Bootstrapping

Giri Iyengar

Cornell University

gi43@cornell.edu

April 11, 2018

Overview

1. Bias-Variance trade-off and Cross Validation

2. Bootstrapping

3. Little Bag of Bootstraps

Overview

1. Bias-Variance trade-off and Cross Validation

2. Bootstrapping

3. Little Bag of Bootstraps

Bias-Variance Trade-off

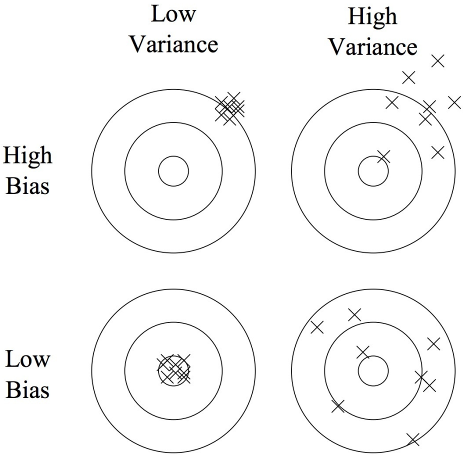


Figure: Image showing Bias Variance Trade-off - courtesy Quora

Bias-Variance Trade-off

- In machine learning, we are trying to learn $y = f(x) + \epsilon$
- In addition to intrinsic noise, ϵ, the models have their own sources of error
- Bias: The tendency of the algorithm to be consistently incorrect
- Variance: The algorithm's tendency to fit to the noise in the data in addition to the signal
- Models with high bias tend to **underfit**. E.g. represent a linear relationship with the mean
- Models with high variance tend to **overfit** E.g. represent a linear relationship with a higher-order polynomial

Bias-Variance Trade-off: Mathematical definition

Bias Variance Mathematical definition

We can represent the error of a model as $Err(x) = E[(y - \hat{f}(x))^2]$.

Decompose this down as $(E[\hat{f}(x)] - f(x))^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] + \epsilon^2$.

In other words, $Err(x) = Bias^2 + Variance + Noise$.

Given infinite data, we can construct models that drive both bias and variance down to zero. However, we live in an imperfect world with finite data, noisy measurement tools, and finite resources. Typically there is a trade-off between Bias and Variance and we try to find the best balance between the two.

Bias Variance Trade-off example

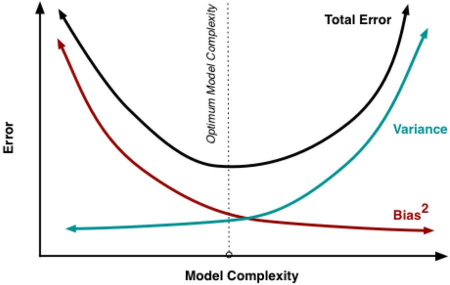


Figure: Bias and Variance vs Model Complexity

Overview

1. Bias-Variance trade-off and Cross Validation
2. Bootstrapping
3. Little Bag of Bootstraps

Some ways to understand and work with bias-variance trade-off

- Cross-validation
- Bootstrapping
- Little Bag of Bootstraps

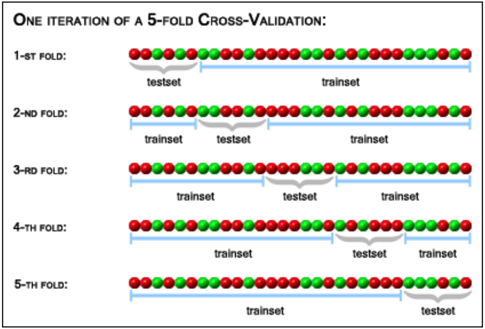


Figure: Cross Validation of a Model

Cross-validation

- Doesn't use the entire training set
- Typically the error is *biased* upwards
- Variance estimates of Θ is not strictly correct (K splits are not *independent*)

Cross-validation mistakes

Cross-validation

Consider a simple classifier applied to some two-class data:

1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
2. We then apply a classifier such as logistic regression, using only these 100 predictors.
3. How do we estimate the test set performance of this classifier?

Can we apply cross-validation in step 2, forgetting about step 1?

Cross-validation mistakes

- **Wrong**: Running CV only on Step 2
- **Right**: Running CV on both Step 1 and 2

Bootstrap

- Powerful technique for estimating Bias and Variance
- Very simple, applies to most situations (e.g. except for Power Law/non-finite variance)
- Make inference about the population from a single sample
- Approximate population distribution by empirical distribution

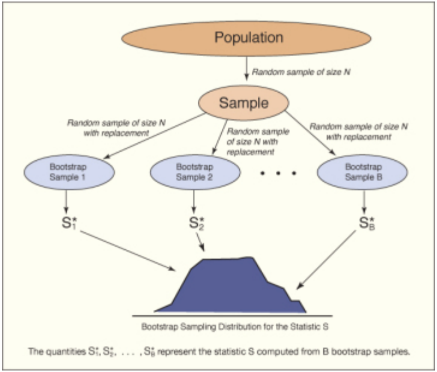

Population
Random sample of size N
Sample
Random sample of size N with replacement
Random sample of size N with replacement
Random sample of size N with replacement
Bootstrap Sample 1
Bootstrap Sample 2
. . .
Bootstrap Sample B
S_1^*
S_2^*
S_B^*
Bootstrap Sampling Distribution for the Statistic S
The quantities $S_1^*, S_2^*, \ldots, S_B^*$ represent the statistic S computed from B bootstrap samples.

