛盾。（上の図のようなことはない）
7. 従って、式(1)を満たす解で、fの値が最大となるもの$\boldsymbol{c}=(c_1, \ldots, c_n)$を選べばよい。



数学の復習（ラグランジュの未定乗数法）

- ラグランジュの未定乗数法のやや一般化
- $f(x_1, \ldots, x_n)$, $g_k(x_1, \ldots, x_n)$は関数 $(1 \leqq k \leqq m)$。 $\boldsymbol{x}=(x_1, \ldots, x_n)$ とおく。条件 $g_k(x_1, \ldots, x_n)=0$のもと、$f(x_1, \ldots, x_n)$を最大化（最小化）する解 $\boldsymbol{c}=(c_1, \ldots, c_n)$ は、

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = f(x_1, \ldots, x_n) - \sum_{k=1}^{m} \lambda_k g_k(x_1, \ldots, x_n)$$

としたとき、

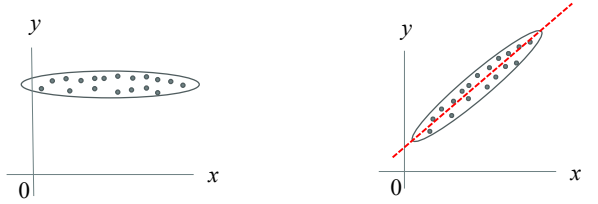
$$\frac{\partial L(\boldsymbol{x}, \boldsymbol{\lambda})}{\partial x_1} = \cdots = \frac{\partial L(\boldsymbol{x}, \boldsymbol{\lambda})}{\partial x_n} = \frac{\partial L(\boldsymbol{x}, \boldsymbol{\lambda})}{\partial \lambda_1} = \cdots = \frac{\partial L(\boldsymbol{x}, \boldsymbol{\lambda})}{\partial \lambda_m} = 0 \qquad (2)$$

の解（つまり停留点）でもある。

23 主成分分析 (PCA)

主成分分析

- これまでの例では次元はとても高そうである。一方で、使われない変数もありそう。
- 仮に使われている変数でも、データを区分する（クラスタリングする）のに重要（不要）な変数はありそうだ。
  - イメージ：たとえば、左下では、x軸の値の方が重要そうだ。右の方は、x,y軸共に重要そうだが、新しく軸を赤線（点線）のようにとれば、一つの軸で区別がある程度できそうである。

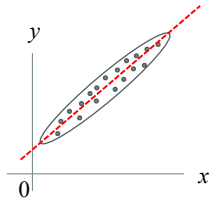


主成分分析

- n変数 (n次元)のデータがN個ある。これを少ない次元mでなるべくデータの区分が可能なように表現したい。ただし、$m \leq n$である。
  - このように少ない変数で表現し直すことを低次元化という。
  - （必要であれば簡単な一次変換を施して）なるべく有効な変数を残し、重要で無さそうなものは消して、簡単に（低次元で）表したい。
  - 通常は新しい直交座標を$(w_1, \ldots, w_m)$として、重要なものから順に第1成分、第2成分…と呼ぶ。この順に区分けが明確になるよう決める。
  - クラスタリングとは別の意味の区分になるので、教師無し学習とも考えられる。
  - 初めに主成分分析で次元を落としてからクラスタリングをする、逆にクラスタリングののち各クラスタの特徴を探るためにそのデータのみに着目して主成分分析を行う、両者平行して行い比較するなど、組み合わせて用いることもある。

主成分分析 (principle component analysis, PCA)

- 基本方針
  - データを「ある軸」に射影したときに、なるべく区分けができるように分散が大きい軸を「ある軸」$\boldsymbol{w}_1$としてを選択する (第1主成分)。
  - 次に、第1主成分と直交する軸で、同様にその軸に射影したときに分散値が大きくなるものを第2主成分$\boldsymbol{w}_2$として選択する。
  - これまで選んだ第k−1主成分まですべてと直交し、射影したときに分散値が大きくなる軸を第k主成分$\boldsymbol{w}_k$として選択する。
  - 上記を第m主成分$\boldsymbol{w}_m$を得るまで繰り返す。
  - なお軸$\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$の大きさは1、つまり
    $$\| \boldsymbol{w}_i \|^2 = \boldsymbol{w}_i \cdot \boldsymbol{w}_i^T = 1$$
    としておく（一般性を失わない）。



