第六章
線形識別関数

11T4066Y
松村拓真

本章の内容

- 識別関数の値によってクラスを決定する
  識別規則について解説
- 超平面の方程式を理解することが
  線形識別関数を理解する第一歩

➢線形識別関数の係数ベクトルとバイアスを
学習データを用いて決定する方法

↓

2乗誤差最小標準化基準、フィッシャーの判別関数
ロジステック回帰

6.1 線形識別関数の定義

6.1.1 線形識別関数の定義

2クラス問題(C_1, C_2)の識別関数は、d次元入力ベクトル$x = (x_1, \ldots, x_d)^T$、係数ベクトルを$w = (w_1, \ldots, w_d)^T$、バイアス項をw_0とすると

$$f(x) = w^T x + w_0$$

で表される。識別境界を$f(x) = 0$とすれば、識別規則は

$$\text{識別クラス} = \begin{cases} C_1 \ (f(x) \geq 0) \\ C_2 \ (f(x) < 0) \end{cases}$$

となり、識別境界では

$$w^T x = -w_0$$

が成り立つ。

6.1 線形識別関数の定義

- 両辺を係数ベクトルのノルム$\|w\|$で正規化すれば

$$\frac{w^T}{\|w\|}x = -\frac{w_0}{\|w\|}$$

となる。ここで、$n = w/\|w\|$、$\Delta_w = -w_0/\|w\|$とおけば

$$n^T x = \Delta_w$$

が得られ、識別境界は

$$f(x) = n^T x - \Delta_w = 0$$

と表される。

6.1 線形識別関数の定義

- 識別境界上の任意の点の位置ベクトルPを考えると

$$f(P) = n^T P - \Delta_w = 0$$

が成り立つので

$$f(x) = n^T x - \Delta_w = n^T (x - P) = 0$$

となる。

⇒識別境界は**単位法線ベクトル**nをもつ超平面となる。

$\Delta_w = n^T P$は、原点から識別超平面までの距離を表す。

(Δ_wは**正規化されたバイアス**とも呼ばれる)

6.1　線形識別関数の定義

6.1.2　多クラス問題への拡張

K(3以上)クラスの線形識別関数の作り方

ｉ)**一対他(one-versus-the-rest)**

一つのクラスと他の全てのクラスを識別する$K-1$個の2クラス識別関数$f_j(x)(j=1,\dots,K-1)$を用意し、

$$\text{識別クラス}\begin{cases} C_j\ (\text{ある}j\text{について}f_j(x)>0\text{の場合}) \\ C_K\ (\text{全ての}j\neq K\text{について}f_j(x)<0\text{の場合}) \end{cases}$$

の規則に従って識別する方法。

この方法は、複数の識別関数が>0となる空白領域のクラスが決定できないこと、正のクラスの学習データ数が負のデータ数に比べて極端に少なくなることが欠点である。

6.1　線形識別関数の定義

- ii)**一対一(one-versus-one)**

クラスiとjを識別する$K(K-1)/2$個の2クラス識別関数$f_{ij}(x)(1 \leq i < j \leq K)$を用意し、それを用いた多数決で識別クラスを決める。

この方法でも識別関数間で識別クラスの矛盾が生じる空白領域のクラスが決定できない。また入力データに関係のない識別関数が混入している場合もあり、多数決でも過半数を取れなかったり、関係のない票が入ったりしてしまう。

6.1 線形識別関数の定義

- 前述の二つの方法で生じる問題

⇒**最大識別関数法**を用いて回避することができる

$$\text{識別クラス} = \text{argmax} f_j(x) = \text{argmax} f_j\left(w_j^T x + w_{j0}\right)$$

