

bootstrap

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

The bootstrap estimates how much the output of a machine learning (ML) method moves when its dataset is drawn again, using only the one dataset at hand. It reads the data points as realizations of independent and identically distributed (i.i.d.) random variables (RVs) and replaces their unknown probability distribution by the empirical distribution of the dataset, so that drawing a new dataset becomes drawing with replacement from the old one. Each of the B datasets so drawn has the same size as the original, repeats some data points and omits others, and holds about $0.63m$ distinct ones. Model training on each gives B learned hypotheses whose spread estimates the variance of the method. Resampling a test set with replacement, with the learned hypothesis held fixed, gives a confidence interval for a reported accuracy. The overlap with the original dataset is what distinguishes the bootstrap from k -fold cross-validation (k -fold CV), whose blocks do not overlap, and it makes a naive bootstrap estimate of prediction error optimistic. Bootstrap aggregating (bagging) puts the same resampling to a different use, averaging the hypotheses instead of measuring their spread.

Definition

Forty days of weather recordings fix one straight line through them. Had forty other days been recorded, the line would have come out slightly different, and how much it would move is what decides whether its slope means anything. Only one set of forty days was ever recorded, so the repetitions that would answer the question do not exist. The bootstrap manufactures them from the forty days in hand.

To manufacture them, the forty days are read as realizations of independent and identically distributed (i.i.d.) random variables (RVs). A dataset $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$ is then a draw of m such RVs with a common probability distribution $\mathbb{P}$. That $\mathbb{P}$ is unknown and must be estimated from $\mathcal{D}$ itself. The bootstrap estimates it by the empirical distribution $\mathbb{P}^{(\mathcal{D})}$ of $\mathcal{D}$ (Efron, 1979; Efron and Tibshirani, 1993; Hastie et al., 2009, Sect. 7.11),

$$\frac{1}{m} |r : \mathbf{z}^{(r)} \in \mathcal{A}| \approx \mathbb{P}(\mathcal{A}).$$

Once $\mathbb{P}$ has been replaced by $\mathbb{P}^{(\mathcal{D})}$, the shortage of data is gone: $\mathbb{P}^{(\mathcal{D})}$ is a fully specified probability distribution and can be sampled as often as wanted, so as many data points and as many datasets as needed can be generated from the one dataset that was collected. The gain is replicates, not information. Each replicate keeps the size m of $\mathcal{D}$, because the quantity being imitated is the spread of a dataset of that size drawn from $\mathbb{P}$. Nor does a replicate reach new data points: $\mathbb{P}^{(\mathcal{D})}$ places all its mass on the m data points observed, so no draw can produce one that $\mathcal{D}$ does not already contain.

Sampling m data points from $\mathbb{P}^{(\mathcal{D})}$ is the same as drawing m times from $\mathcal{D}$ with replacement: every draw is made from the whole of $\mathcal{D}$, so a data point can be drawn twice and another not at all (Fig. 1). Let B be the number of such draws, a number limited only by the computation they cost. Repeating the draw B times yields datasets $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(B)}$, each of m data points.

$\mathcal{D}$	<table><tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr></table>	1	2	3	4	5	6	7	8	
1	2	3	4	5	6	7	8			
$\mathcal{D}^{(1)}$	<table><tr><td>2</td><td>5</td><td>6</td><td>6</td><td>7</td><td>7</td><td>8</td><td>8</td></tr></table>	2	5	6	6	7	7	8	8	omits 1, 3, 4
2	5	6	6	7	7	8	8			
$\mathcal{D}^{(2)}$	<table><tr><td>1</td><td>1</td><td>3</td><td>3</td><td>4</td><td>7</td><td>7</td><td>8</td></tr></table>	1	1	3	3	4	7	7	8	omits 2, 5, 6
1	1	3	3	4	7	7	8			
$\mathcal{D}^{(3)}$	<table><tr><td>1</td><td>2</td><td>3</td><td>3</td><td>3</td><td>4</td><td>7</td><td>7</td></tr></table>	1	2	3	3	3	4	7	7	omits 5, 6, 8
1	2	3	3	3	4	7	7			

Figure 1: Three bootstrap datasets drawn from a dataset $\mathcal{D}$ of $m = 8$ data points (shaded), each by drawing 8 times with replacement. Every row has the same length as $\mathcal{D}$, repeats some data points and omits others; each of the three happens to contain 5 of the 8 distinct data points, against the $0.63 \cdot 8 \approx 5$ that the calculation below predicts

Running model training on each, e.g., via empirical risk minimization (ERM), gives learned hypotheses $\hat{h}^{(1)}, \dots, \hat{h}^{(B)}$. Their spread estimates how much the output of the machine learning (ML) method moves when the dataset is drawn again, which is what the variance of the method means; the same hypotheses also estimate its bias and its generalization gap (Hastie et al., 2009, Sect. 7.11). For the forty weather days, the forty-day slope is reported together with the range that the B bootstrap slopes cover, which is a confidence interval for it.

