Machine Learning

Pawel Wocjan

University of Central Florida

Spring 2019

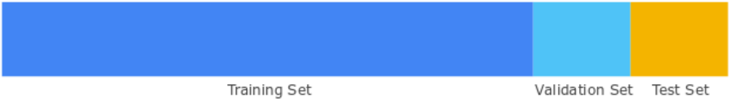
Validation set

- We introduced previously the partitioning a data set into a training set and a test set.
- This partitioning enabled you to train on one set of examples and then to test the model against a different set of examples.
- With two partitions, the workflow would look as follows:

```
Train model on Training Set --> Evaluate model on Test Set
       ^                                            |
       |                                            |
       +---------- Tweak model according <----------+
                   results on Test Set
                            |
                            V
                    Pick model that does best
                    on Test Set
```

Validation set

- Dividing the data set into two sets is a good idea, but it is not enough.
- You can greatly reduce the chances of overfitting by partitioning the data into three subsets shown below:



- Use the validation set to evaluate results from the training set.
- Then, use the test set to double-check your evaluation after the model has “passed” the validation set.

Validation set

- With three partitions, the workflow looks as follows:

```
Train model on Training Set --> Evaluate model on Validation Set
   ^                                                       |
   |                                                       |
   +---------------- Tweak model according <---------------+
                     results on Validation Set
                            |
                            V
               Pick model that does best --> Confirm model
               on Validation Set             on Test Set
```

- In this improved workflow:
  - Pick the model that does best on validation set.
  - Double-check that model against the test set.
- This is a better workflow because it creates fewer exposures to the data set.