識別関数値が最大のクラスを識別クラスにすれば良い。この場合、クラスiとjの間の識別境界は$f_i(x) = f_j(x)$となるため、

$$f_{ij}(x) = (w_i - w_j)^T x + \left(w_{i0} - w_{j0}\right) = 0$$

を満たす$K-1$個の識別境界ができる。この識別境界は2クラスの場合の識別境界と同じになる。

前述の方法との相違点は、どの点でも必ずどれかの識別関数が最大となり、クラスが決定できない領域が生じないことである(識別境界上を除く)。

また、各クラスの占める領域は**単連結**で**凸**となる。

6.2 最小2乗誤差基準によるパラメータの推定

- **6.2.1 正規方程式**

  最小2乗誤差基準…識別関数の出力値と教師入力との差を最小にするパラメータを求める手法

  2クラスの場合

  係数ベクトルはバイアスを含めて$w = (w_0, w_1, \ldots, w_d)^T$、$i$番目の学習用入力ベクトルはバイアスに対応する項を含めて$x_i = (w_{i0}, w_{i1}, \ldots, w_{id})_T$と定義する。線形識別関数は、

$$f(x) = w_0 + w_1 x_1 + \cdots + w_d x_d = w^T x$$

  と表すことができる。入力ベクトルx_iが所属するクラスは、教師入力t_iにより

$$\begin{cases} +1 \ (x \in C_1) \\ -1 \ (x \in C_2) \end{cases}$$

  のように与えられるものとする。

6.2 最小2乗誤差基準によるパラメータの推定

- 学習データ数をNとし、学習用の入力ベクトルを並べたデータ行列Xと、教師入力を並べた教師ベクトルtを

$$X = (x_1, \ldots, x_N)^T、\ t = (t_1, \ldots, t_N)^T$$

で定義する。識別関数の出力値と教師入力の誤差を2乗誤差で評価すれば、その評価関数$E(w)$は

$$E(w) = \sum_{i=1}^{N} (t_i - f(x_i))^2 = (t - Xw)^T (t - Xw)$$

で表される。評価関数を微分して0とおいたものが関数を最小にするパラメータwであるので、

$$\frac{\partial E(w)}{\partial w} = -2X^T(t - Xw) = 0$$

$$\Rightarrow \hat{w} = (X^T X)^{-1} X^T t$$

が得られる。この式を**正規方程式**という。

6.2 最小2乗誤差基準によるパラメータの推定

- 学習データに対する予測値$\hat{t}$は、

$$\hat{t} = X\hat{w} = X(X^T X)^{-1} X^T t$$

によって得られる。行列$X(X^T X)^{-1} X^T$は、教師データtを予測値$\hat{t}$に変換する行列(ハット行列)である。

6.2 最小2乗誤差基準によるパラメータの推定

6.2.2 多クラス問題への拡張

最小2乗誤差基準を多クラス問題に拡張することは容易

⇒最大識別関数法では、K個の識別関数

$f_k(x) = w_k^T x (k = 1 \ldots, K)$を用意して、二乗誤差を最小にするパラメータ$w_k$を同様に求めれば良い。

2クラスの場合との違いは、一つの学習データに、クラス数と同数の教師入力が必要になること。

→i番目の教師入力を$t_i = (t_{i1}, \ldots, t_{iK})^T$で表現する(各$t_{iK}$は、ダミー変数表示にすれば良い)。

6.2 最小2乗誤差基準によるパラメータの推定

- N個の学習データ$X = (x_1, \ldots, x_N)^T$に対する教師データを並べた行列を$T = (t_1, \ldots, t_N)^T$、$t = (0, \ldots, 1, \ldots, 0)^T$($K$個の要素)とすれば、2乗誤差を最小にするパラメータ$\widehat{W}$は、

$$\widehat{W} = (X^T X)^{-1} X^T T$$

で与えられ、識別関数は、

$$G(x) = \widehat{W}^{\mathrm{T}} x = (w_1, \ldots, w_K)^T \mathrm{x} = (g_1(x), \ldots, g_K(x))^T$$

となる。また、識別規則は、

$$\text{識別クラス} = argmax g_j(x)$$

となる。(上手くいく場合とそうでない場合がある。)