主成分の求め方(1)

- 各データx_iは、n個の値のベクトルで表現でき、このようなデータがN個あるとする。これを縦に並べて以下のように表す。
$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nn} \end{pmatrix}$$
- まず求めたい第1主成分を$\boldsymbol{w}_1 = (w_{11}, w_{12}, \ldots, w_{1n})$と表す。また、ベクトルの各要素 $x_{1j}, \ldots, x_{Nj}$ の値の標本平均 μ_jを求め、平均値が0となるように $\bar{x}_{ij} \leftarrow x_{ij} - \mu_j$と変換し、以下のようにおく。
$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_N \end{pmatrix} = \begin{pmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{x}_{N1} & \cdots & \bar{x}_{Nn} \end{pmatrix}$$
- 点$\bar{x}_i$の$\boldsymbol{w}_1$への射影は、内積で表せるので、
$$p_i = \bar{x}_i \cdot \boldsymbol{w}_1^T \quad \left(= \sum_{k=1}^{n} \bar{x}_{ik} w_{1k} \right)$$

主成分の求め方(2)

- これをまとめれば、
$$\begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix} = \bar{X} \cdot \boldsymbol{w}_1^T$$
- データを射影したp_iの不偏標本分散σ^2を求めると標準化してあるので
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} p_i^2 = \boldsymbol{w}_1 \cdot \left(\frac{1}{N-1} \bar{X}^T \bar{X} \right) \cdot \boldsymbol{w}_1^T$$
（標本分散を使っても良いこととする、そのときは$1/N$をつかう）。
- ここでで、$\frac{1}{N-1} \bar{X}^T \bar{X}$の$ij$成分を$\bar{\sigma}_{ij}$とおけば、これは
$$\bar{\sigma}_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} \bar{x}_{ki} \bar{x}_{kj} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$
となり、共分散行列の不偏推定（統計の講義）となる。この行列を$\mathcal{S}$とおく。

主成分の求め方(3)

- $\| \boldsymbol{w}_1 \| - 1 = 0$のもと、 $\sigma^2 = \boldsymbol{w}_1 \mathcal{S} \boldsymbol{w}_1^T$ を最大化する $\boldsymbol{w}_1$ を求めればよい。
- ラグランジュ未定乗数法を適用する。

$$L(\boldsymbol{w}_1, \lambda) = \boldsymbol{w}_1 \mathcal{S} \boldsymbol{w}_1^T - \lambda(\boldsymbol{w}_1 \cdot \boldsymbol{w}_1^T - 1)$$

- 上記より、

$$L(\boldsymbol{w}_1, \lambda) = \left(\sum_{i=1}^{n} w_{1i}\sigma_{i1}, \dots, \sum_{i=1}^{n} w_{1i}\sigma_{in} \right) \cdot \boldsymbol{w}_1^T - \lambda(\boldsymbol{w}_1 \cdot \boldsymbol{w}_1^T - 1)$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} w_{1i}\sigma_{ij} w_{1j} - \lambda(\boldsymbol{w}_1 \cdot \boldsymbol{w}_1^T - 1)$$

- $\frac{\partial L(\boldsymbol{w}_1, \lambda)}{\partial \boldsymbol{w}_1} = 0$を考えると、

$$\frac{\partial L(\boldsymbol{w}_1, \lambda)}{\partial \boldsymbol{w}_1} = 2(\mathcal{S} - \lambda I)\boldsymbol{w}_1^T = 0$$

つまり、$\mathcal{S}\boldsymbol{w}_1^T = \lambda \boldsymbol{w}_1^T$となり、 $\mathcal{S}$の固有値を求めればよい。

