目次

★12　パターン認識と機械学習 (3) — 識別のための教師あり学習の例

「識別」のための教師あり学習の手法を取り上げる．

★12.1　素朴なアイデア — 最短距離法

最も単純な識別の手法は，どのクラスに所属するかがわかっている見本データ (これを**プロトタイプ** (☆ 1) という) をクラス毎に 1 つずつ用意しておき，未知のデータが入力されたら，そのデータと最も「近い」プロトタイプを探して，そのプロトタイプと同じクラスに所属すると判断する，というものである．これは，次節で解説する最近傍法の特殊な場合といえる．

☆ 1) プロトタイプ: prototype.

Q1. (身長 [cm], 体重 [kg]) の二次元のデータが与えられたときに，上記の手法で「人間」と「ほげ星人」を識別してみよう．「人間」のプロトタイプは $(170, 65)$，「ほげ星人」のプロトタイプは $(100, 100)$ とする．また，2 次元ベクトル間のユークリッド距離で「近さ」を測ることにする．次の 2 人はどちらに識別されるか．A さん: $(135, 45)$，B さん: $(135, 82)$.

Q2. Q1 の場合，(身長, 体重)-平面は「人間」に所属する点の集合と「ほげ星人」に所属する点の集合に二分される．その境界はどのような図形になるか．ヒント: 2 点からの距離が等しい点の集合は何？

Q2 の結果が示すように，識別とは，入力データの空間を与えられたクラスのそれぞれに対応する領域に分割することであるといえる (☆ 2)．異なる 2 つのクラスに対応する領域の境界を**決定境界 (識別境界)** (☆ 3) という．上記の例では決定境界は直線であるが，識別手法によっては曲線 (より高次元のデータに対しては平面，曲面) となることもある．

☆ 2) この授業では説明しないが，入力の各点を確定的に 1 つのクラスに所属させるのではなく，各クラスへの所属確率を算出するような方法もある．したがってこの記述はちょっと説明不足．

☆ 3) 決定境界: descision boundary.

★ 12.2 最近傍法と k 最近傍法

上述の手法を一般化すると，1 クラスあたり複数のプロトタイプを用意する手法が考えられる．このような手法は，**最近傍法** (☆ 4) と呼ばれる．最近傍法の手順は，次のようになる．

1. 学習データ $\{(\boldsymbol{x}_n, y_n)|n = 1, 2, \ldots, N\}$ を用意する．ここで y_n は，プロトタイプ $\boldsymbol{x}_n$ の所属するクラスを表す (クラスラベルという)．例えば，‘ほげ’, ‘ふが’, ‘へな’ という 3 つのクラスを識別する問題の場合，$y_n \in \{$‘ほげ’,‘ふが’,‘へな’$\}$ とすればよい．
2. 所属クラスが未知のデータ $\boldsymbol{x}$ が与えられたら，最も「近い」プロトタイプの番号 n^* を求める (☆ 5)．ここで，$d(\boldsymbol{x}, \boldsymbol{y})$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の距離を表す (☆ 6)．

$$n^* = \underset{n=1,2,\ldots,N}{\operatorname{argmin}} \; d(\boldsymbol{x}, \boldsymbol{x}_n) \tag{1}$$

3. データ $\boldsymbol{x}$ をクラス y_{n^*} (n^* 番目の学習データの所属クラス) に識別する．

最近傍法はさらに，未知データに最も近いプロトタイプを 1 つ選ぶかわりに，最も近い k 個を選んで多数決をとる，というように一般化することができる．このような手法は，k **最近傍法**と呼ばれる (☆ 7)．**最近傍法**は，$k = 1$ の場合に相当する．図 1 に，2 次元のデータを k 最近傍法で識別した例を示す．

最近傍法や k 最近傍法は，与えられた学習データを全て記憶しておき，未知データと全ての学習データとの距離を計算する手法である．そのため，識別に必要な計算コスト・記憶コストともに高くつく．しかし，コンピュータ性能の向上により，近年では大規模データに対しても実用されることもある．また，学習データが大量に得られる場合には，その全てをプロトタイプとはせず，教師なし学習手法であるクラスタリング等の手法によってプロトタイプを選別することもある．

