Bootstrap & Bagging

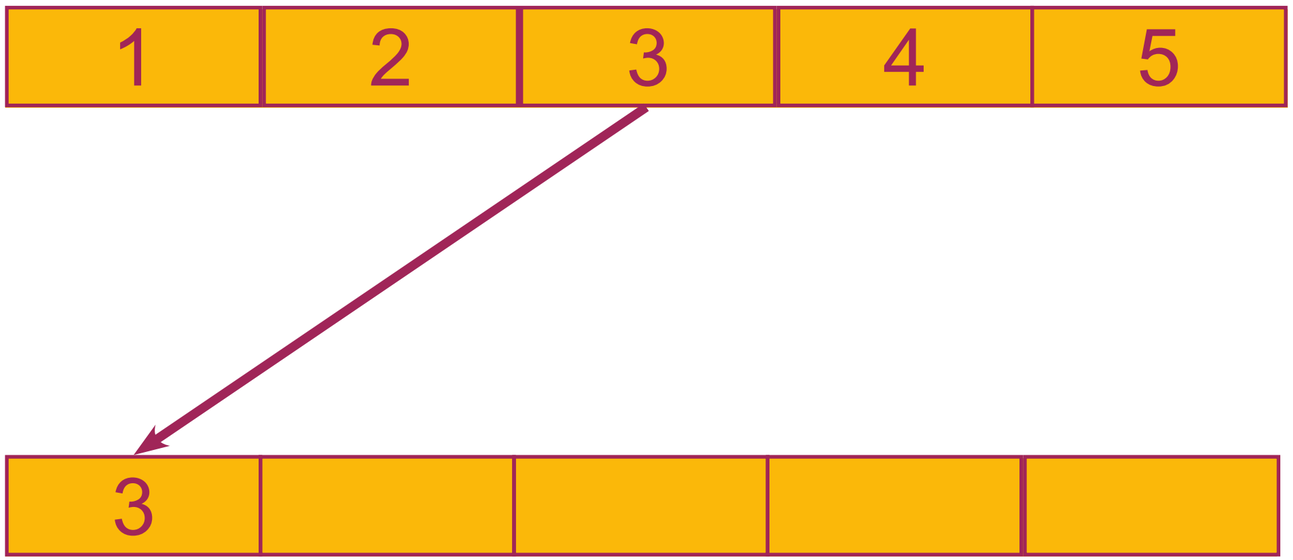
Bootstrap

Consider a dataset $Z=\{(x_1, y_1),\ldots,(x_n, y_n)\}$

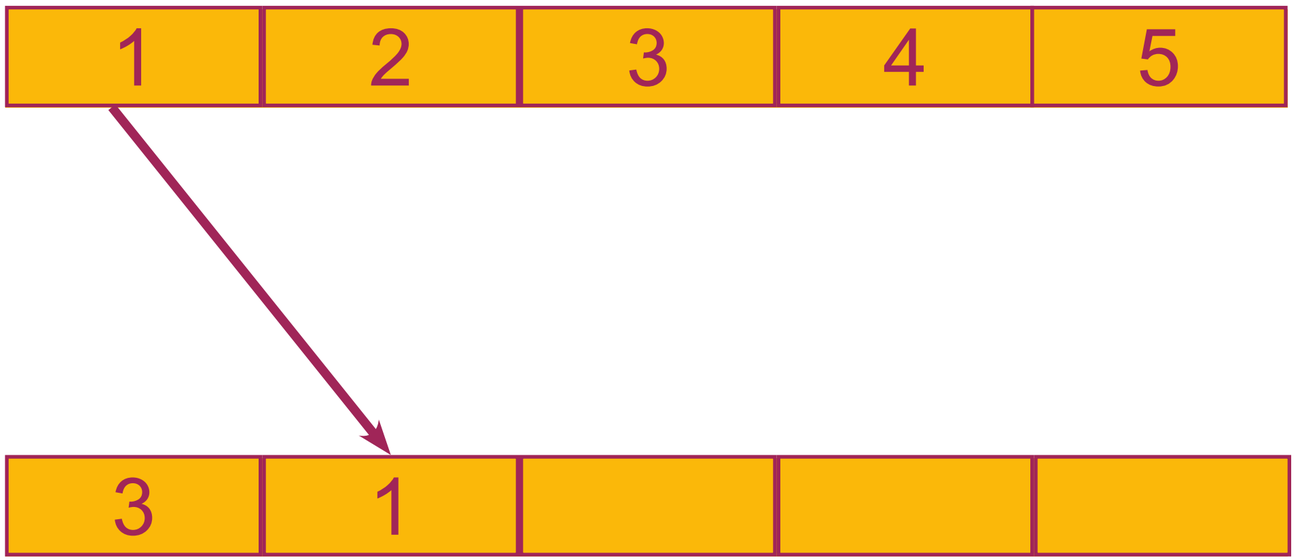
Bootstrapped dataset Z^* – is a modification of the original dataset Z, produced by random sampling with replacement