6.3　線形判別分析

- 線形識別関数を、「d次元ベクトルxを、ベクトルw上のスカラー関数$f(x)$に写像している」と考える

↓

線形判別方法

⇒クラス間の分布がなるべく重ならないような写像方法

6.3 線形判別分析

- **6.3.1 フィッシャーの線形判別関数**

2クラスの場合

各クラスのデータ数はN_1、N_2、全データ数は$N = N_1 + N_2$とする。線形識別関数は線形変換であるため、平均ベクトル

$$\mu_k = \frac{1}{N_K} \sum_{i \in C_k} x_i \ (k = 1,2)$$

を$m_k = w^T \mu_k$に写像する。このとき、平均値の差

$$m_1 - m_2 = w^T (\mu_1 - \mu_2)$$

が大きいほどクラスの分離が良くなり、クラスごとのデータ分布の広がり方が小さい方が、クラス間の重なりが小さくなる望みがある。

6.3 線形判別分析

- 平均の差の2乗を**クラス間変動**という
  線形識別関数で写像された1次元空間内での**クラス内変動**は

$$S_k^2 = \sum_{i \in C_k} (y_i - m_k)^2$$

  で定義される。
  (**全クラス変動**は$S_1^2 + S_2^2$)クラス間変動とクラス内変動の比

$$J(w) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2}$$

  を最大にするwを見つけること

  ⇒フィッシャーの定義

6.3 線形判別分析

- クラス間変動は、

$$(m_1 - m_2)^2 = (w^T(\mu_1 - \mu_2))^2$$
$$= w^T(\mu_1 - \mu_2)\,(\mu_1 - \mu_2)^T w = w^T S_B w$$

と表現可能(S_Bは線形変換される前の学習データのクラス間ン変動行列)。クラス内変動も同様に導出することで

$$S_k^2 = w^T S_k w$$

と表現できるので、全クラス内変動は

$$S_1^2 + S_2^2 = w^T S_W w$$

となる。よって、フィッシャーの基準は

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

となる。

6.3 線形判別分析

- 前述の式を最大化する解は

$$S_B w = \lambda S_W w$$

を解けば得られる(**一般化固有値問題**)。上式より

$$S_W^{-1} S_B w = \lambda w$$

と書ける。ここで

$$S_B w = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w \propto (\mu_1 - \mu_2)$$

となることに注意すると

$$w \propto S_W^{-1} S_B w \propto S_W^{-1}(\mu_1 - \mu_2)$$

が、フィッシャー基準による最適なwとなる。

6.3　線形判別分析

- 線形識別関数⇒ $f(x) = w^T x + w_0$

  w_0…識別境界を与えるバイアス項

  →フィッシャーの基準では消去されてしまうため、直接求められない。

↓

  w上に写像されたデータのクラス条件付き分布を1次元正規分布関数で近似し、事後確率が同じになる点をw_0とする。

6.3 線形判別分析

- **6.3.2 判別分析法**

フィッシャーの基準において、線形変換$y = w^T x + w_0$のバイアスw_0を明示的に扱うことが出来る定式化を考える。

線形変換後のyの平均値と分散は、クラス$k = 1,2$について

$$y = w^T \mu_k + w_0$$

$$\sigma_k^2 = w^T \Sigma_k w$$

で定義される。**クラス分離度**の評価関数$h(m_1, \sigma_1^2, m_2, \sigma_2^2)$を最大にする$w$と$w_0$は、それらで$h$を微分して0とおけば求まる。

$$\frac{\partial h}{\partial w} = 0, \frac{\partial h}{\partial w_0} = 0$$

上式より

$$2\left(\frac{\partial h}{\partial \sigma_1^2}\Sigma_1 + \frac{\partial h}{\partial \sigma_2^2}\Sigma_2\right)w = \frac{\partial h}{\partial m_1}\mu_1 + \frac{\partial h}{\partial m_2}\mu_2 \ldots (1)$$

$$\frac{\partial h}{\partial m_1} + \frac{\partial h}{\partial m_2} = 0 \ldots (2)$$

6.3 線形判別分析

$$s = \frac{\frac{\partial h}{\partial \sigma_1^2}}{\frac{\partial h}{\partial \sigma_1^2} + \frac{\partial h}{\partial \sigma_2^2}} \cdots (3)$$

とおいて式(1)を整理すると

$$2\left(\frac{\partial h}{\partial \sigma_1^2} + \frac{\partial h}{\partial \sigma_2^2}\right)(s\Sigma_1 + (1-s)\Sigma_2)w = \frac{\partial h}{\partial m_1}(\mu_2 - \mu_1)$$