☆ 4) 最近傍法: nearest neighbor method. NN 法とも．画素値の補間の話で同じ名前が出て来たが，別もの．

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

☆ 6) ユークリッド距離以外の距離を用いる場合もあるので，このように一般化して書いている．データが数値で表せないような問題でも，それらの間の距離さえ定義できれば最近傍法は用いることができる．

☆ 7) k 最近傍法: k-nearest neighbor method. k-NN 法とも．

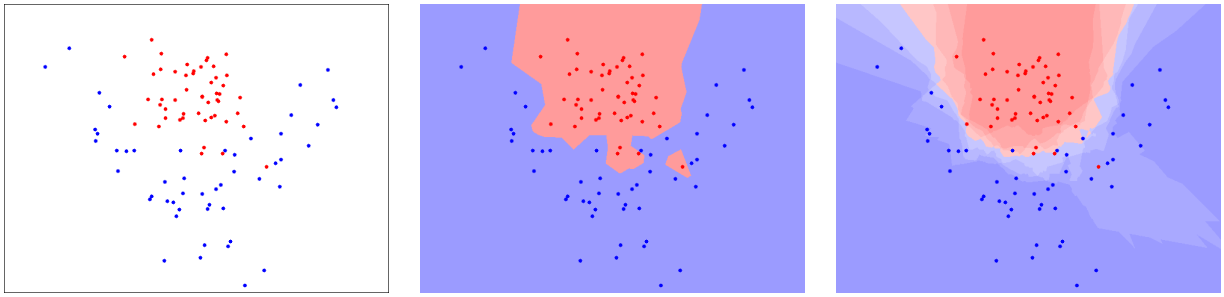


図 1: k 最近傍法による識別．左: 学習データ．赤と青の 2 クラス．中: $k = 1$ での識別結果に応じて領域を塗り分けたもの．右: $k = 7$ での結果．色の濃さは，k 票中の多数票の多さに対応している．

★ 12.3　ロジスティック回帰

識別のための手法の別の例として，ロジスティック回帰 (☆ 8) を紹介する．名前に「回帰」とあって紛らわしいが,「識別」のための手法である (☆ 9).

☆ 8) ロジスティック回帰: logistic regression.

☆ 9) この手法の背景には，その他の多くの機械学習手法と同様に，データ等の確率的・統計的性質を考慮した問題設定や議論があるのだが，この授業では説明を省略する．

★ 12.3.1　2 クラス識別問題の場合の定式化

簡単のため，まずは識別すべきクラスの数が 2 に限定された場合を考える．2 つのクラスのうち一方を ‘positive’（正）クラス，他方を ‘negative’（負）クラスと呼ぶことにする．学習データは $\{(\boldsymbol{x}_n, y_n)|n = 1, 2, \ldots, N\}$ という形とする．$\boldsymbol{x}_n$ は D 次元のデータ（特徴ベクトル）である．y_n は，$\boldsymbol{x}_n$ の所属クラスの正解を表し，$\boldsymbol{x}_n$ が positive クラスに属すべきものなら $y_n = 1$，さもなくば $y_n = 0$ とする．

ここで，$(D + 1)$ 個のパラメータ $w_0, w_1, \ldots, w_D$ を持つ次のような関数を考える．

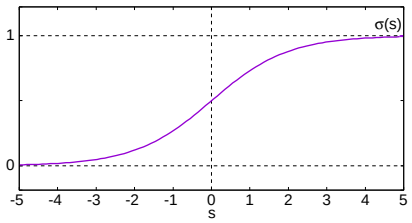
$$f(\boldsymbol{x}; w_0, w_1, \ldots, w_D) = \frac{1}{1 + \exp\left(-\left(w_0 + \sum_{d=1}^{D} w_d x_d\right)\right)} \tag{2}$$

