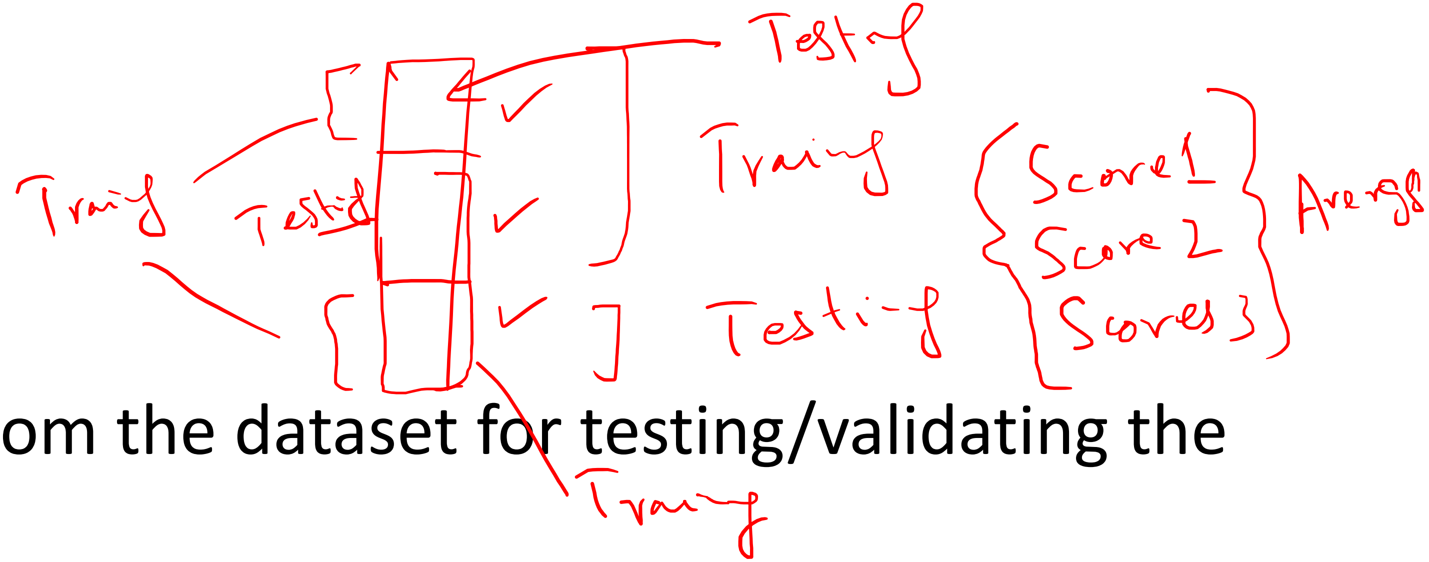
Cross Validation

Dr. Muhammad Wasim

Cross Validation



- Which data should be split from the dataset for testing/validating the data?
- Will 70-30 split do a good job help determine how well the model is?
- May be the 30 percent data from the train_test_split function does not provide a good indication of the model performance.
- A better way would be to evaluate the model using cross validation.
- The simplest form of cross validation is leave-one-out.

Leave One Out

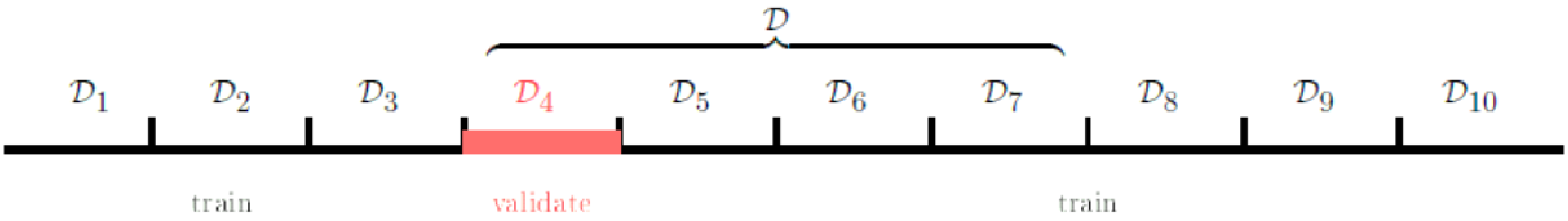
- Set N-1 points for training and 1 point for validation.
- Repeat this for all points. Every estimate is out of sample with respect to hypothesis that's used to evaluate.
- Now define cross validation error as following which is a decent estimate.

$$\underline{\underline{E_{CV}}} = \frac{1}{N}\sum_{n=1}^{N} \underline{\underline{e_n}}$$

- Note, if you have large dataset (one million examples), it will not be possible for you to use this method (You will have to train the model one million times!)

K-fold Cross Validation

- Leave-One-Out results in N training session
- Instead break the data to a number of K folds.
- K training session on the remaining points each time.
- Rule of Thumb:10-fold cross validation => 10 training sessions


$\mathcal{D}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_3$
$\mathcal{D}_4$
$\mathcal{D}_5$
$\mathcal{D}_6$
$\mathcal{D}_7$
$\mathcal{D}_8$
$\mathcal{D}_9$
$\mathcal{D}_{10}$
train
validate
train