Sampling with Replacement

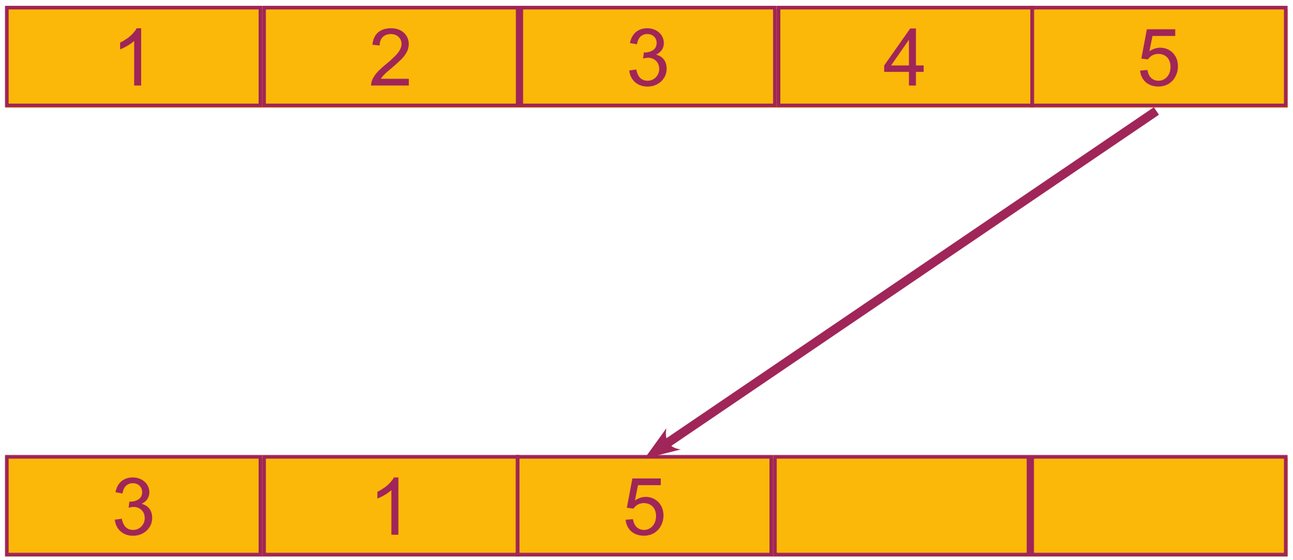
Sampling with Replacement



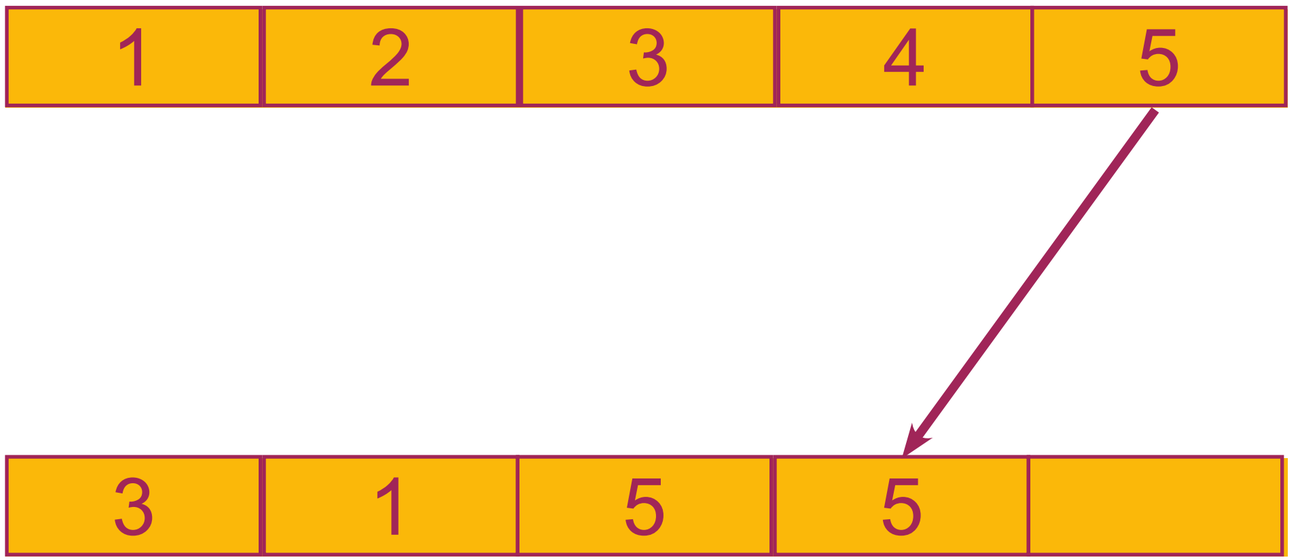
Sampling with Replacement



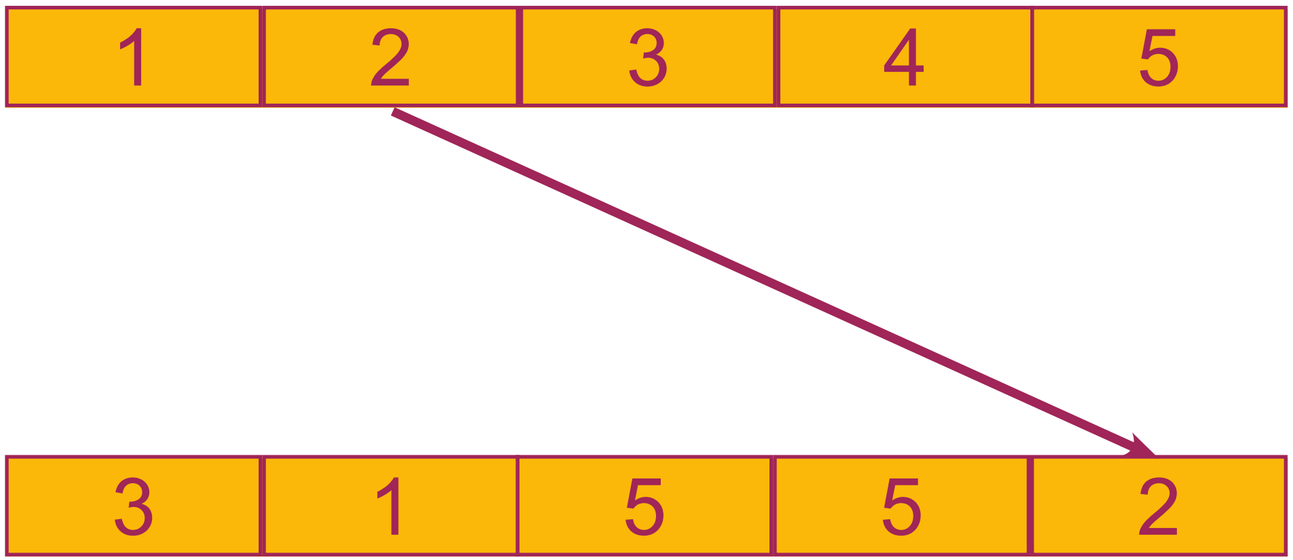
Sampling with Replacement



Sampling with Replacement



Sampling with Replacement



Sampling with Replacement

original dataset

bootstrapped dataset

Bagging

Bagging (**b**ootstrap **agg**regation) – a method for averaging predictions and reducing prediction's variance

Algorithm: Bagging

Input: training set $Z=\{(x_1, y_1),\ldots,(x_n, y_n)\}$,
B – number of iterations
Machine learning method M

1. For b=1…B:
2. Draw a bootstrap sample Z^{*b} of size n from training data
3. Apply method M to the dataset Z^{*b} and obtain a model $f(x)^{*b}$
4. Return: ensemble $\{f^{*1}\ldots f^{*B}\}$

Prediction with ensemble:

- Regression: $\mathbf{f(x)} = \frac{1}{n}\sum_{b=1}^{B} f^{*b}(x)$
- Classification: majority vote of all predictions $f^{*b}(x)$, $b=1\ldots B$

Model f(x) has higher predictive power than any single $f^{*b}(x)$, b=1,…, B

Why does bagging work? Bias-variance trade-off. One may consider the training dataset to be **random**.

Bagging – is an averaging over a set of possible datasets, removing noisy and non-stable parts of models.

Summary

- **Bootstrap** – a method for generating different replicas of the dataset

- **Bagging** (**b**ootstrap **agg**regation) – a method for averaging predictions and reducing prediction's variance

- **Bagging** improves the quality of almost any machine learning method

- **Bagging** is time consuming for large datasets