$s = w_0 + \sum_{d=1}^{D} w_d x_d = w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ とおくと，式 (2) は

$$\sigma(s) = \frac{1}{1 + \exp(-s)} \tag{3}$$

という形をしている．この関数 $\sigma(s)$ のことを**シグモイド関数** (☆ 10) という．式の形と下のグラフからわかるように，この式は任意の s に対して 0 より大きく 1 より小さい値をとる．そこで，式 (2) の $f(\boldsymbol{x})$ の値が,「$\boldsymbol{x}$ は positive クラスに所属するものである」ということの確信度を表すと考えることにする．

☆ 10) シグモイド関数: sigmoid function. 次回の話題であるニューラルネットワークでも登場します．



このように考えると，学習データのうち $y_n = 1$ であるものについては positive クラス所属なのだから $f(\boldsymbol{x}_n)$ が 1 に近くなるように，$y_n = 0$ であるものについては逆に $f(\boldsymbol{x}_n)$ が 0 に近くなるようにうまくパラメータを調節すれば，式 (2) の値で 2 クラスの識別ができそうである．未知のデータ $\boldsymbol{x}$ に対しても，例えば $f(\boldsymbol{x}) > \frac{1}{2}$ なら positive クラス，さもなくば negative クラスと判断すればよい．

そこで，学習データからパラメータ $w_0, w_1, \ldots, w_D$ を定めるために，パラメータの「悪さ」を表す関数を定義する．ロジスティック回帰では，次式で表される**交差エントロピー** (☆ 11) を用いる．

☆ 11) 交差エントロピー: cross entropy. これが何者かは，この授業では説明しません．パターン認識や情報理論を勉強するとわかるかも．

$$H(w_0, w_1, \ldots, w_D) = \sum_{n=1}^{N} h_n \tag{4}$$

$$h_n = -y_n \log z_n - (1 - y_n) \log(1 - z_n) \tag{5}$$

ただし， $z_n = f(\boldsymbol{x}_n)$ である．この交差エントロピー H がなるべく小さくなるようなパラメータが「良い」と考えて，H を最小化するパラメータを探す．これが，2 クラス問題の場合のロジスティック回帰の考え方である．

★ 12.3.2　勾配法によるパラメータの逐次修正

上記のような問題設定は，最小二乗法による回帰の場合にも考えた．最小二乗法の場合，学習データに対する二乗誤差の和で誤差関数を定義し，それを最小にするパラメータを求めたのだった．このように，最小二乗法もロジスティック回帰も，目的とする関数（二乗誤差や交差エントロピー）を最小化する解（特定のパラメータの値）を探す問題である，というところは共通である．このような問題は，**最適化問題** (☆ 12) と呼ばれる．

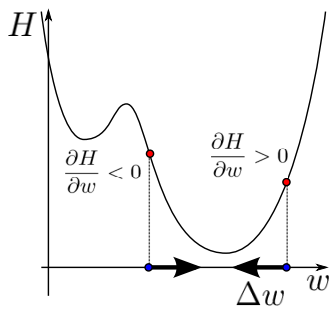
最小二乗法の場合，最適化問題を解くことは比較的容易である．誤差関数をパラメータについて微分して $\mathbf{0}$ とおくと線形の連立方程式が得られるので，それを解けばよいのだった．しかし，ロジスティック回帰の場合，H の微分を $\mathbf{0}$ とおいて得られる連立方程式は非線形であり，簡単に解くことはできない．それでも，ロジスティック回帰の場合，H のパラメータに関する微分が求まるので，それを利用した最適化手法である**勾配法** (☆ 13) が使える．ここでは，その最も単純な方法の一つである**最急降下法** (☆ 14) を説明し，これによってパラメータを逐次修正していく学習アルゴリズムを構成できることを示す．

2 クラス問題のロジスティック回帰における $(D+1)$ 個のパラメータのうちの一つを w と表して考える．H を最小にするパラメータ w の値を求めるための最急降下法の手順は次のようになる．

