Re-Sampling Methods

WHY WE NEED CROSS-VALIDATION?

- R squared, also known as coefficient of determination, is a popular measure of quality of fit in regression. However, it does not offer any significant insights into how well our regression model can predict future values.
- When an MLR equation is to be used for prediction purposes it is useful to obtain empirical evidence as to its generalizability, or its capacity to make accurate predictions for new samples of data. This process is sometimes referred to as “validating” the regression equation.

- One way to address this issue is to literally obtain a new sample of observations. That is, after the MLR equation is developed from the original sample, the investigator conducts a new study, replicating the original one as closely as possible, and uses the new data to assess the predictive validity of the MLR equation.
- This procedure is usually viewed as impractical because of the requirement to conduct a new study to obtain validation data, as well as the difficulty in truly replicating the original study.

- An alternative, more practical procedure is *cross-validation*.

CROSS-VALIDATION

- In cross-validation the original sample is split into two parts. One part is called the training (or *derivation)* sample, and the other part is called the *validation (or validation + testing)* sample.

1) What portion of the sample should be in each part?

If sample size is very large, it is often best to split the sample in half. For smaller samples, it is more conventional to split the sample such that 2/3 of the observations are in the derivation sample and 1/3 are in the validation sample.

CROSS-VALIDATION

2) How should the sample be split?

The most common approach is to divide the sample randomly, thus theoretically eliminating any systematic differences. One alternative is to define matched pairs of subjects in the original sample and to assign one member of each pair to the derivation sample and the other to the validation sample.

- Modeling of the data uses one part only. The model selected for this part is then used to predict the values in the other part of the data. A valid model should show good predictive accuracy.
- One thing that R-squared offers no protection against is overfitting. On the other hand, cross validation, by allowing us to have cases in our testing set that are different from the cases in our training set, inherently offers protection against overfitting.

CROSS VALIDATION – THE IDEAL PROCEDURE

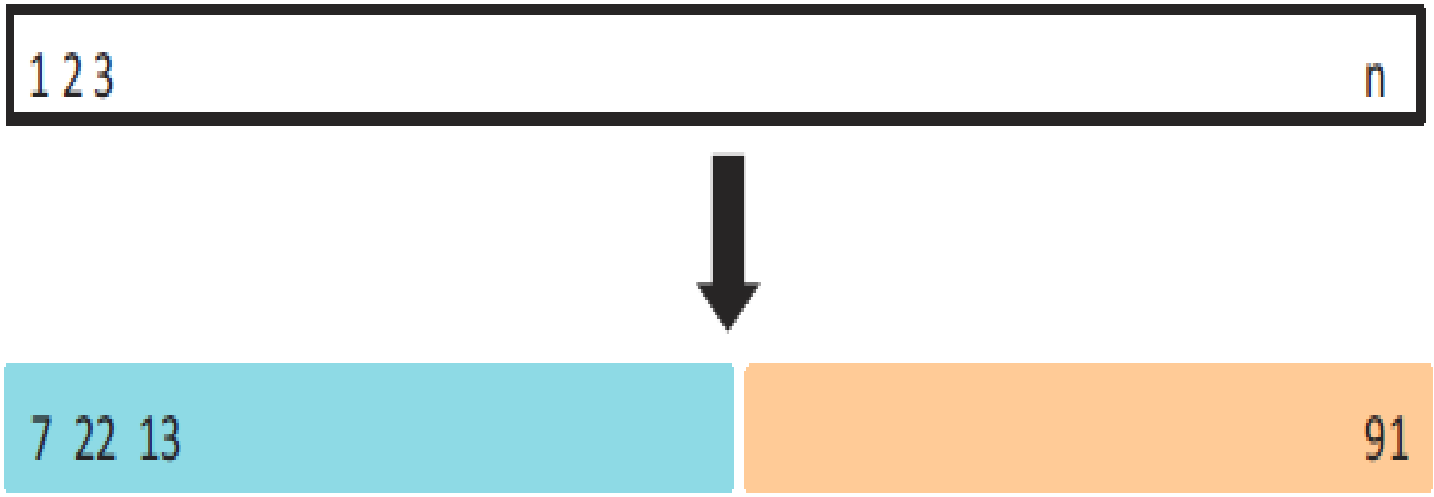
1.Divide data into three sets, training, validation and test

