はじめてのパターン認識
第2章　識別規則と学習法の概要

大内克之

識別規則の構成法

識別規則とは、入力データxからクラスC_iへの写像のこと。
代表的な識別規則の構成法には以下のものが挙げられる。

- 事後確率による方法
- 距離による方法
- 関数値による方法
- 決定木による方法

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する。

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が近い代表ベクトルのクラスに分類する。

識別規則の構成法

関数値による方法

　関数f(x)の正負、あるいは最大値でクラスを決める。識別のために用いられる関数f(x)を識別関数という。

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用し、決定木の形でクラスを決める。

教師付き学習

識別規則の学習は、写像を表す関数であるu=f(x)を学習データを用いて決めることである。

写像の性質を決めるパラメータをωで表すと、識別規則は、

$$y=f(x;w)=\omega_1 x_1+\ldots+\omega_d x_d=\omega^T x$$

のように、パラメータωと入力ベクトルxの線形関数を用いて表現される。

学習の目的はこのパラメータωを調節することである。

教師付き学習

学習をするためには、「入力データ」と、そのクラスを指定したデータである「教師データ」を対にした学習データが必要となる。学習データの集合を学習データセットという。

ωを改良するには、

- 学習データセットの関数としてωを得る方法
- 学習データを一つずつ用いて少しずつ修正していく方法

などがある。後者の場合、同じ学習データを何度も使用して学習していく。学習を終えたら、学習に使用しなかったテストデータセットを用いて性能評価を行う。

教師付き学習と線形回帰

ここまでの学習法は、教師がいるので教師付き学習と呼ばれる。

教師として任意の関数値が指定された場合、識別関数には関数値を近似できる能力が必要となる。この能力は識別関数の複雑さに対応し、このような問題は関数近似（回帰）と呼ばれる。線形関数での近似する場合、線形回帰と呼ばれる。

汎化能力

学習データ未知のデータに対応するために、テストデータを用いた性能評価が行われている。

未知のデータに対する識別能力を汎化能力といい、その誤差を汎化誤差という。

手元にあるデータを学習データとテストデータに分ける方法には、以下のようなものがある。

・ホールドアウト法

・交差確認法

・一つ抜き法

・ブートストラップ法

学習データとテストデータ

ホールドアウト法

手元のデータを二つに分割し、一方を学習に、もう一方はテストのために取り置いておく方法。

学習用を増やすと性能評価の精度が悪くなり、テスト用を増やすと学習そのものの精度が悪くなる。

交差確認法

ホールドアウト法の欠点を補うものとしてよく使用される。

手元のデータをm個のグループに分割し、一つをテストに使い、残りを学習する。これをm回繰り返す。すべてのデータを学習とテストに利用するので、よい性能予測を行えるが、分割によって偏りが生じる可能性がある。

学習データとテストデータ

一つ抜き法

交差確認法で、グループとデータの数を等しくした場合をいう。ジャックナイフ法とも呼ばれている性能予測の基本的な方法である。

ブートストラップ法

学習データを使ってテストを行い誤り率を予測すると、再代入誤り率が得られる。真の誤り率と再代入誤り率の差をバイアスという。このバイアスをブートストラップサンプルというものを用いて修正する。

汎化能力の評価法とモデル選択

学習データでパラメータの調節を終えても誤り率が目標より小さくならない場合、識別関数を変えることになる。その方法には、線形識別関数を非線形に変える方法や、パラメータの数を変える方法がある。

パラメータの数を変え、誤り率が最も小さくなるパラメータを選択する方法をモデル選択という。

教師なし学習

教師がいない学習では、入力データの距離や類似度などに基づき、クラスを自動的に生成する必要がある。これをクラスタリングという。

教師なし学習では、クラスタリングが主目的となる。