Machine learning with Python: random forest

November 17, 2017

Outline

Boostrap for estimating the generalization error

- Generalization error
- Boostrap

Bagging Breiman 1996

- Bootstrap

Random forests

Boostrap for estimating the generalization error

The data

Supervised learning

- Features $X_i \in \mathbb{R}^d$
- Labels $Y_i \in \mathbb{R}$ (regression) $Y_i \in \{-1, 1\}$ (classification)
- $(X_i, Y_i) \sim \mathcal{P}$ unknown

- With $\mathcal{D}_n = \Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$,
- and a loss function $\ell(y, f(x))$

we construct a learner $\hat{f}$, that minimize the empirical risk (or a penalized version, or a majoration...)

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

Generalisation error

What we are actually seeking is to minimize the **generalization error**

$$\mathcal{R}(f) = \mathbb{E}\{\ell(Y_+, f(X_+))\}$$

where (X_+, Y_+) has the distribution $\mathcal{P}$ and is independent of all data in $\mathcal{D}_n$.

We know that minimizing the empirical risk instead of the generalization error leads to **over-fitting**, in other words the empirical risk is a too **optimistic version** of the generalization error.

To estimate the generaralisation error

- a solution is to consider cross-validation,
- an other one is to consider the bootstrap.

Efron's bootstrap (1)

From the data $\mathcal{D}_n = \Big((X_1, Y_1), \dots, (X_n, Y_n)\Big)$, we construct new sample

$$\mathcal{D}_1^\star = \Big((X_{1,1}^\star, Y_{1,1}^\star), \dots, (X_{1,n}^\star, Y_{1,n}^\star)\Big)$$

$$\dots$$

$$\mathcal{D}_b^\star = \Big((X_{b,1}^\star, Y_{b,1}^\star), \dots, (X_{b,n}^\star, Y_{b,n}^\star)\Big)$$

$$\dots$$

by sampling the observations $(X_{b,i}^\star, Y_{b,i}^\star)$ randomly and with replacement for the data $\mathcal{D}_n$.

We know that, conditionnaly to $\mathcal{D}_n$

- each $(X_{b,i}^\star, Y_{b,i}^\star)$ has the distribution $\widehat{\mathcal{P}}$, the empirical distribution constructed from $\Big((X_1, Y_1), \dots, (X_n, Y_n)\Big)$.
- the boostraped samples $\mathcal{D}_1^\star, \dots \mathcal{D}_B^\star$ are independent.

Boostrap estimation of the generalization error

An idea would be to consider a bootstrap sample $\mathcal{D}_b^\star$ as a learning sample and the initial sample as a testing sample. But they are too many dependences here.

What we can do is to measure how optimistic is the empirical error in the estimation of the generalization error. It suffices to find an equivalent of

$$\mathcal{R}(f) - \mathcal{R}_n(f)$$

constructed with the boostraped samples.

Boostrap estimation of the generalization error (1)

1. Consider a boostrap sample $\mathcal{D}_b^\star$ and construct a learner $\hat{f}_b^\star$ on it.
2. Compute the empirical risk of the learner

$$\mathcal{R}_b^\star(\hat{f}_b^\star) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_{b,i}^\star, \hat{f}_b^\star(X_{b,i}^\star)\big)$$

3. Compute the estimation

$$\frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_i, \hat{f}_b^\star(X_i)\big) - \mathcal{R}_b^\star(\hat{f}_b^\star)$$

Boostrap estimation of the generalization error (2)

With the B estimates $\frac{1}{n}\sum_{i=1}^{n}\ell\big(Y_i,\hat{f}_b^\star(X_i)\big)-\mathcal{R}_b^\star(\hat{f}_b^\star)$, the generalization error of a learner $\hat{f}$ is estimated by

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{\frac{1}{n}\sum_{i=1}^{n}\ell\big(Y_i,\hat{f}_b^\star(X_i)\big)-\mathcal{R}_b^\star(\hat{f}_b^\star)\Big\}+\mathcal{R}_n(\hat{f}).$$

Bagging Breiman 1996

The data

Supervised learning

- Features $X_i \in \mathbb{R}^d$
- Labels $Y_i \in \mathbb{R}$ (regression) $Y_i \in \{-1, 1\}$ (classification)
- $(X_i, Y_i) \sim \mathcal{P}$ unknown

