第１章　パターン認識の基礎

1-1 線形代数

線形代数は、機械学習だけでなくありとあらゆる分野（工学、経済学等）で現れる分野である。機械学習の分野では主に、複雑なデータを表現するための都合がいい道具として使われることが多い。1-1 では準備体操として行列の基本的な演算と便利な性質を見ていく。

1-1 ベクトル、行列とは

1-1 基本的な演算

1-1 逆行列

$A\vec{x} = \vec{b}$

$\vec{x} = A^{-1}\vec{b}$

1-1 転置

1-1 対称行列

$n \times n$ 行列（以下は全て $n \times n$ 行列の議論である）において

$$A = A^T$$

が成立するとき、その行列を対称行列と呼ぶ。

1-1 行列式

1-1 内積、1 次形式、2 次形式

ベクトルの内積を次のように表記する。

$$f = (\vec{a}, \vec{x})$$

ただし、ベクトル $\vec{a}, \vec{x}$ は

$$\vec{a} =, \vec{x} =$$

とし、f を以下のようにする。

$$f = a_1 x_1 + \cdots + a_n x_n = \sum_{i=1}^{n} a_i x_i$$

変数の 1 次の項のみからなる式を 1 次形式という。変数の 2 次の項のみからなる式もあり、これを 2 次形式と呼び、以下のようになる。

$$f = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{(n-1)n}x_{n-1}x_n = \sum_{i,j=1}^{n} a_{i,j}x_ix_j$$

これを（）のように内積表記すれば、

$$f = (\vec{x}, A\vec{x})$$

となる。ただし A は対称行列である。

(1) 任意のベクトル $\vec{x}, \vec{y}$, 行列 A に対して、$(A\vec{x}, \vec{y}) = (\vec{x}, A^T\vec{y})$ が成立することを示せ。

(2) 任意の $n \times n$ の行列 A, B に対して $(AB)^T = B^TA^T$ が成立することを示せ。

1-1 行列の微分

行列の微分は機械学習で頻繁に登場する。代表的な 1 次形式と 2 次形式の 2 つの微分方法を説明する。

線形結合と独立

1-1 固有値、固有ベクトル

大抵のベクトルは行列 A をかけると予期しない方向に回転する。ここでは行列 A を作用させてもベクトルの向きが変化しない（A に固有な）特別なベクトルを考える。このベクトルを $\vec{u}$, ベクトルの拡大率を λ とすると、

$$A\vec{u} = \lambda\vec{u}$$

とかける。この $\vec{u}$ を固有ベクトル、λ を固有値という。

（）から単純な式変形で以下の方程式、

$$det|A - \lambda I| = 0$$

を利用すれば A の固有値、固有ベクトルが求まることがわかる。

(3) $A = ()$ の固有値、固有ベクトルを求めよ。

A に n 個の独立な固有ベクトル $\vec{x_1} \cdots\cdots \vec{x_n}$ があればそれらは S の列に入り、以下のように対角化される。ここで S は固有ベクトルを列に持つ、$n \times n$ 行列である。

$$A = S\Lambda S^{-1}$$

(4) 対角化 $A = S\Lambda S^{-1}$ ができるとし、A^k を求めよ。

上記問で、対角化を用いて A^k を求めたが、これは行列微分方程式や、フーリエ変換と行った工学上重要な分野で利用されている。

1-1 対称行列の対角化

固有値が重複することによって固有ベクトルが足りなくなることがある。このとき行列 A の対角化は不可能になってしまう。しかし、対称行列の場合、必ず対角化することが可能であり、A の要素が全て実数（実対称行列）なら、

1. 固有値が全て実数
2. 固有ベクトルも（都合がいいことに）必ず全て直交

という性質を持つ。

(5) 実対称行列の全ての固有値が実数であることを示せ。

(6) 実対称行列の異なる λ に対応する固有ベクトルは必ず直交することを示せ。

2. の性質を利用し、固有ベクトルを正規直交するように定数倍すると、() の対角化をさらに簡単に行うことができる。

（スペクトル定理）全ての対称行列は、実数固有値からなる Λ と正規直交する固有ベクトルからなる $S = Q$ を用いて

$$A = Q\Lambda Q^T$$

と分解される。ここで Q は直交行列と呼ばれ、以下の性質を満たす。（これは物理、工学の分野では主軸定理と呼ばれている。）

一部文献では、以下のような表記も行われることがある。

$$A = \lambda_1 \vec{x_1}\vec{x_1}^T + \lambda_1 \vec{x_1}\vec{x_1}^T + \cdots + \lambda_n \vec{x_n}\vec{x_n}^T$$

（注意）固有値が重複した場合、固有ベクトルが足りず対角化不可能のように思われるが、重複した固有値にはその重複度分の独立な固有ベクトルが存在することが保証される。（証明割愛）これらは互いに直交していないことが多いが、シュミットの直交化を用いて、独立なベクトルを正規直交するベクトルに取り直すことができるので、結局全ての固有ベクトルを直交するようにとることができる。（を p 次元固有空間と 呼ぶ。）

$$q_i q_j = \delta_{i,j}$$

$$\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$Q^{-1} = Q^T$$

これらの性質から対称行列は工学上よく利用される。

1-1 2 次形式の標準形

２次形式で表される 2 変数の 2 次関数

$$ax^2 + 2bxy + cy^2 = 1$$

を考える。これは（傾いた）楕円を表す。x, y に行列 U を作用し、

$$\vec{x} = U\vec{\tilde{x}}$$

とする。このとき 2 次形式の内積と対称行列の対角化を利用すると以下のようになる。

$$(\vec{x}, A\vec{x}) = (U\vec{\tilde{x}}, AU\vec{\tilde{x}}) = (\vec{\tilde{x}}, U^T AU\vec{\tilde{x}}) = (\vec{\tilde{x}}, \Lambda\vec{\tilde{x}}) = \lambda_1 \tilde{x_1}^2 + \lambda_2 \tilde{x_2}^2 + \cdots + \lambda_n \tilde{x_n}^2$$