主成分の求め方(4)とPCAの意味

- 固有値は、$\det(\mathcal{S} - \lambda I)$=0を解けばよいが、一般に固有値は複数個ある。どれを選ぶか？これまでは第1主成分を考えてきたが、その他についても同様で、 $\mathcal{S}$の固有値問題に帰着できる。
- $\mathcal{S}$の固有値を$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$とおき、対応する固有値ベクトルを$\omega_1, \dots, \omega_m$とする。

$$\sigma^2 = \boldsymbol{w}_1 \mathcal{S} \boldsymbol{w}_1^T = \sigma^2 = \boldsymbol{w}_1 \lambda_k \boldsymbol{w}_1^T = \lambda_k$$

- 以上から第1主成分は$w_1 = \omega_1$、第2主成分は$w_2 = \omega_2$ などとなる。
- 主成分分析の固有値ベクトルからの意味
  - ω_1 に射影したとき分散が最大となり、データを広く区分けできる。
  - 一方、それと直交する軸のデータは失われる。失われたデータの中でデータを最も広く区分するの軸がω_2 （以下同様）。

PCAの利点と欠点

- 利点
  - 相関のある変数の除去。
  - 次元を下げることができる。これによる単純化と高速化の利益は大きい
- 欠点
  - 線形性に強く依存している。これに合わないものは適用できない。

応用例：画像圧縮

- 画像は多数の点でできていて、各点は色 (RGBなど) で表されている。
  - たとえば、1024x640=655360の点 (pixel) でできているが、これを64x64に分割し、16x10の画像が4096集まっていると考える。各点の色はRGBで表現(3次元）。分割した画像を16x10x3=480次元の点で表現。これが4096ある。
  - この4096の480次元データをPCAで次元圧縮する。
  - $\boldsymbol{x}_i = (x_{i1}, \dots, x_{i480})$ (480次元) $\Rightarrow \boldsymbol{y}_i = (y_{i1}, \dots, y_{im})$ (m次元)
  - $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m$と合わせて、 $y_{i1}\boldsymbol{w}_1 + \cdots + y_{im}\boldsymbol{w}_m$として元の次元に戻す（色に戻すと言うこと。もちろん結果は前とは違う）
    - たくさんの色が使われている領域では、大きく劣化するかもしれない。
    - 他にも方法がある。画像圧縮については、いろいろな研究があり、それらは専門の講義で。

課題６－１

- 以下の手順に従って、主成分分析に関するプログラムを書き、実験を行いなさい
  - 正規分布$N(0, 4)$に従う乱数を適当な数M個 (Mは50~100ぐらい) 生成し、それをa_i (i =1, …, M)とおく。
  - $N(3, 2)$に従う乱数をM個生成し、これをb_iとおく。
  - 0~4の範囲で一様に発生する乱数をM個用意し、これをc_iとおく。
1. $x_i=(a_i, b_i, c_i)$としてM個のデータを作る。このデータから不偏共分散行列を求めよ。
2. この行列の固有値を求めよ。
3. PCAの第1主成分、第2主成分をもとめ、その結果について考察せよ。

- なお、正規分布の平均値をμ分散をσ^2とすると、$N(\mu, \sigma^2)$と表します。

課題６－２

- 以下の手順に従って、K-means方でクラスタリングするプログラムを書き、実験を行いなさい
  - 0~5の範囲で一様に発生する乱数を数M個 (Mは50~100ぐらい) 生成し、それをa_i (i =1, …, M)とおく。
  - 正規分布$N(7, 4)$と$N(-3, 3)$に従う乱数を同数、合わせてM個生成し、ランダムに並べ替える。これをb_iとおく。
  - $N(5, 5)$と$N(-1, 4)$に従う乱数を同数、合わせてM個生成し、ランダムに並べ替える。これをc_iとおく。
1. $x_i=(a_i, b_i, c_i)$としてM個のデータを作る。このデータをK個のクラスタに分けてみよ。 K=2, 3, 4, 5 ..として試してその重心をもとめてみなさい。
2. Kはいくつが良さそうか、考えなさい。

