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。

$$f = a_1x_1 + \cdots + a_nx_n = \sum_{i=1}^{n} a_ix_i$$

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、

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

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。

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

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) 式 () の正規方程式を導け。

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 M で収束することが保証されている。(パーセプトロンの収束定理)

(2) パーセプトロンの収束定理

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

を証明せよ。

1-4 ロジスティック回帰

パターン認識を行う際、単に分類結果だけではなく、「クラス C_1 に属する確率 70%、クラス C_2 に属する確率 30%」といった確率値を出力してくれた方が便利である。これを実現する手法で代表的なものが「ロジスティック回帰」である。

1-4-1 一般化線形モデル

ロジスティック回帰を考える前に、1-3 節でも扱った以下の線形モデル

$$y = w^T x$$

を考える。左辺の予測したい値 y は「応答変数」、右辺は「線形予測子」と呼ばれている。応答変数が正規分布であるような問題を考える場合はこの定式で十分であるが、応答変数が非負の実数である場合や離散変数である場合は (例として気温や、カテゴリーデータなど)、右辺の線形予測子との対応関係が取れないため、意味のある出力結果が得られない。こういった場合には、式 () に数学的な操作を行って両辺の対応関係がうまくとれるように変形することができる。

例として、応答変数 y が非負の実数である場合、

$$\ln(y) = w^T x$$

とすれば、$0 < \ln(y) < \infty$ であるため、両辺の対応関係がとれる。

両辺の対応関係をとるために用いる関数は「リンク関数」と呼ばれている。このようにして、応答変数が正規分布以外の分布に従うものに対して線形モデルを拡張したものを「一般化線形モデル」と呼ぶ。

1-4-2 ロジット関数とロジスティック回帰

出力結果を確率値 $(0, 1)$ にするためには、以下のようにリンク関数にロジット関数を適用すれば良い。

$$\ln\Big(\frac{p}{1-p}\Big) = w^T x$$

これを p について解くと、

$$p = \frac{1}{1 + \exp(w^T x)}$$

(回帰という名前がついていて紛らわしいが、線形関数を $(0, 1)$ の間に回帰するという意味で、ロジスティック回帰自体は分類に使われる。)

1-4-3 最尤推定

2 クラス分類を考える。n 番目のデータをロジスティック回帰に適用したときに得られる出力確率 P_n は、教師データ (ラベル変数)t_n を用いて以下のように表せる。(このような分布はベルヌーイ分布と呼ばれている。)

$$P_n = p_n^{t_n}(1-p_n)^{1-t_n}$$

この P_n を学習する全データについて計算して掛け合わせたものを考え、これを $L(w)$ とすると、

$$L(w) = P_1 P_2 \cdots P_N = \prod_{n=1}^{N} p_n^{t_n}(1-p_n)^{1-t_n}$$

この $L(w)$ は「尤度関数」と呼ばれている。(尤度、最尤推定については Column 参照。) この尤度関数を最大化するような w を求めることが、ロジスティック回帰における「学習」である。実際は、数値計算の観点からこの尤度関数ではなく、次の負の対数尤度関数、

$$E(w) = -\ln L(w) = -\sum_{n=1}^{N}(t_n \ln p_n + (1-t_n)\ln(1-p_n))$$

を最小化する。この $E(w)$ を「交差エントロピー誤差関数」と呼び、$E(w)$ を最小とするパラメータ w は勾配降下法といった最適化手法が用いられる。

(1) $E(w)$ を w で微分して以下を示せ。

$$\frac{\partial E(w)}{\partial w} = \sum_{n=1}^{N} x_n(y_n - t_n)$$

1-5 SVM（サポートベクトルマシン）

1-5-1 SVM の主問題

次のようなクラス分類の問題を考える。

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

この問題を解くために、p 次元データ $x = (x_1, x_2, ..., x_p)$ と超平面 $w^T x_i + b = 0$ の距離を

$$d = \frac{|w_1 x_1 + w_2 x_2 + w_3 x_3 + b|}{\sqrt{w_1^2 + w_2^2 + ...w_p^2}} = \frac{|w^T x_i + b|}{||w||}$$

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。

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

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

また、線形分離可能でない場合にスラッグ変数 ϵ_i を導入してこの最適化問題の制約を緩めることができる。

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

これはソフトマージン SVM と呼ばれている。またこれらの最適化問題は、不等式制約条件最適化問題の「主問題」と呼ばれており、主問題に対して「双対問題」という形式の問題を導くことができる。

1-5-2 SVM の双対問題と非線形分離

ラグランジュ未定乗数法を用いると、主問題を双対問題に変形することができる。

$$argmax_{\alpha} \left\{ L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=j}^{n} \alpha_i \alpha_i y_i y_i x_i^T x_j \right\}$$