これを 2 次形式の標準形と呼ぶ。これは元々の楕円 $(\vec{x}, A\vec{x})$ を回転させ、新しい軸の方向に長軸、短軸をとることができることを意味している。

$\lambda_1, \lambda_2, \cdots \lambda_n$ が正の値をとるとき、$(\vec{x}, A\vec{x}) > 0$ となる。固有値が全て正の値をとる行列 A のことを正定値行列と呼び、このとき行列 A は、$(\vec{x}, A\vec{x}) > 0$ を満たす。またこの形式の 2 次曲面、

$$f(x, y) = ax^2 + 2bxy + cy^2$$

を考えると（）より、

$$(\vec{x}, A\vec{x}) = \lambda_1 \tilde{x_1}^2 + \lambda_2 \tilde{x_2}^2$$

であり、$\lambda_1 > 0, \lambda_2 > 0$ のとき、この 2 次曲面は唯一の最小値を持つはずである。このことから、2 次曲面が最小値（または最大値）を持つか持たないか（これはたびたび機械学習の分野で話題になる）は、2 次形式 $(\vec{x}, A\vec{x})$ で式を表した際の行列 A が正定値行列か（もしくは負定値行列か）どうかを求めれば判定することができる。

1-2 単回帰

正規方程式は次のようになる

$$X^T X \vec{w} = X^T \vec{y}$$

$X^T X$ が正則なとき上記方程式の解は

$$\vec{w} = (X^T X)^{-1} X^T \vec{y}$$

となる。これが単回帰における重みの値となる

1-3 パーセプトロン

パーセプトロンの学習規則は i 番目のデータ x_i を入力したときの出力 $f(x_i)$ に応じて以下のようになる。

$$w_{i+1} = \begin{cases} w_i & f(x_i) \geqq 0 \\ w_i + \eta x_i & f(x_i) < 0 \end{cases}$$

また、パーセプトロンは有限の学習回数で収束することが保証されており、学習回数を M とすると

$$M \frac{D^2(\vec{w^*})\eta}{d(\eta + 2\alpha)} \leq \phi \leq 1 \Longrightarrow M \leq d \frac{1 + 2\alpha/\eta}{D_{max}^2}$$

と上から評価することができる。これは「パーセプトロンの収束定理」と呼ばれている。

1-4 ロジスティック回帰

1-5 SVM（サポートベクトルマシン）

$$t_i(w^Tx_i+b)>0,\quad t_i=1,2,...n,\quad t_i=\begin{cases}1 & x_i\in K_1\\ -1 & x_i\in K_2\end{cases}$$

p 次元データ $x=(x_1,x_2,...,x_p)$ と超平面 $w^Tx_i+b=0$ の距離を

$$d=$$

とし、2 つのクラスを分ける超平面とそれに最も近いデータ (サポートベクトル) との間の距離 (マージン M) を最大化するように w,b を最適化する。

$$argmax_{w,b}M,\quad \frac{t_i(w^Tx_i+b)}{||w||}\geq M,\quad i=1,2,...n$$

$$argmin_{w,b}\frac{1}{2}||w||^2,\quad t_i(w^Tx_i+b)\geq 1,\quad i=1,2,...n$$

スラッグ変数 ϵ_i を導入してこの最適化問題を書き換えると、

$$argmin_{w,b}\left\{\frac{1}{2}||w||^2+C\sum_{i=0}^{n}\epsilon_i\right\},\quad t_i(w^Tx_i+b)\geq 1-\epsilon_i,\quad \epsilon_i\geq 0\quad i=1,2,...n$$

双対問題

$$argmax_\alpha\left\{L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=j}^{n}\alpha_i\alpha_iy_iy_ix_i^Tx_j\right\}$$

$$\sum_{i=1}^{n}\alpha_iy_i=0,\quad 0\leq\alpha_i\leq C,\quad i=1,2,...n$$

前処理、高次元特徴空間

$$argmax_\alpha\left\{L(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{i=j}^{n}\alpha_i\alpha_iy_iy_i\Phi(x)_i^T\Phi(x)_j^T\right\}$$

$$\sum_{i=1}^{n}\alpha_iy_i=0,\quad 0\leq\alpha_i\leq C,\quad i=1,2,...n$$

カーネル法

1-6 教師なし学習

教師データを使わずに学習を行うことを「教師なし学習」と呼ぶ。教師なし学習は、主にクラスタリングの用途で使われ、データに隠れている構造を発見したり、教師あり学習の前処理として用いられる。

クラスタリング (教師なし学習) によく使われる手法に「K-means アルゴリズム」がある。K-means アルゴリズムでは以下の目的関数を最小化する。

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\vec{x}_n - \vec{\mu}_n||^2$$

$\vec{\mu}_k$ について解くと、

$$\vec{\mu}_k = \frac{\sum_n r_{nk} \vec{x}_n}{\sum_n r_{nk}}$$

この式の分母は k 番目のクラスタに割り当てられたデータの数に等しいので、$\vec{\mu}_k$ は、k 番目のクラスタに割り当てられた全てのデータ点 $\vec{x_n}$ の平均となっている。これが K-means アルゴリズムと呼ばれている理由である。なお、K-means アルゴリズムは次章の混合ガウス分布に対する EM アルゴリズムの非確率的極限となっている。