36 混合分布

ここはやや難しいので概略だけ。たぶん３年の後期か大学院で。

人工知能の概論と言うより他の応用に結びつけて。

準備（EMアルゴリズム）

- Expectation maximization algorithm：基本的には隠れ変数（値が未知の変数）とある統計量について、隠れ変数の推定と統計量の推定（最尤推定）を交互に行い、値を徐々に更新する方法。
  - ある統計量θ（たとえば平均値とか確率とか、何か知りたいもの）を最尤推定する。最尤推定は、観測$X=(x_1, …, x_n)$に対して、その観測が最大となる確率$P(X|\theta)$が最大となる$\tilde{\theta}$を求めれば（つまり微分して0となるときの値を調べる）それがθの推定値であるとするもの。
  - ただし、観測できない変数もあり（隠れ変数という）、これを$z_1, …, z_k$とする。
  - $\tilde{\theta}$を適当に決め、そこから隠れ変数$z_1, …, z_k$を推定する。次に決めた$z_1, …, z_k$に基づいてθの推定値$\tilde{\theta}$を求める。その$\tilde{\theta}$を利用して、隠れ変数$z_1, …, z_k$を推定。これを収束条件が満たすまで繰り返す。

EMアルゴリズム (1)

- 少し確率・統計の復習。
  - 観測できる変数$\boldsymbol{x}$と観測できない変数$\boldsymbol{z}$を仮定し、確率$p(\boldsymbol{x}, \boldsymbol{z})$を考える。$\boldsymbol{x}$の周辺分布を求める。

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \; (= \sum_{\boldsymbol{z}} p(\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x}))$$

  - これに推定したいパラメータθを表現に加えて、$p(\boldsymbol{x}|\theta)$と置換え。これは推定したいパラメータをθとしたときの$\boldsymbol{x}$が観測できる確率。

$$p(\boldsymbol{x}|\theta) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\theta)p(\boldsymbol{x}|\boldsymbol{z}, \theta)$$

  - なお各確率の値はθに依存するものと考える。

EMアルゴリズム (2)

- EMアルゴリズムを形式的に。
  1. $\theta(0)$を初期値として選ぶ。
  2. E: $p(\boldsymbol{z}|\boldsymbol{x}, \theta(t))$を計算する。($E$ = expectation)
  3. M: $J = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \theta(t)) \ln p(\boldsymbol{x}, \boldsymbol{z}|\tilde{\theta})$を最大化する$\tilde{\theta}$を求めこれを$\theta(t+1)$とする。($M$ = maximization)
  4. EとMを収束するまで繰り返す。
- なお、θの関数 $Q(\theta; \theta(t)) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \theta(t)) \ln p(\boldsymbol{x}, \boldsymbol{z}|\theta)$とおいて、

$$\tilde{\theta} = arg \max_{\theta} Q(\theta; \theta(t))$$

  と表現することもある。
- Mの式で自然対数が使われているが、最尤推定ではよく使われる。同時確率は、それらの積となるので対数をとると和となり簡易化
- Mの式で自然対数値の期待値が使われているが、これは$\boldsymbol{z}$を直接観測できないため、Eで求めた$\boldsymbol{z}$の確率を利用して求めている。

EMアルゴリズムの適用例

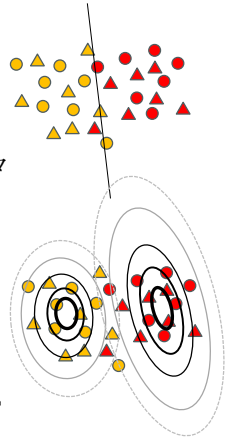
- K-means法
  - $\Omega=\{x_1, \ldots, x_n\}$、$\theta=(c_1, \ldots, c_K)$. 隠れ変数は$\boldsymbol{\gamma}_i=(\gamma_{i1}, \ldots, \gamma_{iK})$, ただし

$$\gamma_{ik} = \begin{cases} 1 & \text{if } x_i \text{が} c_k \text{に一番近いとき} \\ 0 & \text{その他} \end{cases}$$

