知能機械情報学 課題 2

03223008 坂本光皓

2024 年 7 月 15 日

1 AdaBoost

1.1 概要

AdaBoost とは適応的ブースティング（Adaptive Boosting）の略であり，弱学習器を増強することで精度の高い強学習器を構成するブースティングの一種である．この手法ではそれぞれの識別器を, 前段の識別器の性能に応じて重み付けしたパターンを用いて学習させる．すなわち, ある識別器で誤識別されたパターンには大きな重みが付与され，後段の識別器の学習において重視される．すべての識別器が学習された後，各識別器の信頼度に応じた重み付け多数決で最終的な識別を行う．

1.2 データセット

実装したアルゴリズムの評価には Iris データセットを用いる．これは 3 種類のアヤメの花びら（petal），がく片（sepal）それぞれの長さと幅という特徴量から構成されている． このデータセットはサンプル数が 150 件と少ないが，異なる種類の植物を正確に分類するための初歩的な手法の理解に役立てることができる．

1.3 性能評価

Iris データセットのうち versicolor，virginica の 2 クラスについて，花びらの長さおよび幅を特徴量として AdaBoost で学習を行う．なお，弱学習器には決定株を用いた．

Fig. 1 に識別器の数を 10 としたとき得られた決定境界を示す．正解率は 98 %，学習時間は 22ms であった．ハイパーパラメータである識別器の数は 10 より大きくしても性能の向上が見られなかったため，このデータセットに対しての適切な値であるといえる．

また，データの分布から明らかなように 2 クラスは線形分離不可能であるが，AdaBoost はこのような非線形分類問題にも適用可能である．

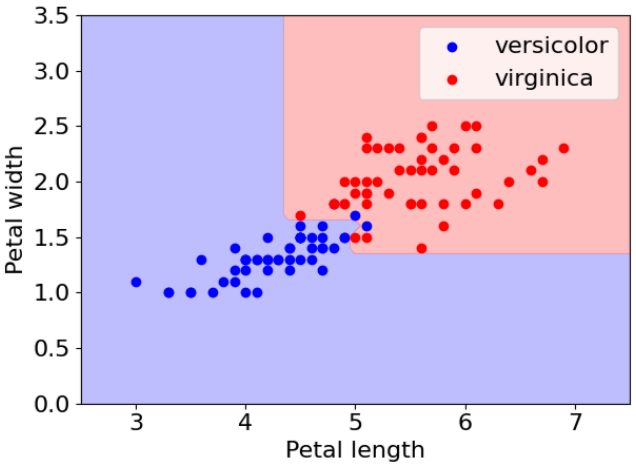


Fig. 1: 学習で得られた決定境界

2 Kernel K-means

2.1 概要

Kernel K-means は教師なしクラスタリング手法の一つである K-means をカーネル法を用いて拡張したものである．パターン $\boldsymbol{x}_n$ の特徴空間における距離をカーネル関数 $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を用いて

$$d(\phi(\boldsymbol{x}_n), \boldsymbol{\nu}_k) = k(\boldsymbol{x}_n, \boldsymbol{x}_n) - \frac{2}{|\mathcal{C}_k|}\sum_{j\in\mathcal{C}_k} k(\boldsymbol{x}_n, \boldsymbol{x}_j) + \frac{1}{|\mathcal{C}_k|^2}\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{C}_k} k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{1}$$

と定めることで，陽に特徴量 $\phi(\boldsymbol{x}_n)$ や特徴空間における代表点を計算せずクラスタを分離できる．

2.2 データセット

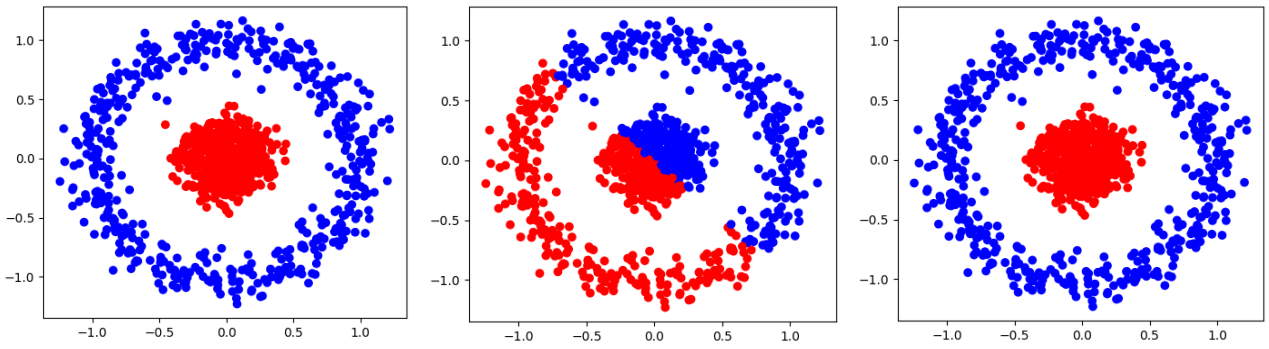
実装したアルゴリズムの評価には，円状に生成したノイズを含むデータを用いる．内側と外側の円の 2 つの領域にデータが分布しているため線形分離不可能であり，Kernel K-means や Kernel SVM，ニューラルネットワーク等の非線形分類器のテストに適する．

2.3 性能評価

Fig. 2 a にサンプル数が 1000, ガウシアンノイズの標準偏差が 0.1, 外径に対する内径の比が 0.2 となるように生成したデータを示す．これに対し，線形カーネル $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_j$ およびガウスカーネル $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$ を用いたクラスタリングを行ったところ，Fig. 2 b, c のような結果が得られた．分類に要した時間は線形カーネルが 68ms，ガウスカーネルが 115ms であった．なお，各パターンに割り当てるクラスタは完全に収束しなかったため，1000 個のパターンのうち割り当て先が変化したものが 50 個未満の場合に収束したとみなして処理を終了している．

線形カーネルにはハイパーパラメータはなく，線形分類器となるため Fig. 2 b のように内側と外側の区別ができていない．一方，ガウスカーネルはハイパーパラメータとして決定境界の分散（の逆数）γ があり，値を試行錯誤的に調整することでクラスタリング結果が Fig. 2 c のように入力データと一致した．ただし，γ に対する結果の変化が激しいため，膨大なデータに対するパラメータの調整は相当の時間が必要であると考えられる．

また，適切なカーネル関数を用いることで通常の K-means では分離できないデータに対してもクラスタリングを行うことができることがわかる．



a) 入力データ　b) 線形カーネル　c) ガウスカーネル

Fig. 2: Kernel K-means によるクラスタリング結果