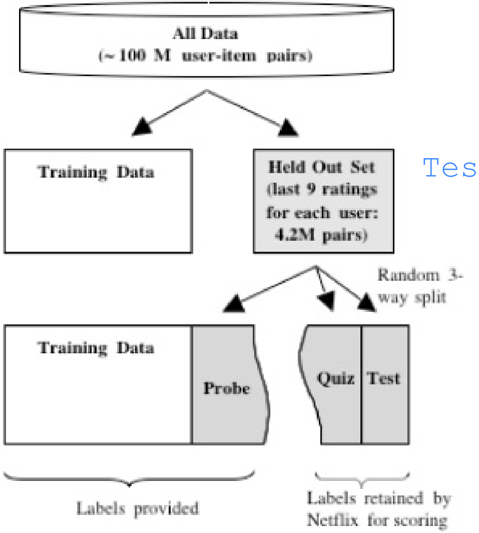
Cross validation

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Study design


All Data
(~ 100 M user-item pairs)
Training Data
Held Out Set (last 9 ratings for each user: 4.2M pairs)
Random 3-way split
Training Data
Probe
Quiz
Test
Labels provided
Labels retained by Netflix for scoring

Test data set

http://www2.research.att.com/~volinsky/papers/ASAStatComp.pdf

Key idea

1. Accuracy on the training set (resubstitution accuracy) is optimistic
2. A better estimate comes from an independent set (test set accuracy)
3. But we can't use the test set when building the model or it becomes part of the training set
4. So we estimate the test set accuracy with the training set. using Cross validation

Cross-validation

Approach:

1. Use the training set
2. Split it into training/test sets i.e. we split it again!
3. Build a model on the training set a subset of the original training set
4. Evaluate on the test set a subset of the original training set
5. Repeat and average the estimated errors

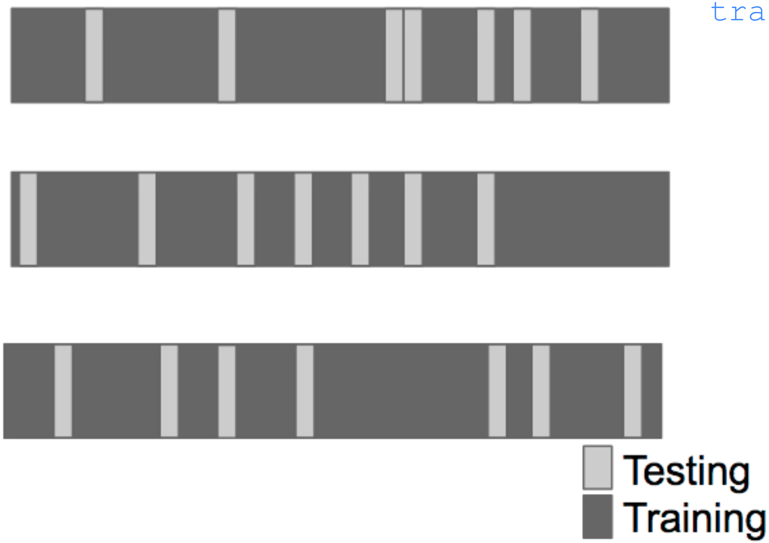
Used for: so the original test set is never used.

1. Picking variables to include in a model
2. Picking the type of prediction function to use
3. Picking the parameters in the prediction function
4. Comparing different predictors

Random subsampling

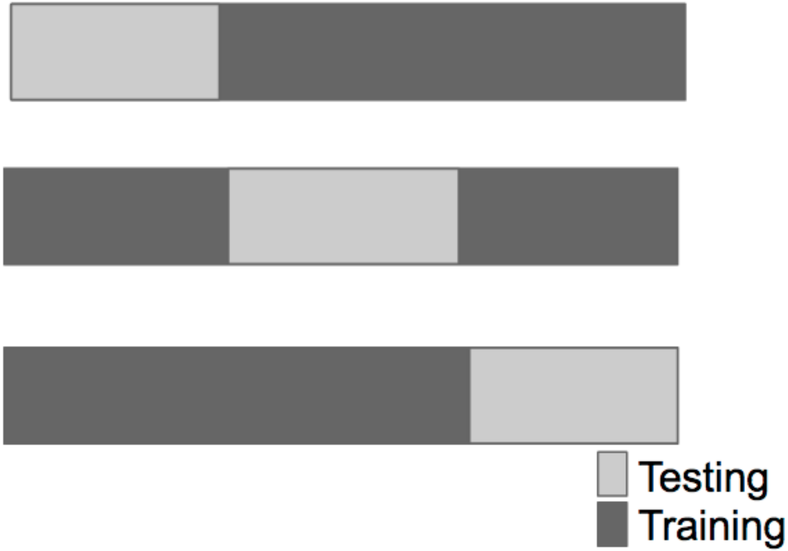
this is only still from the original training set

different runs, in the end: averaging the errors

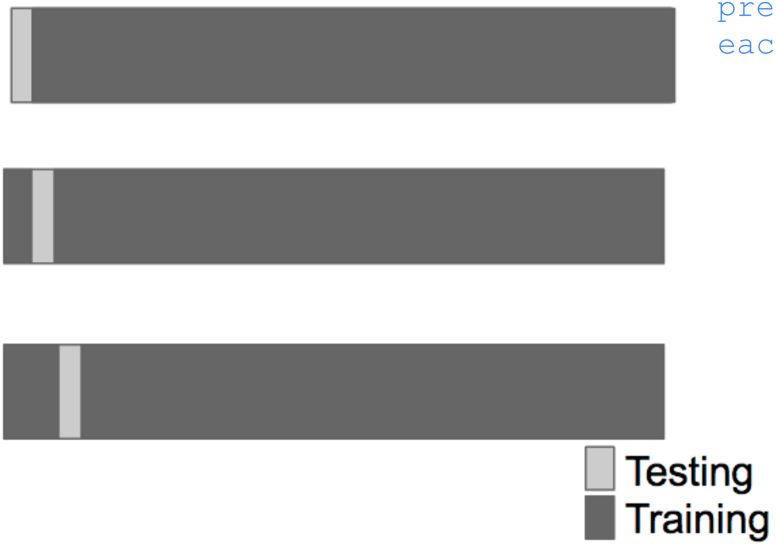


K-fold

break training set up into k equal sized sets.



Leave one out



predict only one in each run.

Considerations

- For time series data data must be used in "chunks"

because time t might depend on time t-1 (or -2,..)

- For k-fold cross validation
  - Larger k = less bias, more variance
  - Smaller k = more bias, less variance

bias=ungenaues Modell — under fitting
variance=zu genaues Modell — over fitting

- Random sampling must be done *without replacement*
- Random sampling with replacement is the *bootstrap*
  - Underestimates of the error
  - Can be corrected, but it is complicated (0.632 Bootstrap)
- If you cross-validate to pick predictors estimate you must estimate errors on independent data.