  - 当然隠れ変数は分からない（どちらかと言えば、これを求めたい）
1. ランダムに$\theta(0) =(c_1(0), \ldots, c_K(0))$.を選ぶ。
2. E: $\theta(t)$を使って$\gamma_{ik}(t)$を求める。これが隠れ変数の推定値。具体的には、x_iに一番近い$c_k(t)$を求めることに相当。
3. M: $c_k(t+1)$を求める。$Q(c_k; c_k(t)) = \sum_{i=1}^{n} \gamma_{ik}(t) \parallel x_i - c_k \parallel^2$が最小となる c_k を $c_k(t+1)$ とする。なおここでは最小値だが、これが、クラスタに属する確率を最大化するものと定義している。具体的には、$Q(c_k; c_k(t))$をc_kで微分して0と置けば新しい $c_k(t+1)$ が平均値（重心）として求められる。

混合ガウス分布（混合正規分布）

- これまでの方法の欠点
  - 実データでは境界付近が綺麗に分かれるとは限らない。これを強制的に一方に分けていた。
  - どこのクラスタに属すかを確率的に表現し、境界の部分は曖昧性をもって表現する。
- このためにデータありきではなく、データが生成された過程を考える。
  - そもそも異なるクラスタのデータは、異なるメカニズム（ここでは確率分布）にしたがって生成されていると仮定。
  - クラスタがK個あるとすると、K個の異なる分布$f_k(\boldsymbol{x})$が確率α_kに従って選択され($1 \leqq k \leqq K$)、それに基づいてm次元データ$\boldsymbol{x}$が生成されると考える。



混合ガウス分布（混合正規分布）

- ここで分布$f_k(\boldsymbol{x})$が正規分布$N(\boldsymbol{\mu}_k, \Sigma_k)$に従うと仮定する。
  - もちろん、平均$\boldsymbol{\mu}_k$と（分散）共分散行列分散$\boldsymbol{\Sigma}_k$は不明なので、これを知る必要がある。
  - ここでK-means法を思い出すと、クラスタの中心c_iに相当するのはクラスタの平均$\boldsymbol{\mu}_k$に相当する。（K-means法との関係）
- 考え方（最適化の講義で主に扱うので簡単に）
  - データの集合を$\Omega = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$とする。$\boldsymbol{x}_i$は、$f_1, \ldots, f_K$のどれかに従うので、最も尤もらしいものを対応づける→最尤推定法 (EM)
  - x_iが分布f_jに従って生成される確率は、$P(\boldsymbol{x}_i, f_j)=P(\boldsymbol{x}_i|f_j)P(f_j)=P(\boldsymbol{x}_i|f_j)\alpha_j$である。$f_j$は正規分布を仮定しているので、$P(\boldsymbol{x}_i|f_j)$は正規分布にしたがった確率となる。つまり$f_j(\boldsymbol{x}_i)$.

$$P(\boldsymbol{x_i}) = \prod_{j=1}^{K} \alpha_j \cdot f_j(\boldsymbol{x_i})$$

混合ガウス分布（混合正規分布）

- この先は最適化の講義で詳しくやると思うが、

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik} \cdot x_i$$

ただし、$N_k = \sum_{i=1}^{n} \gamma_{ik}$

- $\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{n} \gamma_{ik} \cdot (x_i - \mu_k)^2$
- $\alpha_k = \frac{N_k}{N}$
- となり、EMアルゴリズムで解く

- 混合分布は最適化問題やパターン認識でよく使われるので、そちらの講義で詳しくあるはず（大学院かも）。

混合ガウス分布（混合正規分布）

- 一方ベイズの定理から、

$$P(f_j|\boldsymbol{x}_i) = \frac{P(\boldsymbol{x}_i, f_j)}{\sum_{k=1}^{K} P(f_k)P(\boldsymbol{x}_i|f_k)} = \frac{\alpha_j f_j(\boldsymbol{x}_i)}{\sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{x}_i)}$$

これをγ_{ij}とおく。ただし

$$f_k(\boldsymbol{x}_i) = \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{|\Sigma_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\right) \quad (\in [0,1])$$

- ここから$\alpha_k, \boldsymbol{\mu}_k, \Sigma_k$を求める。
- しかし$\boldsymbol{x}_i$がどの正規分布に従ったものかは分からない。これを隠れ変数とする。つまり、
- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$であり、ただ一つの要素のみ1で、残りは0。もちろん分からない。