2.Find the optimal model on the training set, and use the test set to check its predictive capability

3.See how well the model can predict the test set

4.The validation error gives an unbiased estimate of the predictive power of a model

Cross Validation: Validation Set



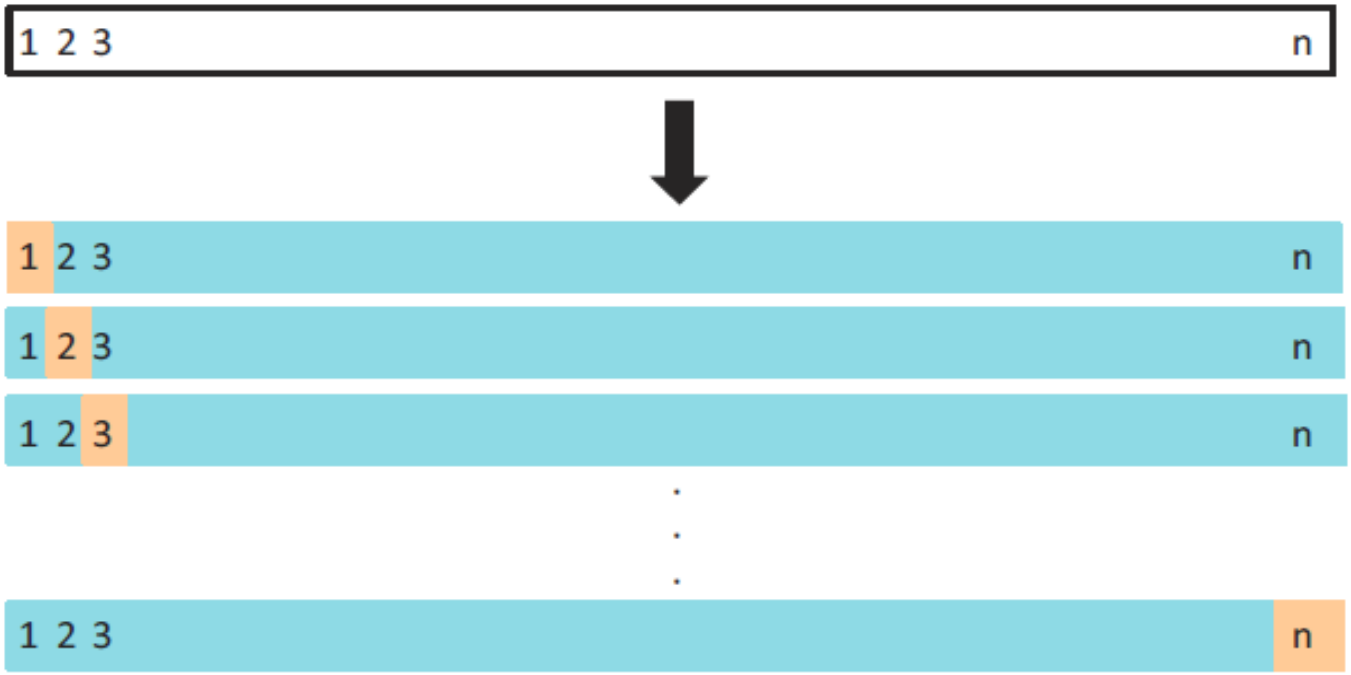
A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

Drawbacks of Cross Validation: Validation set

The validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

In the validation approach, only a subset of the observations—those that are included in the training set rather than in the validation set—are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

