CMPS 460 – Spring 2022

MACHINE LEARNING

Tamer Elsayed

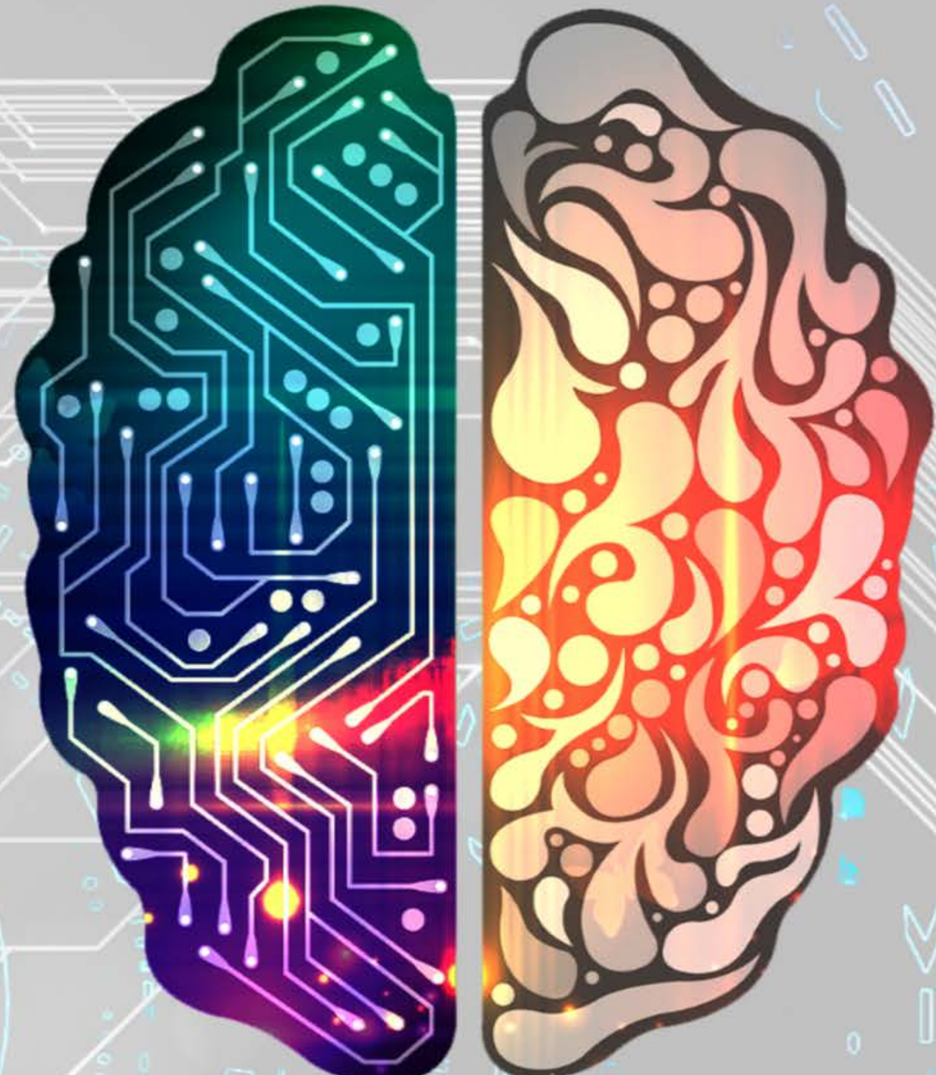
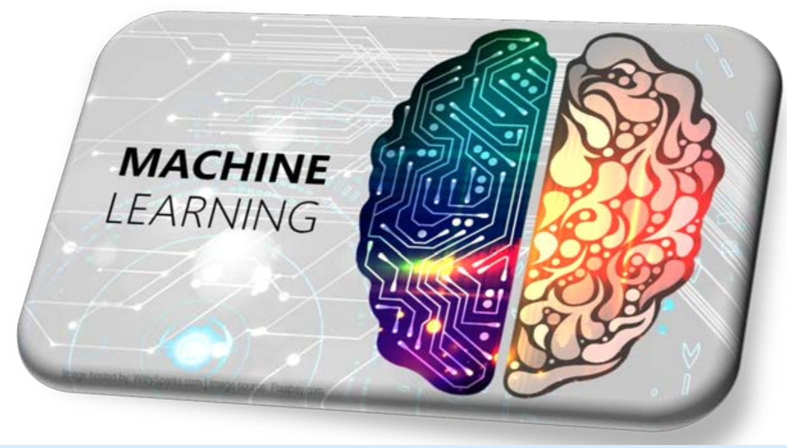


Image hosted by: WittySparks.com | Image source: Pixabay.com

5.c

Practical Issues: Cross Validation & Debugging

Sec 5.6, 5.8



Cross-Validation

Disadvantage of Train-Test split?