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

学習データが線形識別関数で分離できない場合は、高次元非線形空間にデータ点を写像し、その空間内で線形識別関数を用いると、線形分離可能となる可能性がある。主問題を双対問題に変形しておくと、学習データの高次元非線形空間への写像は単に

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

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

と問題を変えるだけでよいので定式化が容易であり、さらにカーネル法 (次節参照) を適用できる形式になる。

(1) 式 () の最適化問題にラグランジュ未定乗数法を適用することで得られる以下のラグランジュ関数

$$L(w, b, \epsilon, \alpha, \beta) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} \alpha_i \{ t_i(w^T x_i + b) - 1 + \epsilon_i \} - \sum_{i=1}^{n} \beta_i \epsilon_i$$

について、主変数 w, b, ϵ の偏微分を考え、双対問題 () を導け。

1-5-3 カーネル法

双対問題に現れる値 $\Phi(x)_i$ は計算量が多く最適化問題を解くが困難な形になってしまっているが、$\Phi(x)_i$ の内積 $\Phi(x)_i^T\Phi(x)_j$ は $\Phi(x)_i$ を計算しないで求めることができることが知られている。

$$K(x_i, x_j) = \Phi(x)_i^T\Phi(x)_j$$

$K(x_i, x_j)$ をカーネル関数と呼び、このようにして内積を計算する手法を「カーネル法」とも呼ぶ。(これを決めてしまえば内積計算 $\Phi(x)_i^T\Phi(x)_j$ を簡単に行うことができてしまうため、この一見魔法のような手法は「カーネルトリック」も呼ばれている。)

以下にいくつかのカーネル関数をあげる。(実際は、解きたい問題に合わせてこれらを使い分ける。)

ガウスカーネル

$$K(x_i, x_j) = \exp\left\{-\frac{||x_i - x_j||^2}{2\sigma^2}\right\}$$

多項式カーネル

$$K(x_i, x_j) = (x_i^T x_j + c)^d$$

シグモイドカーネル

$$K(x_i, x_j) = tanh(bx_i^T x_j + c)$$

1-6 教師なし学習

1-6-1 k-means アルゴリズム

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。)

1-6-2 次元削減

教師なし学習の応用の一つに「次元削減」があげられる。学習するパラメータの数を減らすことができるため、計算時間を減らすために主に用いられる。

次元削減は数学的に見れば、データ空間でより次元が低い部分空間にデータ点を射影することで実現される。しかしその過程でデータが元々持っていた情報が失われてしまうため、できるだけ情報を残しながらデータ空間の次元を下げる必要がある。これを実現するための手法として、データの持つ分散 (ばらつき) が大きくなるようにデータを射影するものがあり、これは「主成分分析」と呼ばれる。

学習用データを表す行列を $X(NM)$ とおく。このときデータの共分散行列は以下のように表すことができる。

$$\Sigma = \frac{1}{N}X^T X$$

ある特定の方向をもつ単位ベクトル $\vec{e}$ を考える。このときこの単位ベクトルにデータ点を射影してできる新しいデータ点に対して分散 σ を考えると、

$$\sigma = \vec{e}^T \Sigma \vec{e}$$

となる。この σ に対して、ラグランジュ未定乗数法を適用して σ の最大値を求める (分散最大化) をしようとすると、以下の最適化問題を解くことになる。

$$argmax_{\vec{e}} \left\{ L(\vec{e}, \lambda) = \sigma - \lambda(||\vec{e}||^2 - 1) \right\}, \quad ||\vec{e}||^2 = 1$$

この最適化問題を解くと主成分分析は以下の固有値問題、

$$\Sigma\vec{e} = \lambda\vec{e}$$

となる。分散が最大になるような単位ベクトルを「第一主成分」と呼び、上記の固有値問題のうち、固有値が最大になる固有ベクトルが第一主成分に対応する。

(1) 3 つの 2 次元データ $\vec{a_1}, \vec{a_2}, \vec{a_3}$ を考え、主成分分析が式 () の固有値問題で定式化されることを確認せよ。

(略解) $\vec{a_1}, \vec{a_2}, \vec{a_3}$ を並べた 3×2 行列 X を考え、

$$\Sigma = \frac{1}{3} X^T X$$

が成立することを確認する。次に 2 次元単位ベクトル $\vec{e} = (e_x, e_y)^T$ を考え、$\vec{a_1}, \vec{a_2}, \vec{a_3}$ それぞれとの内積を考えることで分散 σ を求めて式 () を示す。最後にこの σ を用いて式 () を構成してラグランジュ未定乗数法を用いて以下の値を計算してその結果を用いて固有方程式 () を導出する。

$$\frac{\partial L(e_x, e_y, \lambda)}{\partial e_x}, \quad \frac{\partial L(e_x, e_y, \lambda)}{\partial e_y}$$