DLI Accelerated Data Science Teaching Kit

Lecture 14.5 - Overfitting and Cross Validation

An ideal model should correctly estimate:

- known or seen data examples' labels
- unknown or unseen data examples' labels

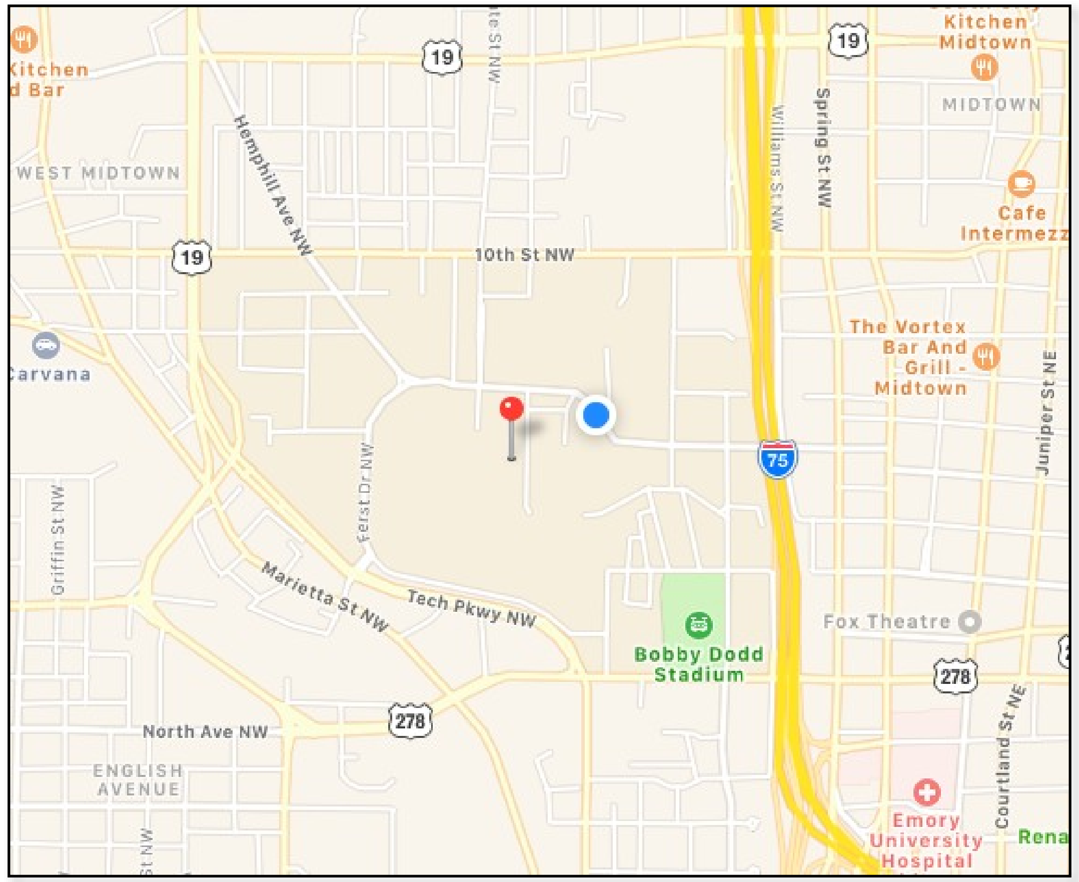
Song name	Artist	Length	...	Like?
Some nights	Fun	4:23	...	🙂
Skyfall	Adele	4:00	...	☹
Comf. numb	Pink Fl.	6:13	...	🙁
We are young	Fun	3:50	...	🙂
...	...	...	...	...
Chopin's 5th	Chopin	5:32	...	??

Training a classifier = building the “model”

Q: How do you learn appropriate values for parameters a, b, c, ... ?

(Analogy: how do you know your map is a “good” map?)

- $y_i = f_{(a,b,c,....)}(x_i), i = 1, ..., n$
  - Low/no error on training data (“seen” or “known”)
- $y = f_{(a,b,c,....)}(x)$, *for any new x*
  - Low/no error on test data (“unseen” or “unknown”)



Screenshot from Apple Maps

It is very easy to achieve perfect classification on training/seen/known data. Why?