Leave-One-Out Cross-Validation

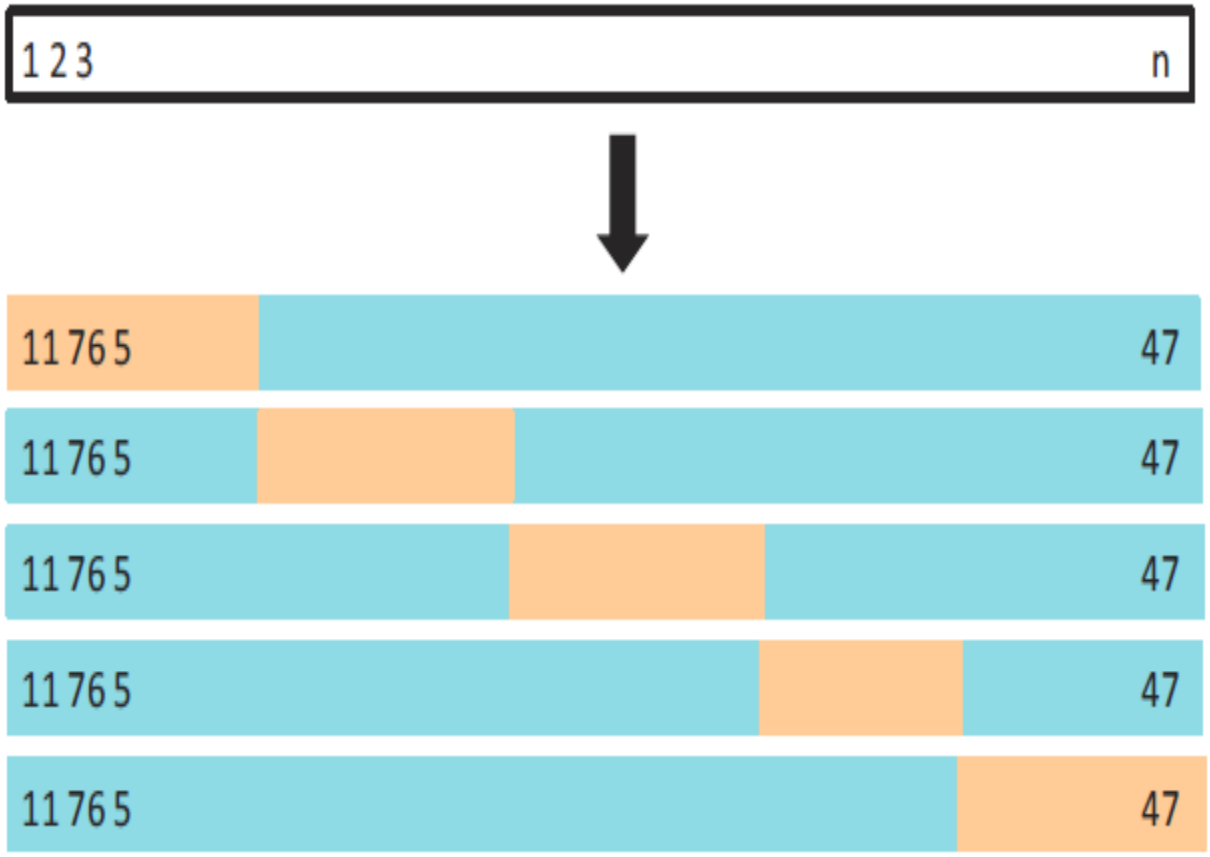


A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

K-fold CV

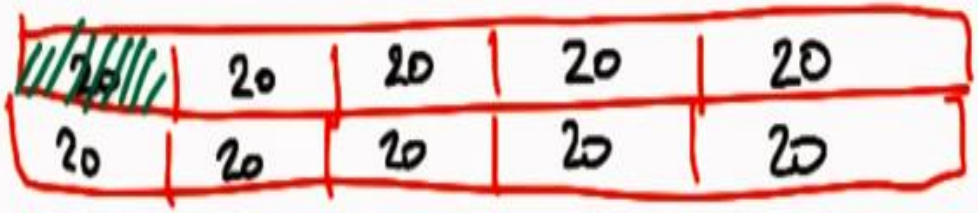
An alternative to LOOCV is *k-fold CV*. This approach involves randomly dividing the set of observations into k groups, or *folds*, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. The mean squared error, MSE_1, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, $\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_k$. The k-fold CV estimate is computed by averaging these values,

