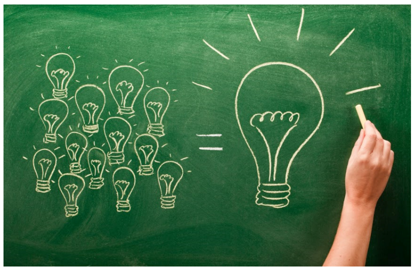
Bagging
Random Forest
Boosting

櫻井彰人

3人寄れば文殊の知恵



http://www.leanleader.org/2012/04/many-heads-are-better-than-one.html

アンサンブル法

- 決定木一本では性能が不十分
- しかし、学習は非常に速い
- では、多数の木を学習したらどうだろうか？
  - 勿論、同じ木を作ったのでは話にならない

Bagging

データをランダムに異なる仕方で分割し、決定木を作ると、異なる結果が得られる。 しかし、一般には分散が大きい。そこで、分散を低下させる方法が考えられた

Bagging: Bootstrap **agg**regat**ing**.

原理は: もし多数組の独立なデータ（サンプル）があれば、複数回の予測ができ、「多数の予測の平均をとると、分散が減る」ことを利用することができる

Bagging

つまり、s で s 番目のサンプルを表すとし、各予測器は $\hat{f}^s(x)$ を出力すると書くと:

$$\hat{f}_{avg}(x) = \frac{1}{S}\sum_{s=1}^{S} \hat{f}^s(x)$$

分類器のbootstrap

Bootstrap サンプルを作る（S個（何百個も））

それぞれに分類器を作ることにする（決定木の場合、枝刈り等の後処理はしない）

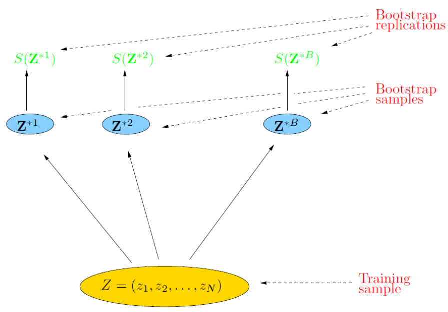
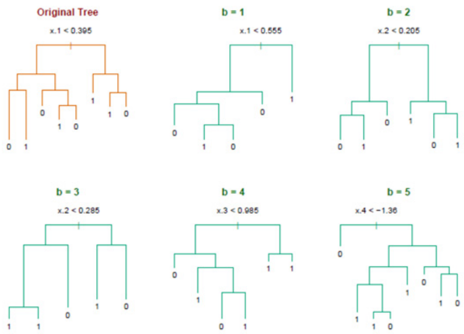


FIGURE 7.12. *Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(\mathbf{Z})$ computed from our dataset. B training sets $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ each of size N are drawn with replacement from the original dataset. The quantity of interest $S(\mathbf{Z})$ is computed from each bootstrap training set, and the values $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$ are used to assess the statistical accuracy of $S(\mathbf{Z})$.*

Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

決定木のbootstrap



Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

Out-of-bag エラー

- cross validation は不要?
- bootstrapでは復元抽出を行う。従って、それぞれのbootstrapサンプルには全ての観測値が含まれているわけではない。平均してその 1/3 は用いられない
- これらをout-of-bag samples (OOB)と呼ぶ
- 第 i 番目の観測値に対し、その応答の予測値を、その観測値がOOBの時の予測器（決定木）を用いて求める。これをn個の観測値に対して行う
- OOBのMSE (mean squared error) や分類誤差を求める

Bagging について

- Bootstrap時、通常は1種類の分類器を用いる
  - 決定木がよく使われる
- 並列化が容易
- 過学習しにくく、分散は低下する
- 1個の分類器に比べ、精度が向上する
- しかし、残念なことに、aggregation の結果得られたモデルの解釈は難しい。つまり、bagging は解釈可能性を捨てて、予測精度を向上させたことになる

Bagging の問題

それぞれの分類器は、identically distributed (i.d.)

分類器B個の平均の期待値は、個々の分類器の期待値と同じである

Bagging後のバイアスは、個々の分類器のバイアスと同じである

Baggingの問題

B 個のi.i.d. 確率変数の期待値の分散は、個々の分散が σ^2 であれば、 σ^2/B となる

しかし i.d. でしかなければ(identically であるが independently ではなければ) 、それぞれの間の相関係数を ρ とすると、分散は

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

B が増加すれば第2項は小さくなるが、第一項は残る

決定木を使う場合、各決定木は独立なのか？

Bagging の問題 – 決定木の場合

非常に強い（つまり応答変数を良く予測する）特徴量があったとしよう。そして、他の特徴量の強さはそこそこであったとしよう

そうすると、どの決定木においても、根で選択される特徴量は、この非常に強い特徴量であり、従って、どの木も似たようなものになってしまう

解決案

最初に選ぶ特徴量が同じになってしまうのが問題であった（実は、2番目に強い特徴量があると、2番目に選ぶ特徴量が同じになってしまう。以下同様）

例えば、特徴量選択をランダムにしたらどうだろうか？

いやいや、それはまずい。決定木の作り方はgreedyなので、ランダムに選んだのでは、かなり不適切な木ができてしまいそう

Random forest

Bagging と同様に、bootstrapサンプルを用いて決定木を多数作る

ただし、ノードに乗せる特徴量を選ぶとき、全特徴量の内、m 個の特徴量をランダムに選び、その中で（通常通り）最良のものを選ぶことにする

Random forest

For $b = 1$ to B:

(a) N 個の学習データから、bootstrap サンプルを作成する（それぞれN個）

(b) それぞれの木を成長させる。ただし、各ノード（その時は葉）で特徴量を選択するとき

i. 全特徴量の中から、ランダムに m 個選ぶ

ii. 通常の決定木作成アルゴリズムと同様に、最適な特徴量を m 個から選ぶ.

全体を出力とする

新規データ点 x に対しては:

回帰の場合: 各木（回帰木）の出力値の平均

分類の場合: 各木（決定木）の出力値の多数決

Boosting

- Adaboostへ