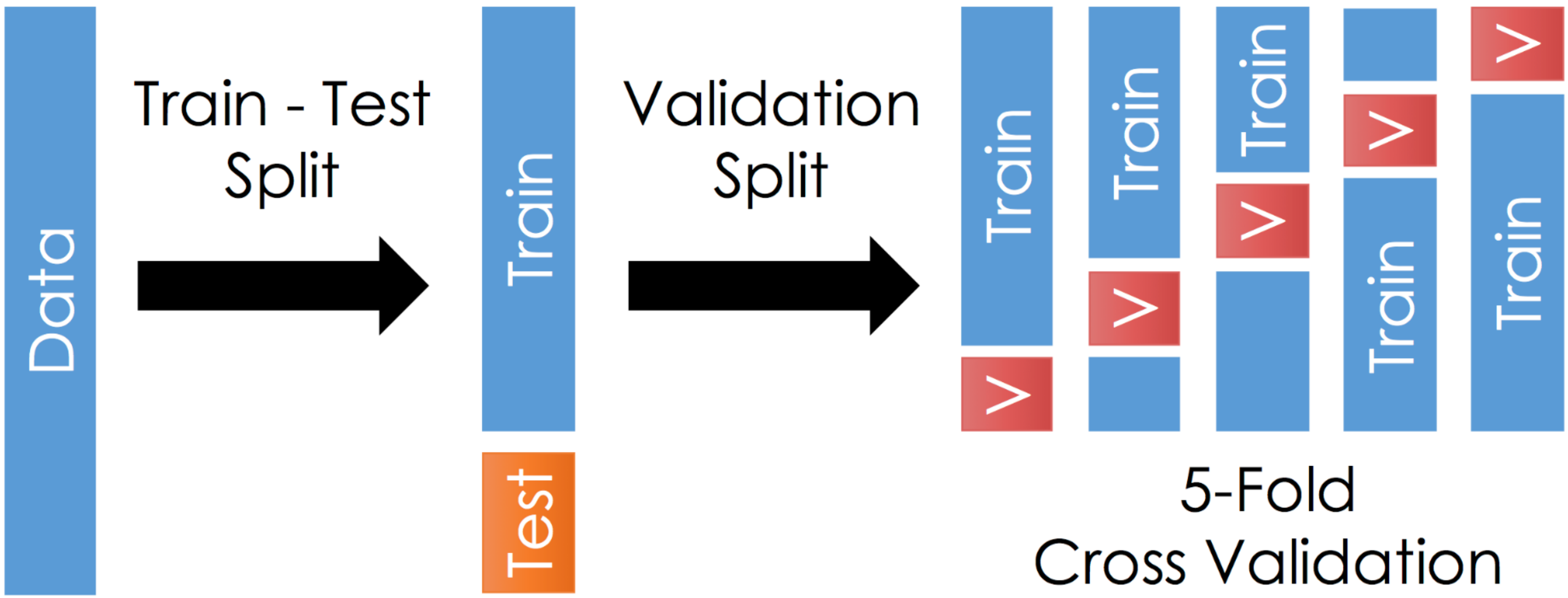
- Random split ==> Data might not be representative!
- Test only on part of the available data!

Solution: Cross-Validation

K-fold Cross-Validation

K=10 here

With Hyper-Parameters?



Algorithm 8 CrossValidate(*LearningAlgorithm*, *Data*, *K*)

```
ε̂ ← ∞                                          // store lowest error encountered so far
α̂ ← unknown                                    // store the hyperparameter setting that yielded it
for all hyperparameter settings α do
    err ← [ ]                                   // keep track of the K-many error estimates
    for k = 1 to K do
        train ← {(x_n, y_n) ∈ Data : n mod K ≠ k − 1}
        test ← {(x_n, y_n) ∈ Data : n mod K = k − 1}      // test every Kth example
        model ← Run LearningAlgorithm on train
        err ← err ⊕ error of model on test      // add current error to list of errors
    end for
    avgErr ← mean of set err
    if avgErr < ε̂ then
        ε̂ ← avgErr                              // remember these settings
        α̂ ← α                                   // because they're the best so far
    end if
end for
```

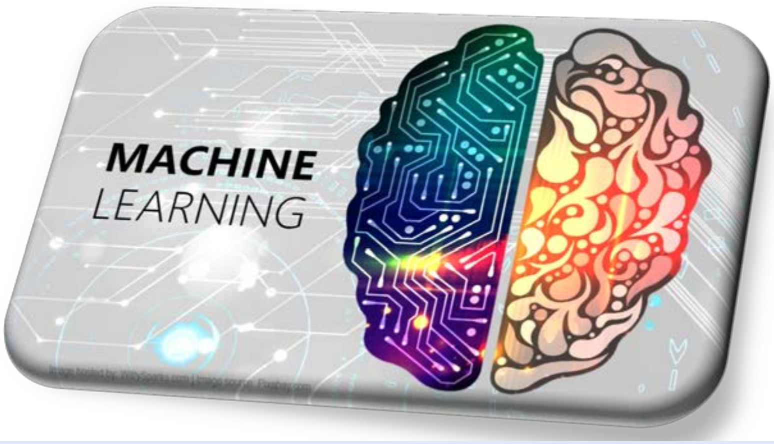
Leave-one-out (LOO) C.V.

- When K=N

Why would we do that?

Any drawbacks?

- Very natural for kNN.



Debugging Learning Algorithms


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Debugging!

- You've implemented a learning algorithm ..
- You try it on some train/dev/test data ..
- You get really bad performance ..

What's going on?!

- Is the data too noisy?
- Is the learning problem too hard?
- Is the implementation of the learning algorithm buggy?
- …

Strategies for Isolating Causes of Errors

- Is the problem with generalization to test data?
  - Can learner fit the training data?
  - Yes: problem is in generalization to test data
    - too complicated model family
    - not enough data
  - No: problem is in representation
    - need better features
    - better data

- Train/test mismatch?
  - Try reselecting train/test by shuffling training and test data together.

Strategies for Isolating Causes of Errors

- Is learning algorithm implementation correct?
  - Measure loss rather than accuracy (make sure it is minimized)
  - Hand-craft a toy dataset
  - (if possible) Compare against reference implementation
- Is representation adequate?
  - If you can’t fit training data, you might need better features.
  - Can you overfit if you add a cheating feature that perfectly correlates with correct class?
    - If not (near) 0% error: too many noisy features OR a bug!
    - If (near) 0% error: work on better features, or change learning model
- Do you have enough data?
  - Try training on less training set, how much does it hurt performance?