Introduction to Machine Learning
Cross-validation

Ramon Fuentes[1,2]

August 6, 2019

[1]Visiting Researcher, Dynamics Research Group
The University of Sheffield

[2]Research Scientist, Callsign Ltd

Generalisation error

- In order to assess generalisation error, we must set aside a sample of our training data for evaluation
- However, training data can be scarce! Holding out data means it does not inform training... so maybe not a good idea?

Cross validation example

The idea in cross validation is to use all the available data as training and validation/testing, to assess generalisation performance

- Split available data into training and testing
- Train model
- Compute generalisation error
- Change the training/testing split

Cross-validation

Ideally, we would want a CV scheme where all samples are used for training and testing. However, this exhaustive search is expensive! Typical CV schemes are

- Leave-one-out (yes,... Loo)
- K-fold

Leave-one-out cross-validation

In Leave-one-out cross validation
for i in range $0 : n$:

- Remove sample i from training set
- Train model
- compute error on sample i (hold-out set)

iterate above for all values of λ
pick a λ that gives you small generalisation errors

K-fold Cross-validation

In K-fold cross-validation
Divide data into K different subsets at random
for k in range $0 : K$:

- Remove fold k from training set
- Train model
- compute error on fold k (hold-out set)

iterate above for all values of λ
pick a λ that gives you small generalisation errors

Cross-validation, what have we gained?

- We have alleviated over-fitting of models by choosing hyeperparameters that give us low generalisation errors.

Cross-validation, what have we gained?

- We have alleviated over-fitting of models by choosing hyeperparameters that give us low generalisation errors.
- We've balanced bias and variance whilst using all of our training data.