Fig. 2 draws that picture with measured data: not forty days but $m = 1096$, the daily minimum and maximum temperature at Krems an der Donau from 2022 to 2024. In the left panel each dot is one data point, its feature the minimum temperature and its label the maximum temperature, and the thick line is the hypothesis learned from all of them by ERM with the squared error loss. The thin lines are learned by the same ERM problem from bootstrap datasets. Their spread is narrow here, and the 95% confidence interval for the slope runs from 1.07 to 1.13. The dots are also all the mass that $\mathbb{P}^{(\mathcal{D})}$ has, which is what the right panel makes visible: smearing each dot into a small bump and adding the bumps gives a density, and its contours show two crowded regions,

the winter days and the summer days. Drawing from that smoothed density rather than from the dots is a variant called the smoothed bootstrap; it produces temperatures that no day of $\mathcal{D}$ recorded, which the bootstrap itself never does.

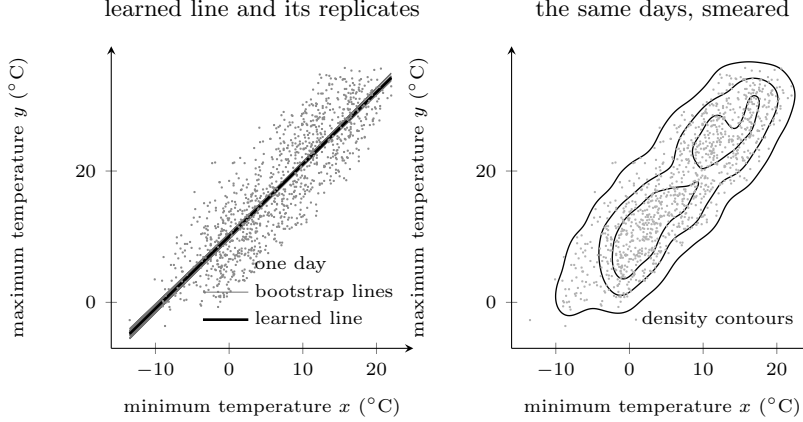


Figure 2: Daily minimum and maximum temperature at Krems an der Donau, 2022 to 2024, $m = 1096$ data points (GeoSphere Austria, 2026). Left: the hypothesis learned from all of them (thick) and 25 hypotheses learned from bootstrap datasets (thin), whose spread gives the confidence interval $[1.07, 1.13]$ for the slope. Right: the same data points with each one smeared into a bump, the contours of the resulting density marking two crowded regions. The empirical distribution itself has no such spread: all its mass sits on the dots. Data generated by `pythondemos/bootstrap.py`

The same resampling turns a single test set into a confidence interval for a reported accuracy. A hypothesis $\hat{h}$ is learned once from the training set $\mathcal{D}^{(\text{train})}$, and the held-out test set $\mathcal{D}^{(\text{test})}$ is resampled B times with replacement, each draw holding as many data points as $\mathcal{D}^{(\text{test})}$. Evaluating the same $\hat{h}$ on each of them gives B values of the accuracy, and the interval between the 2.5th and the 97.5th percentile of those values is a 95% confidence interval for the accuracy (Hastie et al., 2009, Sect. 8.2.1). Only the test set is redrawn here, while $\hat{h}$ is held fixed, so the interval reports how much the measured accuracy depends on which data points happened to be collected for testing. A single number such as 94% accuracy on 200 test data points says nothing about that dependence; the interval it comes with does.

A data point is left out of a bootstrap dataset with probability $(1 - 1/m)^m$, so it appears in one with probability $1 - (1 - 1/m)^m$, which approaches $1 - 1/e \approx 0.63$ as m grows. A bootstrap dataset therefore holds about $0.63m$ distinct data points and overlaps $\mathcal{D}$ heavily. That overlap is what separates the bootstrap from k -fold cross-validation (k -fold CV), which partitions $\mathcal{D}$ into blocks that do not overlap at all. Used naively to estimate prediction error, the bootstrap

therefore evaluates the learned hypotheses on data points that entered their own training, and the estimate comes out optimistic (Hastie et al., 2009, Sect. 7.11). The repair is to imitate k -fold CV: keep, for each data point, only the predictions of those bootstrap datasets that do not contain it. That leave-one-out bootstrap then trains on about $0.63m$ distinct data points, so its bias behaves like that of k -fold CV with $k = 2$, which the 0.632 estimator corrects (Hastie et al., 2009, Sect. 7.11).

Bootstrap aggregating (bagging) is the bootstrap put to a different use: instead of measuring the spread of $\hat{h}^{(1)}, \dots, \hat{h}^{(B)}$, it averages them into one hypothesis, which is how a random forest is built from decision trees.

Cross References

empirical distribution, i.i.d., RV, probability distribution, variance, k -fold CV, bagging, confidence interval, test set, accuracy

References

1. Efron and Tibshirani (1993). An Introduction to the Bootstrap. Chapman & Hall, New York, NY, USA.
2. Efron (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Statist., 7(1), pp. 1–26.
3. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: <https://dataset.api.hub.geosphere.at/v1/station/v2-1d>.
4. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.