第二章
識別規則と学習法の概要

11T4066Y
松村　拓真

本章の内容

券売機を例として…

- 同じ100円硬貨でも特徴ベクトル(重さ、透磁率、サイズ)に違いがあるため、同じ100円に対応付けるため、多量の100円硬貨を用いてデータを収集する⇒**学習データ**

- 学習データには含まれない100円硬貨や、怪しげな100円硬貨が使われても正しく識別する能力⇒**汎化能力**

識別規則と学習法の概要の解説後、
学習データを用いた汎化能力について推定する方法を述べる

2.1識別規則と学習法の分類

2.1.1識別規則の構成法

識別規則…入力データxからクラス$C_i \in \Omega = \{C_1, \ldots C_k\}$への写像

写像の実現方法⇒関数値の正負、ダミー変数表現、事後確率の最大値、
決定木の終端ノードなど様々

2.1識別規則と学習法の分類

代表的な識別規則の構成法

- 事後確率による方法：パターン空間に確率分布を仮定し、事後確率が最大のクラスに分類(ベイズの最大事後確率法)
- 距離による方法：入力ベクトルxと各クラスの代表ベクトルとの距離を計算し、一番近い代表ベクトルのクラスに分類(最近傍法)
- 関数値による方法：関数f(x)(識別関数という)の正負、又は最大値でクラスを決定(パーセプトロン型学習回路、サポートベクトルマシン)
- 決定木による方法：識別規則の真偽に応じて次の識別規則を順次適応し、決定木の形でクラスを決定

2.1識別規則と学習法の分類

2.1.2教師付き学習

識別規則は入力データからクラスへの写像を関数y=f(x)を用いて表現する

写像の性質を決めるパラメータをwで表すとする

ex)第六章で述べる2クラス問題の線形識別関数の場合、識別規則は

$$y = f(x; w) = w_1 x_1 + \cdots + w_d x_d = w^T x \ldots (2.1)$$

上式のように、パラメータwと入力ベクトルxの線形計画(内積)を用いて表現される

2.1識別規則と学習法の分類

　識別規則の学習は、このf()を学習データを用いて決めることであり、その目的は、パラメータwを調節して入力データが正しいクラスに対応する関数値を出すようにすること

　学習するためには、入力データとそのクラスを指定したデータ(**教師データ**)を対にした学習データが必要

　2クラスの場合(識別クラスは関数値yの正負で決めるとする)

$$(x_i, t_i) \in \{-1, +1\} \ldots (2.2)$$

　3クラス以上の場合はダミー変数表現を用いて、例えば

$$t = (0,1,0,0,0,0,0,0,0,0)^T$$

　のような教師データにする(K対1符号化という)

2.1識別規則と学習法の分類

入力データと教師データの対を、$(x_i, t_i)(i = 1, \ldots, N)$の様に表す
学習に用いられる全ての対の集合を**学習データセット**といい、D_Lで表す

学習データセットの関数としてwを得る方法や、学習データセット中の学習
学習データを一つずつ用いて、wを少しずつ修正していく方法などがある
修正が完了した識別関数は、**テストデータセット**を用いて性能評価を行う

⇒以上のような学習法を**教師付き学習**と呼ぶ

2.1識別規則と学習法の分類

2.1.3教師付き学習と線形回帰

教師は、入力が所属するクラスを2値($\{-1,+1\}$等)で指定する

2値でなく、任意の関数値が与えられる場合、識別関数は入力xに対して与えられた関数値を出力するように学習が行われ、そのために、識別関数f(x)は与えられた関数値を近似できるだけの能力を持つ必要がある

⇒このような問題を**関数近似(回帰)**という(線形関数の場合は**線形回帰**)

ex)ある月の平均降雨量と作物の収量の関係

→統計データを元に線形回帰で求めておくと、その年の平均降雨量から収量を予測可能(与えられる作物の収量は**被説明変数**、関数の引数として与えられる平均降雨量は**説明変数**と呼ばれる)

2.1識別規則と学習法の分類

2.1.4教師なし学習

入力データ間の距離や類似度、統計的な性質に基づいて、クラスを自動的に生成(**クラスタリング**)することが主目的になる

⇒**教師なし学習**、もしくは**自己組織型学習**という(詳細は10章で述べる)

最近では一部のデータのみ教師をつけ、他は教師なしで学習を行う**形質導入学習**という学習法も提案されている

2.2汎化能力

学習⇒学習データに対する識別関数の出力値と教師データとの誤差を最小にするように、識別関数のパラメータを調節すること

しかし…未知データに対してうまくはたらくか分からない

↓そこで

未知データに対する動作をテストデータに対する誤り確率という形で予測する

未知のデータに対する識別能力…**汎化能力**

汎化能力の誤差…**汎化誤差**

2.2汎化能力

2.2.1学習データとテストデータの作り方

手元のデータを分割し、学習データセットD_LとテストデータセットD_Tを作成

ex)券売機の設計…10000枚の100円硬貨を

$D_L = 8000, D_T = 2000$ に分割する

- d次元空間内(dは特徴)の分布をp_Lとp_Tで表す

  ⇒平均値や分散などが計算可能

- D_Lで設計し、D_Tでテストしたときの誤り率をε(p_L, p_T)で表すとする

2.2汎化能力

- 国内すべての100円硬貨の集合を**母集合**、そのd次元特徴の分布をpで表し、**真の分布**と呼ぶ
- p_Lとp_T の各特徴の平均値や分散がpと同じになるとは限らない。