Figure: The Bootstrap Method

The Bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

The Bootstrap

- If we had several independent *samples* of the *population*, we could compute *independent* estimates of parameters

The Bootstrap

- If we had several independent *samples* of the *population*, we could compute *independent* estimates of parameters
- In practice, gathering data is *hard* and *expensive*

The Bootstrap

- If we had several independent *samples* of the *population*, we could compute *independent* estimates of parameters
- In practice, gathering data is *hard* and *expensive*
- Bootstrap starts with the assumption that empirical distribution of a single sample closely resembles the population

The Bootstrap

- If we had several independent *samples* of the *population*, we could compute *independent* estimates of parameters
- In practice, gathering data is *hard* and *expensive*
- Bootstrap starts with the assumption that empirical distribution of a single sample closely resembles the population
- Sample **with replacement** from the original data set and derive as many copies of the data as you want

The Bootstrap

- If we had several independent *samples* of the *population*, we could compute *independent* estimates of parameters
- In practice, gathering data is *hard* and *expensive*
- Bootstrap starts with the assumption that empirical distribution of a single sample closely resembles the population
- Sample **with replacement** from the original data set and derive as many copies of the data as you want
- Estimate coefficients on each sample independently and use that to derive variance, and standard error estimates

Overview

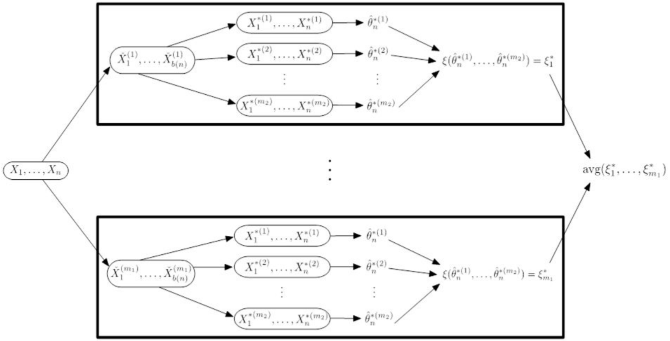
1. Bias-Variance trade-off and Cross Validation
2. Bootstrapping
3. Little Bag of Bootstraps

Little Bag of Bootstraps

- Approximately 63.2% samples covered in each bootstrap round
- When N, the sample size, is large this can be a severe limitation
- 1TB of data → 632GB per bootstrap round
- You need to perform several rounds to get good estimates
- Difficult to parallelize when you have to move that much data around

Little Bag of Bootstraps

The Bag of Little Bootstraps (BLB)



$X_1, \dots, X_n$

$\check{X}_1^{(1)}, \dots, \check{X}_{b(n)}^{(1)}$

$X_1^{*(1)}, \dots, X_n^{*(1)}$ → $\hat{\theta}_n^{*(1)}$

$X_1^{*(2)}, \dots, X_n^{*(2)}$ → $\hat{\theta}_n^{*(2)}$

$\vdots$

$X_1^{*(m_2)}, \dots, X_n^{*(m_2)}$ → $\hat{\theta}_n^{*(m_2)}$

$\xi(\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(m_2)}) = \xi_1^*$

$\vdots$

$\check{X}_1^{(m_1)}, \dots, \check{X}_{b(n)}^{(m_1)}$

$X_1^{*(1)}, \dots, X_n^{*(1)}$ → $\hat{\theta}_n^{*(1)}$

$X_1^{*(2)}, \dots, X_n^{*(2)}$ → $\hat{\theta}_n^{*(2)}$

$\vdots$

$X_1^{*(m_2)}, \dots, X_n^{*(m_2)}$ → $\hat{\theta}_n^{*(m_2)}$

$\xi(\hat{\theta}_n^{*(1)}, \dots, \hat{\theta}_n^{*(m_2)}) = \xi_{m_1}^*$

$\text{avg}(\xi_1^*, \dots, \xi_{m_1}^*)$

Figure: The Little Bag of Bootstraps Method

Little Bag of Bootstraps

- From your sample of N data points, create s samples (without replacement) of size $\approx N^{0.6}$
- On each of these s samples, run r bootstrap iterations
- In the inner bootstraps (r iterations), data is sampled with replacement and resampled back to size N
- Take average of averages and return that as your estimate. Also return confidence intervals

Little Bag of Bootstraps

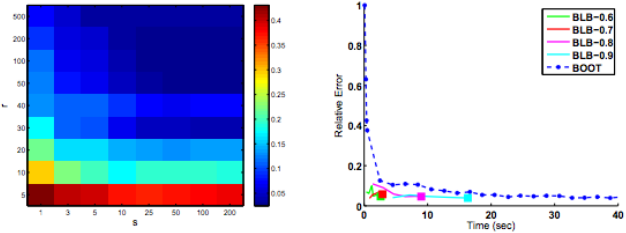


Figure: Little Bag of Bootstraps Performance

рахмат
danke
謝謝
ngiyabonga
Баярлалаа
spas
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
хвала
nanni
nandri
dhanyavad
thank you
gracias
asante
manana
obrigada
bedankt
bayarlalaa
gracie
hvala
maururu
köszönöm
dankon
aciu
akun
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
dékuji
obrigado
mèsi
didi madloba
sagolun
sukriya
kop khun krap
ありがとう
arigatô
takk
dakujem
trugarez
najis tuke
kam sah hamnida
rahmat
terima kasih
tanemirt
rahmet
grazie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci