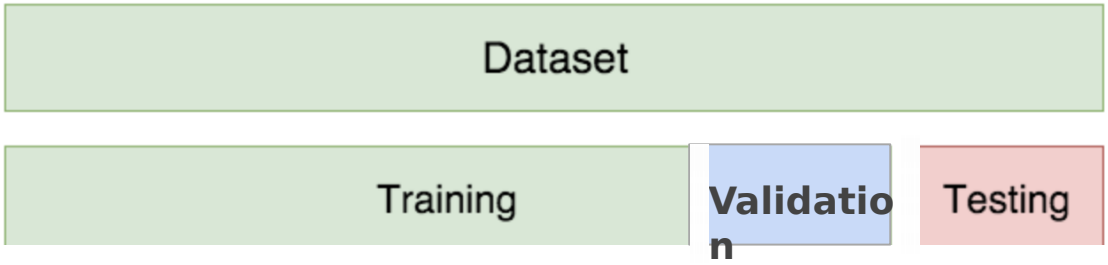
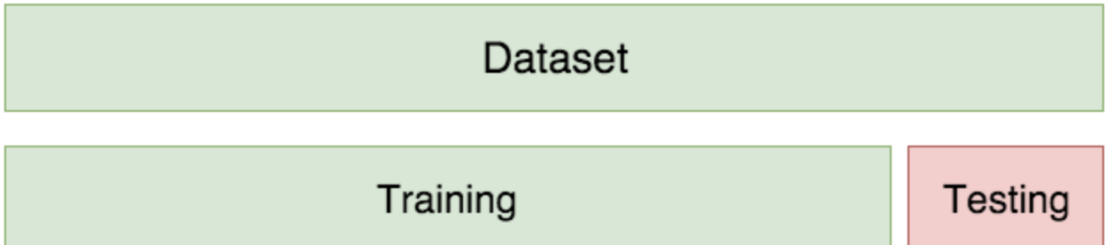
K Fold Cross Validation

K Fold Cross Validation



K Fold Cross Validation



Analytics Vidhya
Learn everything about analytics

K Fold Cross Validation : Procedure

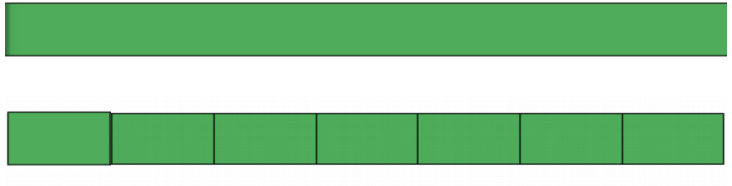
K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.

Analytics Vidhya
Learn everything about analytics

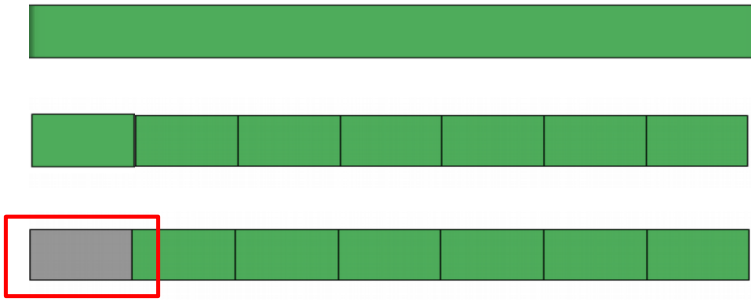
K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.
- Split the dataset into k groups



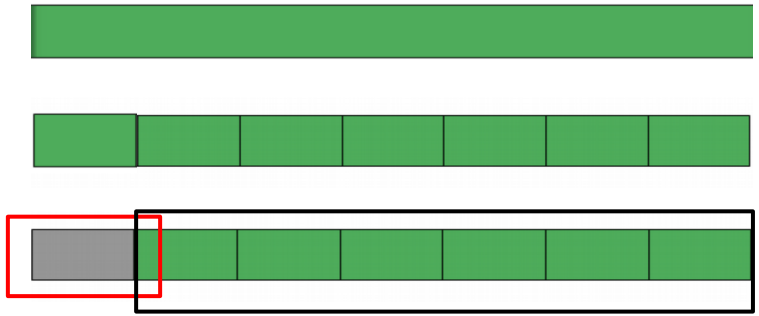
K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.
- Split the dataset into k groups
  - Pick a group as a hold out



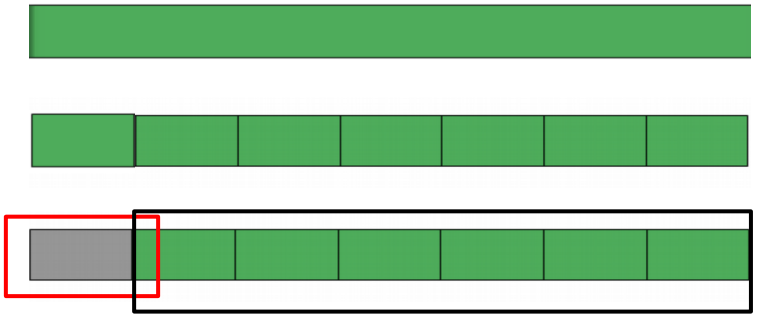
K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.
- Split the dataset into k groups
  - Pick a group as a hold out
  - Take the remaining groups as training and fit a model



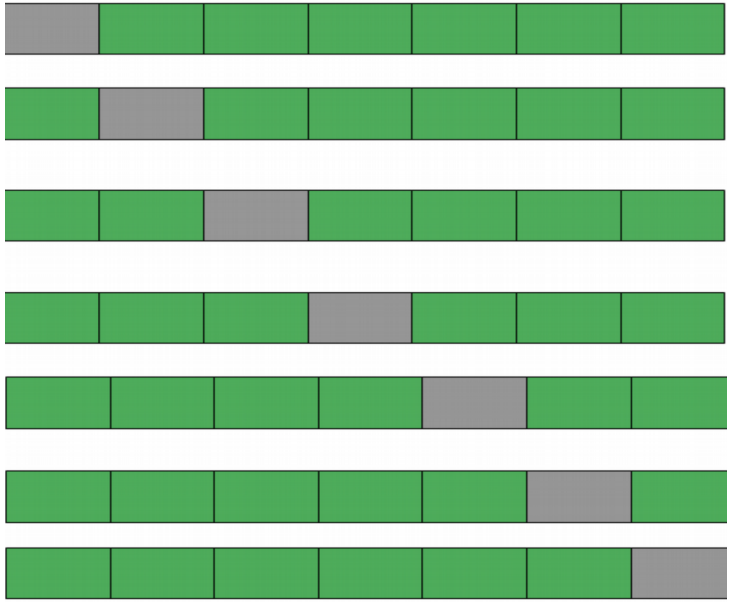
K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.
- Split the dataset into k groups
  - Pick a group as a hold out
  - Take the remaining groups as training and fit a model
  - Predict and evaluate on the hold out



K Fold Cross Validation : Procedure

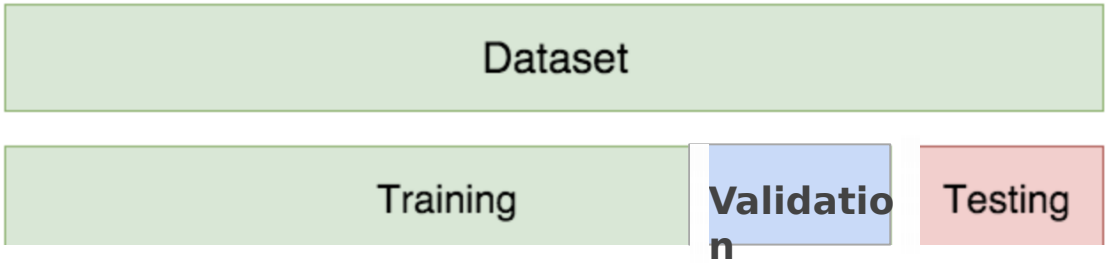
- Shuffle the dataset randomly.
- Split the dataset into k groups
  - pick a group as a hold out
  - Take the remaining groups as training and fit a model
  - Predict and evaluate on the hold out
- Repeat the above procedure with every group



Analytics Vidhya
Learn everything about analytics

K Fold Cross Validation : Advantages

K Fold Cross Validation : Advantages



K Fold Cross Validation : Advantages

- Prevents overfitting

Analytics Vidhya
Learn everything about analytics

K Fold Cross Validation : Advantages

- Prevents overfitting
- Consistency of model

K Fold Cross Validation : Deciding K

K Fold Cross Validation : Deciding K

- Generally: 5 < K < 10

K Fold Cross Validation : Deciding K

- Generally: 5 < K < 10
- Validation size = statistically significant

K Fold Cross Validation : Deciding K

- Generally: 5 < K < 10
- Validation size = statistically significant
- K = 2, 50-50 split, high bias

K Fold Cross Validation : Deciding K

- 5 < K < 10: generally
- Validation size = statistically significant
- K = 2, 50-50 split, high bias
- K > 10 , more data for training. More models

K = 10

K Fold Cross Validation : Deciding K

- 5 < K < 10: generally
- Validation size = statistically significant
- K = 2, 50-50 split, high bias
- K > 10 , more data for training. More models

K = 20

K Fold Cross Validation : Deciding K

- 5 < K < 10: generally
- Validation size = statistically significant
- K = 2, 50-50 split, high bias
- K > 10 , more data for training. More models

K = 30

K Fold Cross Validation : Deciding K

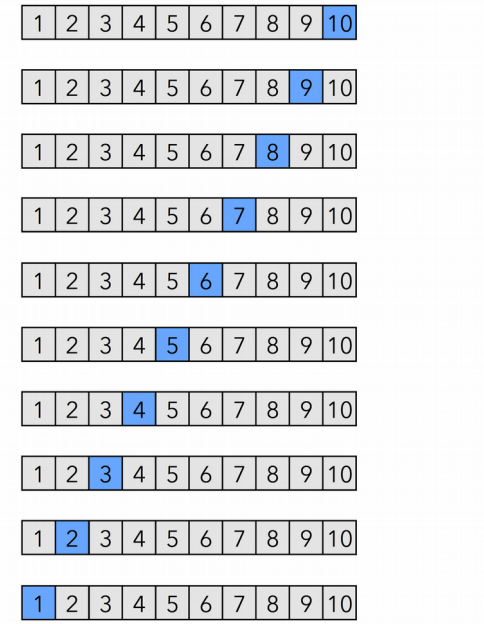
- 5 < K < 10: generally
- Validation size = statistically significant
- K = 2, 50-50 split, high bias
- K > 10 , more data for training. More models
- K = number of observations, extreme case

K = n

Analytics Vidhya
Learn everything about analytics

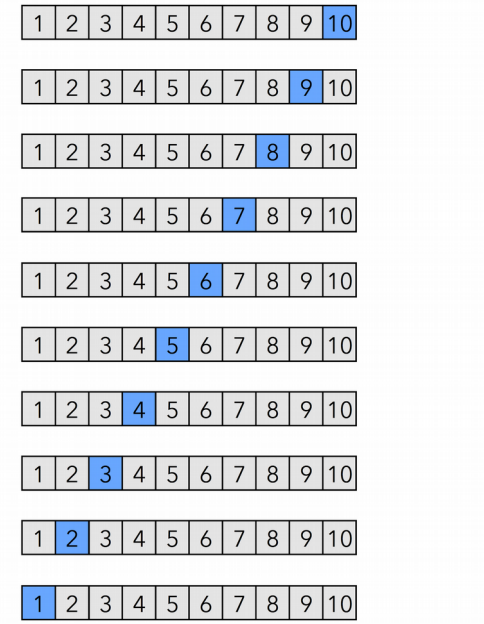
Leave one out Cross Validation

- N instances, n models



Leave one out Cross Validation

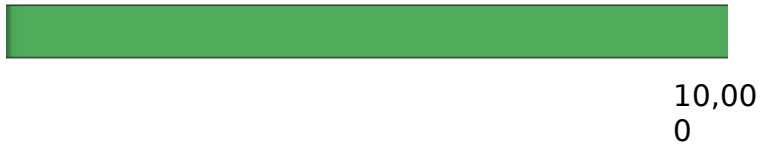
- N instances, n models
- not significant when dataset is large



Analytics Vidhya
Learn everything about analytics

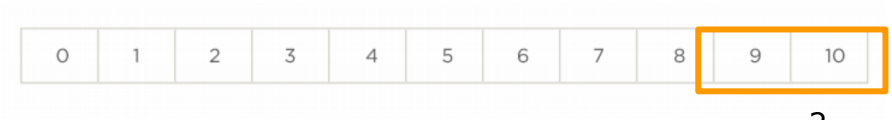
Leave one out Cross Validation

0	1	2	3	4	5	6	7	8	9	10



Analytics Vidhya
Learn everything about analytics

Leave one out Cross Validation



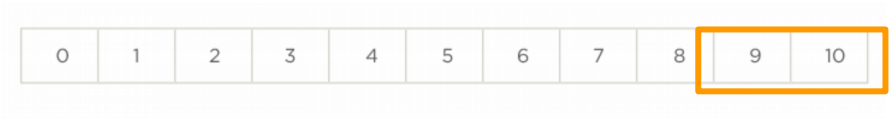
2 instances

2000 instances

K= 5

Analytics Vidhya
Learn everything about analytics

Leave one out Cross Validation



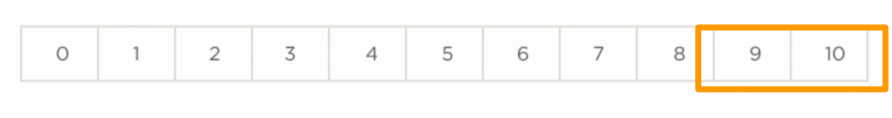
K=5

Very less data to train on

Significant amount of data

Analytics Vidhya
Learn everything about analytics

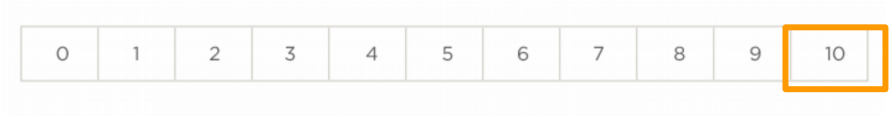
Leave one out Cross Validation



K=5

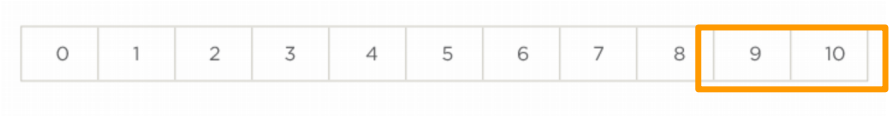
Very less data to train on

Significant amount of data



K=10

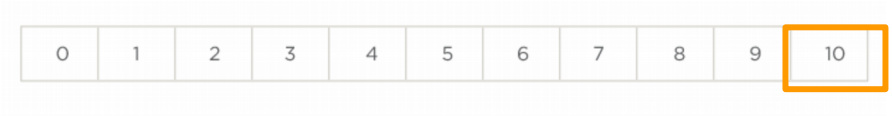
Leave one out Cross Validation



K=5

Very less data to train on

Significant amount of data



K=10

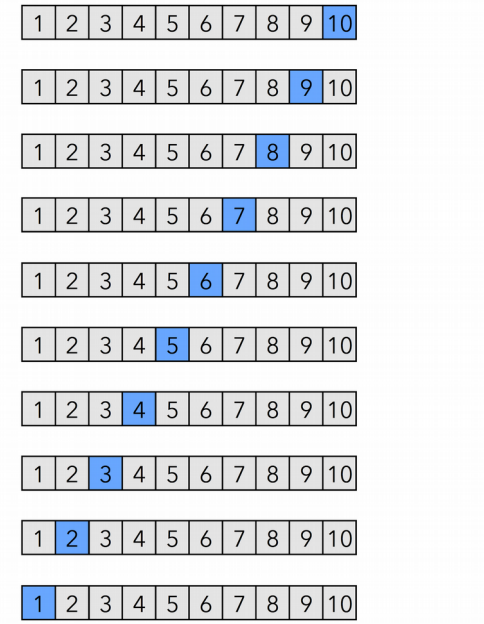
Increase in training is significant

Insignificant increase in training

Analytics Vidhya
Learn everything about analytics

Leave one out Cross Validation

- N instances, n models
- Not significant when dataset is large
- Best practise for small dataset



Leave one out Cross Validation

- N instances, n models
- Not significant when dataset is large
- Best practise for small dataset
- Can use leave ‘p’ out cross validation