→偏り**(バイアス)**という

- 真の分布を元に学習データを設計し、テストしたときの誤り率を**真の誤り率**$\varepsilon(p\ ,p)$とする
- 学習データとテストデータに同じデータを用いてテストをしたときの誤り率を**再代入誤り率**$\varepsilon(p_L,p_L\ )$で表す

2.2汎化能力

- 学習用とテスト用に分割する代表的な方法

(1)ホールドアウト法

データを二つに分割し、一方を学習(p_L)、もう一方をテストに使用(p_T)

⇒ホールドアウト誤り率$\varepsilon(p_L, p_T)$

- 有限のデータを分割するため、互いの割合によって、学習精度と性能評価の精度に偏りが発生する可能性…

→手元に大量にデータがないと、高い性能評価を得られない

2.2汎化能力

(2)交差確認法(交差検証法、交差妥当化法)

- ホールドアウト法を補う方法
- 手元の各クラスのデータをそれぞれm個のグループに分割し、$m-1$個のグループのデータで学習、残りの一つのデータでテスト

↑これをm回繰り返し、誤り率の平均を性能予測値とする

- i番目のグループ以外で学習し、i番目のグループでテストしたときの誤り率ε_{-i}は

$$\frac{1}{m}\sum_{i=1}^{m}\varepsilon_{-i}$$

となる

⇒全てのデータを学習とテストに用いるため、良い性能予測を行える

(分割によっては偏りが出るので、分割を変えて繰り返し、平均から 誤識別率を予測する必要がある)

2.2汎化能力

(3)一つ抜き法(ジャックナイフ法)

- 交差確認法を、データの数=グループの数にしたもの
- 一つを除いたすべてのデータで学習、除いた一つでテスト(データの数だけ繰り返す)
- 組み合わせは一つのため、1回で十分

2.2汎化能力

(4)ブートストラップ法

- 再代入誤り率のバイアスを補正するために利用
- N個のデータを用いて再代入誤り率$\varepsilon(N, N)$を得る

→誤識別率は真の誤識別率より低く出るはず(差をバイアスとする)

⇒バイアスをN個のデータからN回**復元抽出**を行って作った**ブートストラップサンプル**(N^*)を用いて修正

2.2汎化能力

- N^*を学習データ、テストデータに用いて得られた誤識別率を$\varepsilon(N^*, N^*)$で表すとすると、バイアスは

$$\mathrm{bias} = \varepsilon(N^*, N^*) - \varepsilon(N, N)$$

で推定

- 複数のブートストラップサンプル(最低で50程度)を生成し、得られる誤識別率の差の平均値で、バイアスを推定する

→このときの推定値は

$$\varepsilon = \varepsilon(N, N) - bias$$

⇒クラスが複数ある場合はブートストラップサンプルを作成し、バイアスを推定するのが良い

2.2汎化能力

2.2.2**汎化能力の評価法とモデル選択**

　学習データによるパラメータ調整とテストデータによる誤り率の評価でも誤りが目標を下回らない→識別関数を変更する

　誤り率が最少となるパラメータの選択方法…**モデル選択**

2.2汎化能力

1次,3次,6次,10次の多項式で近似した場合…

- 1次関数の近似はデータから大きく外れている(バイアスが大きい)が、近似した直線はほとんど同じ(分散が小さい)
- 3次の多項式は信号線分を正確に近似(バイアス、分散ともに小)
- 6次の多項式ではノイズを追うようになる
- 10次の多項式は完全にノイズに追従する

→関数が複雑になるほど学習データに対する近似能力は上がるが、ノイズ部分を近似するようになってしまう(バイアスは小さくなるが、分散は大きくなる)

バイアスと分散の反比例の関係…**バイアス・分散トレードオフ**

⇒汎化能力を決める大きな要因

2.2汎化能力

多項式の次数を大きくするにつれ、学習誤差は単調に減少するが、テストデータに対する誤差は3次から少しずつ増加している(**過学習**という)

→汎化能力は3次の多項式が最も高い

(バイアスと分散が均衡している)

2.2汎化能力

- 識別関数の構成でもモデル選択と同様に考えられる

　→式(2,1)より、$y = f(x; w)$の$f()$の形やwの要素の数を変更しながら、交差確認法やブートストラップ法により汎化誤差を推定する(データの分布を統計モデルで記述しなくてもいい)

- データの分布に統計モデルを仮定する場合…

　→解析的に汎化誤差を評価でき、その結果を利用してモデル選択が可能

　具体的な手法

◆赤池の情報量基準(AIC)

◆ベイズ情報量基準(BIC)

◆最小距離基準(MDL)　etc.