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i.$$



A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

k-fold CROSS VALIDATION



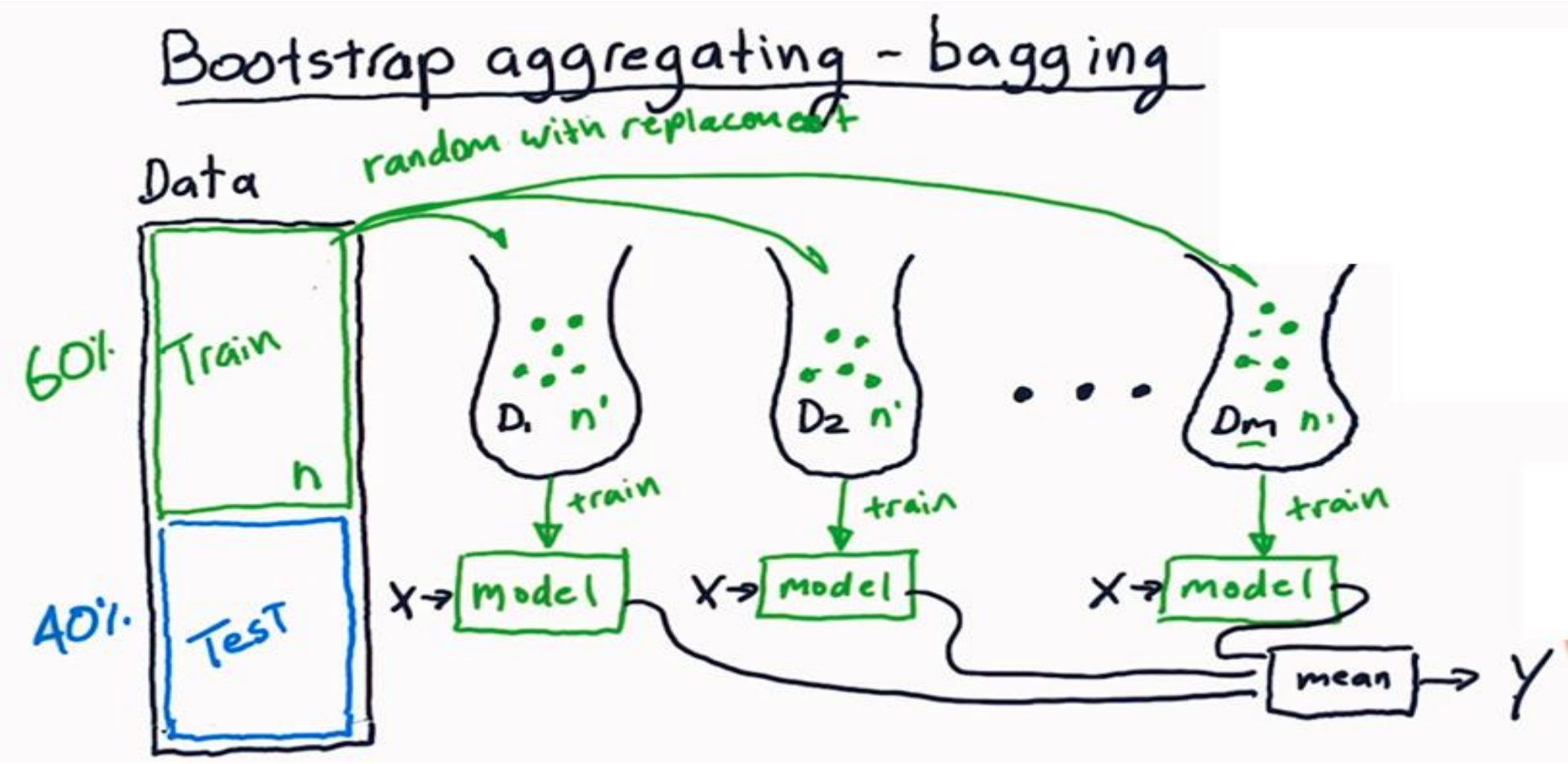
k = 10

Run k seperate learning experiments

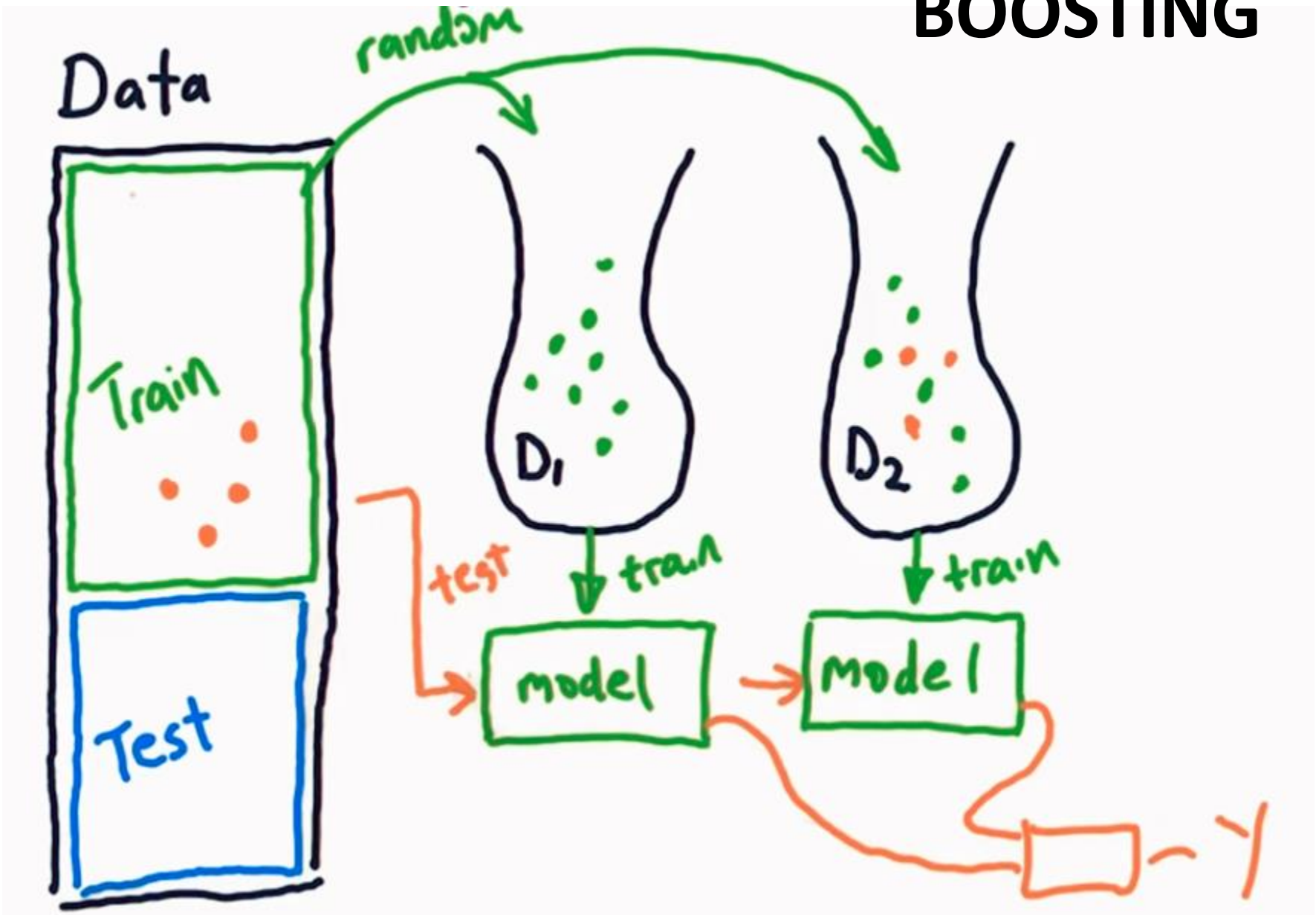
- pick testing set
- train
- test on testing set

Average test results from those k experiments

BAGGING


Bootstrap aggregating - bagging
random with replacement
Data
60%
Train
n
40%
Test
D1 n'
D2 n'
Dm n'
train
train
train
X
model
X
model
X
model
mean
Y

BOOSTING



Boosting Problem

- *crim*
  per capita crime rate by town.
- *zn*
  proportion of residential land zoned for lots over 25,000 sq.ft.
- *indus*
  proportion of non-retail business acres per town.
- *chas*
  Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- *nox*
  nitrogen oxides concentration (parts per 10 million).
- *rm*
  average number of rooms per dwelling.
- *age*
  proportion of owner-occupied units built prior to 1940.
- *dis*
  weighted mean of distances to five Boston employment centres.
- *rad*
  index of accessibility to radial highways.
- *tax*
  full-value property-tax rate per $10,000.
- *ptratio*
  pupil-teacher ratio by town.
- *black*
  1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- *lstat*
  lower status of the population (percent).
- *medv*
  median value of owner-occupied homes in $1000s.
- Source
- Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81–102.
- Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.