Weak learners

- Consider a set "weak learners" $\mathcal{H}$ such that
- each weak leaner $h : \mathbb{R}^d \to \mathbb{R}$ ou $\mathbb{R}^d \to \{-1, 1\}$ is very simple

Efron's bootstrap (1)

From the data $\mathcal{D}_n = \Big((X_1, Y_1), \dots, (X_n, Y_n)\Big)$, we construct new sample

$$\mathcal{D}_1^\star = \Big((X_{1,1}^\star, Y_{1,1}^\star), \dots, (X_{1,n}^\star, Y_{1,n}^\star)\Big)$$
$$\dots$$
$$\mathcal{D}_b^\star = \Big((X_{b,1}^\star, Y_{b,1}^\star), \dots, (X_{b,n}^\star, Y_{b,n}^\star)\Big)$$
$$\dots$$

by sampling the observations $(X_{b,i}^\star, Y_{b,i}^\star)$ randomly and with replacement for the data $\mathcal{D}_n$.

We know that, conditionnaly to $\mathcal{D}_n$, each $(X_{b,i}^\star, Y_{b,i}^\star)$ has the distribution $\widehat{\mathcal{P}}$, the empirical distribution constructed from $\Big((X_1, Y_1), \dots, (X_n, Y_n)\Big)$.

Efron's bootstrap (2)

From each bootstrapped sample $\mathcal{D}_b^\star$, we can construct a learner $\hat{f}_b^\star$:

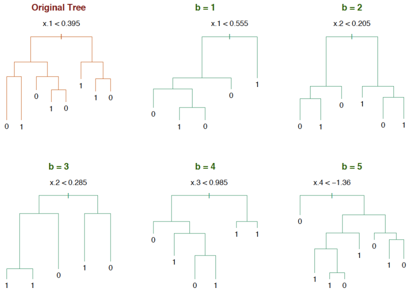


Figure 1: From Friedman, Hastie, and Tibshirani 2001

Efron's bootstrap (3)

The bootstrap aggregation or bagging averages the learners over the bootstrapped samples

$$\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^\star$$

with the intuition that (in the regression case) it would have a smaller variance that $\hat{f}$, hence a smaller generalization error.

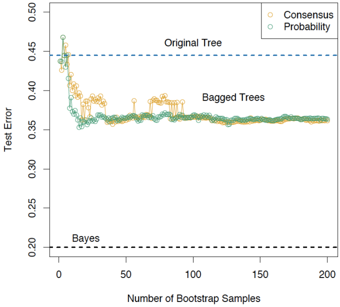


Figure 2: From Friedman, Hastie, and Tibshirani 2001

Simple facts on the bagging

As the boostraped samples are i.i.d. (conditionnally to the initial sample), the expectation of

$$\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^{\star}$$

is the same as the one of $\hat{f}_b^{\star}$, so we cannot hope to reduce the bias.

For the variance on the contrary, we can hope a reduction based on the fact that the boostraped samples are independent. (Just write that if $Z_1, \ldots, Z_B$ are i.i.d. the variance of $1/B\sum_1^B Z_b$ is $\mathbb{V}(Z_b)/B$.)

Random forests

Why this is not so simple !

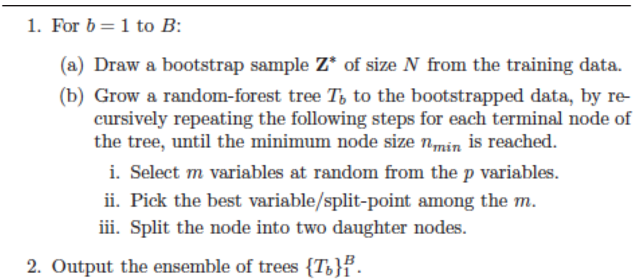
We just gave arguments conditionally to the initial sample, we should reason without this conditionning. In this case, it is clear that the $\hat{f}_b^\star$ are dependent.

If $Z_1, \dots, Z_B$ are i.d. with correlation ρ (we assume it positive) the variance of $1/B \sum_1^B Z_b$ is

$$\rho \mathbb{V}(Z_b) + \frac{1-\rho}{B} \mathbb{V}(Z_b).$$

So what we should be concern about it the correlation between the trees.

Random forestsBreiman 2001



```
1. For $b = 1$ to $B$:
    (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
```

Figure 3: From Friedman, Hastie, and Tibshirani 2001