1. $w(t)$ の初期値（$t=0$ における値）を適当に定める．
2. ステップ t における値 $w(t)$ を用いて微係数 $\frac{\partial H}{\partial w}\big|_{w=w(t)}$ を計算し，次のステップ $t+1$ における値 $w(t+1)$ を次式により求める．

$$\begin{cases} w(t+1) = w(t) + \Delta w \\ \Delta w = -\eta \, \frac{\partial H}{\partial w}\big|_{w=w(t)} \end{cases} \tag{6}$$

   Δw は w の修正量を表す．η は正の小さな定数である．
3. H の値が十分小さくなっていれば終了，さもなくば 2. を繰り返す．



上図からわかるように，最急降下法は「現在地での傾きを調べ，下り方向にちょっとだけ進む」ことを繰り返す手法である．初期値によっては極小解に陥って最小解にたどり着けないこともある．

☆ 12) 最適化: optimization. 負号を付けるだけなので，最大化でも同じこと．機械学習，パターン認識，コンピュータビジョンではよく出てくるが，それ以外の幅広い分野で頻出の問題である．

☆ 13) 勾配法: gradient method.
☆ 14) 最急降下法: steepest descent method.

上記のロジスティック回帰の場合，微係数を具体的に計算してみると，

$$\frac{\partial H}{\partial w_d} = \sum_{n=1}^{N}(z_n - y_n)x_{n,d} \qquad (d = 0, 1, \ldots, D) \tag{7}$$

となる．ただし，$x_{n,d}$ はベクトル $\boldsymbol{x}_n$ の d 番目の要素（$d = 0$ のときは 1 とみなす）である．これを式 (6) に当てはめれば，学習アルゴリズムが得られる．

以下の Q は，式 (7) を導出するためのヒントである．

Q3. シグモイド関数の微分は，シグモイド関数自身を用いて

$$\frac{d\sigma(s)}{ds} = \sigma(s) \times \text{hoge} \tag{8}$$

と表せる．ただし，hoge は $\sigma(s)$ を用いた式である．hoge を求めなさい．

Q4. 上記の結果を利用して，$d = 1, 2, \ldots, D$ の場合の

$$\frac{\partial z_n}{\partial w_d} = \frac{\partial}{\partial w_d}\sigma\left(-\left(w_0 + \sum_{d=1}^{D} w_d x_{n,d}\right)\right) \tag{9}$$

を $x_{n,d}$ と z_n のみを用いた式で表しなさい（$z_n = \sigma(\ldots)$，合成関数の微分…）．

Q5. 上記の結果を利用して，$d = 1, 2, \ldots, D$ の場合の

$$\frac{\partial}{\partial w_d}(y_n \log z_n) = y_n \frac{\partial}{\partial w_d} \log z_n \tag{10}$$

を $x_{n,d}, y_n$ と z_n のみを用いた式で表しなさい（やぱし合成関数の微分…）．

Q6. 上記の結果を利用して，$d = 1, 2, \ldots, D$ の場合の $\frac{\partial h_n}{\partial w_d}$ を $x_{n,d}, y_n$ と z_n のみを用いた式で表しなさい．また，$x_{n,0} = 1$ と考えると，その式が $d = 0$ の場合にも当てはまることを確かめなさい．

Q7. 上記の結果を利用して，式 (7) が成り立つことを示しなさい．

余談だよん

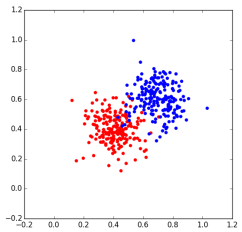
識別に関する手法は，画像や音声などのパターン情報を対象としてよく研究されてきたため，このような研究分野は**パターン認識** (☆ 15) と呼ばれている．この授業で取り上げることができたのは，この分野で知られている代表的な手法の中のほんの一部である．パターン認識の分野では，様々な意味で確率を取り入れたデータの扱い／アルゴリズムの構成をすることが一般的であるが，この授業では全く取り上げなかった．興味のある人は，次のような参考書を参照してほしい．

☆ 15) パターン認識: pattern recognition.