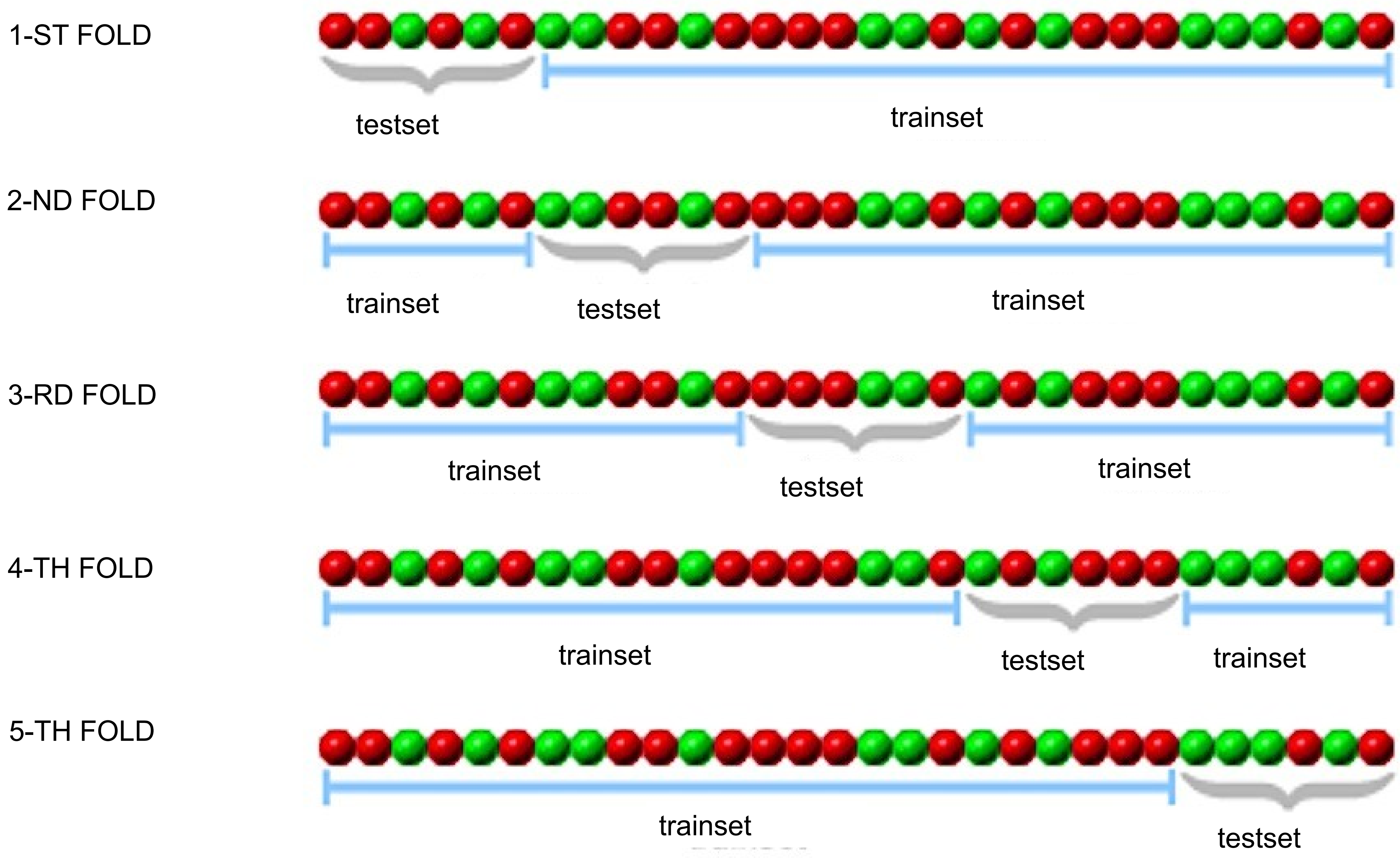
Over fitting

- If your model works really well for training data, but poorly for test data, your model is "over fitting".

- How to avoid over fitting?