が得られるため、最適なwは、スカラー項を無視すれば

$$w = (s\Sigma_1 + (1-s)\Sigma_2)^{-1}(\mu_2 - \mu_1)$$

となる。評価関数をクラス間分散とクラス内分散の比で定義した判別関数を、判別分析法という。

$$h = \frac{\text{クラス間分散}}{\text{クラス内分散}} = \frac{P(C_1)(m_1 - \bar{m})^2 + P(C_2)(m_2 - \bar{m})^2}{P(C_1)\sigma_1^2 + P(C_2)\sigma_2^2}$$

($\bar{m}$は全データの平均　$\bar{m} = (N_1 m_1 + N_2 m_2)/N$)

6.3 線形判別分析

- この場合、

$$\frac{\partial h}{\partial \sigma_k^2} = \frac{P(C_k)P(C_1)(m_1-\bar{m})^2 + P(C_2)(m_2-\bar{m})^2}{(P(C_1)\sigma_1^2 + P(C_2)\sigma_2^2)^2} \quad (k = 1,2)$$

を(3)に代入すると$s = P(C_1)$が得られるため、最適なwは

$$\mathrm{w} = (P(C_1)\Sigma_1 + P(C_2)\Sigma_2)^{-1}(\mu_2 - \mu_1)$$

となる。また、$\frac{\partial h}{\partial m_k}$と(2)より

$$P(C_1)(m_1 - \bar{m}) + P(C_2)(m_2 - \bar{m}) = 0$$

が得られるため、$m_k = w^T\mu_k + w_0$を代入して整理すると、最適なバイアス項

$$w_0 = -w^T P(C_1)\mu_1 + P(C_2)\mu_2$$

が得られる。

6.3 線形判別分析

- **6.3.3 多クラス問題への拡張**

多クラスの場合は、識別境界は計算不可。

各クラスのデータ数を$N_k(k = 1, \dots, K)$とする。2クラスと同様に、各クラスのクラス内変動を

$$S_k = \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \ , \mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i$$

で定義し、全クラスのクラス内変動の和を$S_W = S_1 + \cdots + S_K$と定義する。全データ数を$N = N_1 + \cdots + N_K$とすれば、平均μは

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{k=1}^{K} N_k \mu_k$$

で与えられる。全平均からの全データの変動の和S_Tを、**全変動**という。

6.3 線形判別分析

- $$S_T = \sum_{i=1}^{N}(x_i - \mu_k)(x_i - \mu_k)^T$$

S_Tは以下のように分解することが出来る。

$$S_T = S_W + \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

上式を用いてクラス間変動

$$S_B = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

を定義する。上式と、μ_kとμの制約より、S_Bのランクはバイアス項を除いたデータの次元をdとすれば、たかだか$\min[K-1, d]$。

$d > K$であれば、線形写像

$$y_k = w_k^T x \ (k = 1, \dots, K-1)$$

を考える。$y = (y_1, \dots, y_{K-1})$、$W = (w_1, \dots, w_{K-1})$とすると、線形変換は$y = W^T x$と書ける。

6.3 線形判別分析

- 線形変換後のクラス内変動$\widetilde{S_W}$、クラス間変動$\widetilde{S_B}$、全変動$\widetilde{S_T}$、は各クラスの平均ベクトルm_kと全平均ベクトルmが、

$$m_k = W^T \mu_k \, , m = W^T \mu$$

となるため、

$$\widetilde{S_W} = W^T S_W W$$
$$\widetilde{S_B} = W^T S_B W$$
$$\widetilde{S_T} = \widetilde{S_W} + \widetilde{S_B}$$

となる。

最適な写像行列Wを求める基準は、クラス間変動行列とクラス内変動行列の比を最大化することである。

↓

スカラー量に変換しないと最大値を求められない。

(様々な方法がある)