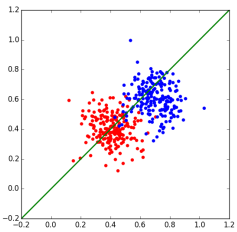
- 「わかりやすいパターン認識」 オーム社，ISBN4–274–13149–1
- 「パターン認識と機械学習 上下」シュプリンガー・ジャパン，ISBN978–4–431–10013–3, 10031–7

★ 12.3.3　2 クラスデータのロジスティック回帰の例

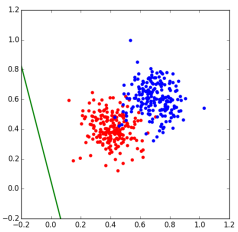
乱数を使って 2 次元の人工データを生成し，ロジスティック回帰によってそれらを 2 クラスに識別する実験を行った結果を示す．以下の図の赤と青の点は 2 つのクラスの学習データを表している．緑色は，$f(x_1, x_2; w_0, w_1, w_2) = \frac{1}{2}$ を満たす点 (x_1, x_2) の集合，すなわち赤クラスと青クラスの識別境界を表している．式 (2) からわかるようにこれらの点は $w_0 + w_1 x_1 + w_2 x_2 = 0$ を満たすので，境界は直線である．



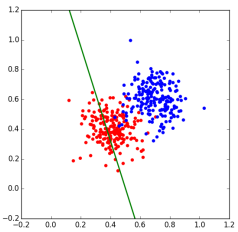
学習データ



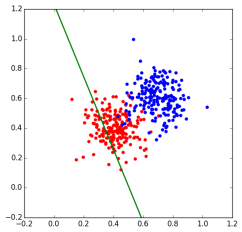
初期値 ($H = 276.4$)



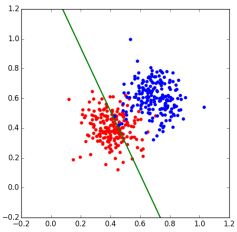
学習 1 回 ($H = 267.7$)



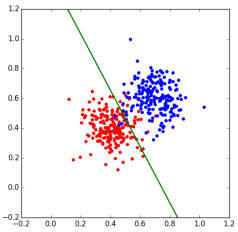
2 回 ($H = 260.0$)



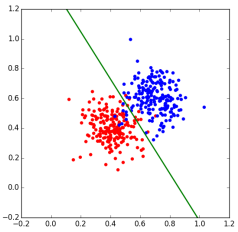
3 回 ($H = 252.9$)



5 回 ($H = 239.8$)

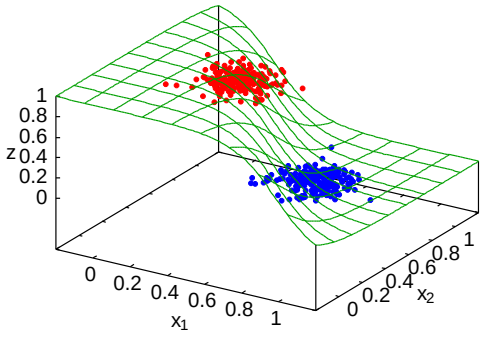


10 回 ($H = 212.5$)



100 回 ($H = 86.2$)

以下の図は，100 回学習後の $z = f(x_1, x_2)$ の表す曲面を可視化したものである．赤クラス青クラスのデータを，それぞれ $z = 1$ および $z = 0$ の平面上に重ねて表示してある．



★ 12.3.4　クラス数が 3 以上の場合のロジスティック回帰

上述の定式化では 2 クラス問題しか扱うことができないが，3 クラス以上の識別ができるように拡張することは容易である．ただし，紙面と時間の都合で，具体的な定式化や実験結果については省略する．