One run of *5-fold* cross validation

You should do a few runs and compute the average
(e.g., error rates if that’s your evaluation metrics)



1-ST FOLD

testset trainset

2-ND FOLD

trainset testset trainset

3-RD FOLD

trainset testset trainset

4-TH FOLD

trainset testset trainset

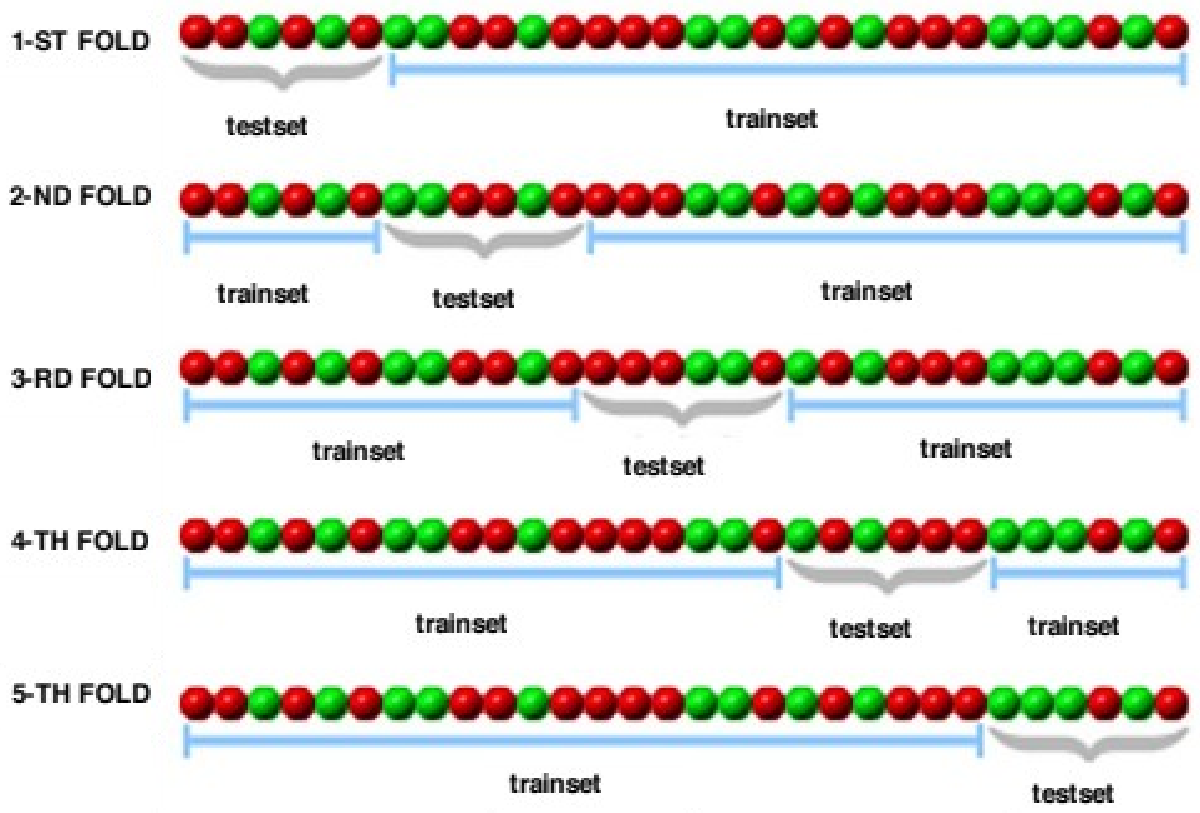
5-TH FOLD

trainset testset

Cross-Validation in plain english? (n.d.). Retrieved December 04, 2017, from http://stats.stackexchange.com/questions/1826/cross-validation-in-plain-english

Cross Validation

1. Divide your data into n parts
2. Hold 1 part as “test set” or “hold out set”
3. Train classifier on remaining n-1 parts “training set”
4. Compute test error on test set
5. Do the above steps n times, once for each n-th part
6. Compute the average test error over all n folds (i.e., cross-validation test error)



Cross-Validation Variations

K-fold cross-validation

- Test sets of size (n / K)
- K = 10 is most common (i.e., 10-fold CV)

Leave-one-out cross-validation (LOO-CV)

- test sets of size 1

DEEP LEARNING INSTITUTE

DLI Accelerated Data Science Teaching Kit

Thank You