6.4 ロジスティック回帰

- **ロジスティック回帰**…関数値を区間$(0,1)$に制限し、確率的な解釈を可能にするもの。

6.4.1 ロジスティック関数

2クラスの場合

クラスC_1の事後確率$P(C_1|x)$は

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$

である。ここで、$a = ln\, p(x|C_1)P(C_1)/p(x|C_2)P(C_2)$とおけば、

$$P(C_1|x) = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

と表せる。

$\sigma(a)$⇒**ロジステック関数(ロジスティック・シグモイド関数)**

6.4 ロジスティック回帰

- **6.4.2 ロジスティック回帰モデル**

ロジスティック回帰モデル…事象の有無{0,1}の2値で表し、
事象の生起確率をロジスティック関数で表す。

例)喫煙量と肺がん発生の有無

N人の喫煙量$\hat{x} = (x_1, \dots, x_N)^T$を観測したとき、肺がんになる確率を

$$p(1|x_1, \dots, x_N) = f(x) = \frac{1}{1 + \exp(-(w_0 + w_1 x_1 + \cdots + w_N x_N))}$$

で表す。$\mathrm{w} = (w_0, w_1, \dots, w_N)^T$、$x = (1, \widehat{x^T})^T$とする。

$a = w^T x$とすれば、$f(x)$はロジスティック関数となる。

$$f(x) = \sigma(a) = \frac{1}{1 + \exp(-a)} = \frac{\exp a}{1 + \exp a}$$

⇒非線形である

6.4 ロジスティック回帰

- ロジスティック関数の逆関数である**ロジット関数**は

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

オッズ(事後確率の比)は、

$$\frac{p(1|x)}{1-p(1|x)} = \frac{p(1|x)}{p(0|x)} = \exp(w^T x)$$

で表される。xの中のx_1が1増えた状態$\tilde{x}$を考えたとき

$$\frac{\frac{p(1|\tilde{x})}{1-p(1|\tilde{x})}}{\frac{p(1|x)}{1-p(1|x)}} = \exp w_1$$

となる。

→ x_1が1単位分増えるとオッズが$\exp w_1$増加

6.4 ロジスティック回帰

- **6.4.3 パラメータの最尤推定**

2クラスロジスティック回帰モデルのパラメータの最尤推定

モデルの出力:t tが1の確率:$p = \alpha$ tが0となる確率:$p = 1 - \alpha$

$$f(t|\alpha) = \alpha^t(1-\alpha)^{1-t} \ (\mathrm{t} = 0\mathrm{or}1)$$

上式よりN回の試行に基づく尤度関数は、

$$L(\alpha_1, \ldots, \alpha_N) = \prod_{i=1}^{N} f(t_i|\alpha_i) = \prod_{i=1}^{N} \alpha_i^{t_i}(1-\alpha_i)^{(1-t_i)}$$

となる。この対数を取り、負の符号を加えた尤度関数

$$L(\alpha_1, \ldots, \alpha_N) = -\ln L(\alpha_1, \ldots, \alpha_N) = -\sum_{i=1}^{N} (t_i ln a_n + (1-t_i)\ln(1-\alpha_i))$$

⇒**交差エントロピー型誤差関数**(cross-entropy error function)

6.4 ロジスティック回帰

- $$\alpha_i = \sigma(x_i) = \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)}$$

を代入して整理すると

$$L(\alpha_1, \ldots, \alpha_N) = L(w) = -\sum_{i=1}^{N} (t_i w^T x_i - ln(1 + \exp(w^T x_i)))$$

が得られる。

↓

この交差エントロピー型誤差関数を

最小にするパラメータwを得ること　⇒**最尤推定法**

→対数尤度関数をwで微分し、その式が0となるwを求める。

6.4 ロジスティック回帰

- **6.4.4 多クラス問題への拡張と非線形変換**

  拡張手段…各クラスごとに、$a_k = w_k^T x \ (k = 1, \ldots, K)$を求め、事後確率を

$$P(C_k|x) = \pi_k(x) \frac{\exp a_k}{\sum_{j=1}^{K} exp a_j} \text{(ソフトマックス関数)}$$

  で計算して、最大事後確率を与えるクラスに分類する

  うまく分離できない場合

→入力ベクトルxを非線形関数$\varphi\ (\ )$で、$M+1$次元データに変換すると分離できる可能性がある

$$\varphi\ (\ ) \Rightarrow \text{非線形基底関数}$$