

平成 22 年度情報生命科学演習，交差確認法

Tsuyoshi Kato

平成 22 年 6 月 30 日

バイオインフォマティクスにおいて，しばしば予測を行う場面に遭遇する．これまでに計算機科学の分野で多数のパターン識別器が開発されてきた．多くのパターン識別機では，ハイパーパラメータと呼ばれるパラメータがあり，最適な値はデータに依存する．ここでは，最適な値を決める方法を学ぶ．

1 データセット

データセットは

<http://www.net-machine.net/~kato/kemba-svm1-ts/software/ex-kemba-svm1-070702.zip>

に含まれているものを使う．解凍すると

- k.X.csv
- k.y-true.csv

という 2 つのファイルが入っている．

k.X.csv には 100 次元の特徴ベクトルを持つ 72 例題が 100×72 の行列として，csv 形式で格納されている．このうち，20 例題のみのクラスラベルが分かっている．10 個が正例で，10 個が負例，52 個がクラス不明であり，この 52 個のクラスを予測したい．k.y-true.csv には，これらの情報が含まれている．72 個の数値が並んでおり，そのうち +1 が正例，-1 が負例，0 がクラス不明である．

2 性能評価

2 クラス識別器の識別性能を測る基準の一つに正解率がある．先に，50 個のクラスは不明と述べたが，実は分かっており，真のクラスラベルは

http://www.net-machine.net/~kato/kemba-svm1-ts/sel02_01/k.y_orac.csv

から得ることができる．しかし，このデータは性能評価だけのために提供したもので，予測時には使ってはならない．

クラス未知の例題の予測結果と真のクラスラベルを照合し，どのくらいの割合の例題で正しく予測できたか評価する．

3 K -最近傍識別器

ここでは、 K -最近傍識別器を用いることとする。 K -最近傍識別器は次のような識別器である。未知の例題から各訓練用例題への距離を測るものによって、もっとも近傍にある K 個の訓練用例題を特定する。 K 個の中で多数決を取り、多いほうのクラスを未知例題に対する予測結果とする。 K -最近傍識別器の場合、この K がハイパーパラメータとなる。

4 最適ハイパーパラメータの推定法

サポートベクトルマシンも K -最近傍識別器もハイパーパラメータを持っている。ハイパーパラメータの値を何らかの手段を使って選ばなくてはならない。そのための方法の一つが交差確認法 (cross validation) である。ここは一つ抜き交差確認法 (leave-one-out cross validation) により、ハイパーパラメータの値を決定する。

一つ抜き交差確認法とは次のような方法である。 ℓ 個訓練用例題がある場合、そのうち 1 個を未知の例題と見立てて、残りの $\ell - 1$ 個の訓練用例題を使って、その 1 個のクラスラベルを予測する。これを ℓ 個すべてに対して繰り返し、正解率を出す。もっとも正解率がよかったハイパーパラメータの値が推定値となる。

K -最近傍識別器は $K = 1, 3, 5, 7$ の 4 通りを試すこととする。

5 課題

交差確認法で求めたハイパーパラメータの推定値は、最適値とどれくらいずれていたか確かめよ。ただし、最適値と言うのは、ハイパーパラメータの値全通りで正解率を出し、正解率が最大になった値と交差確認法で求めた推定値とどれほどずれていたか、もしくは、ぴったり合っていたか確かめよ。