CHAPTER 5: VALIDATION

Grado en Ingeniería Informática

Curso 2014 / 15

Topics

1. Introduction

2. Validation methods

3. Summary

INTRODUCTION

Why validate?

- To test our model under all possible circumstances.
- However, it is not possible to have training data to simulate all possible situations.
- We have to obtain a model that is able to generalize from data available

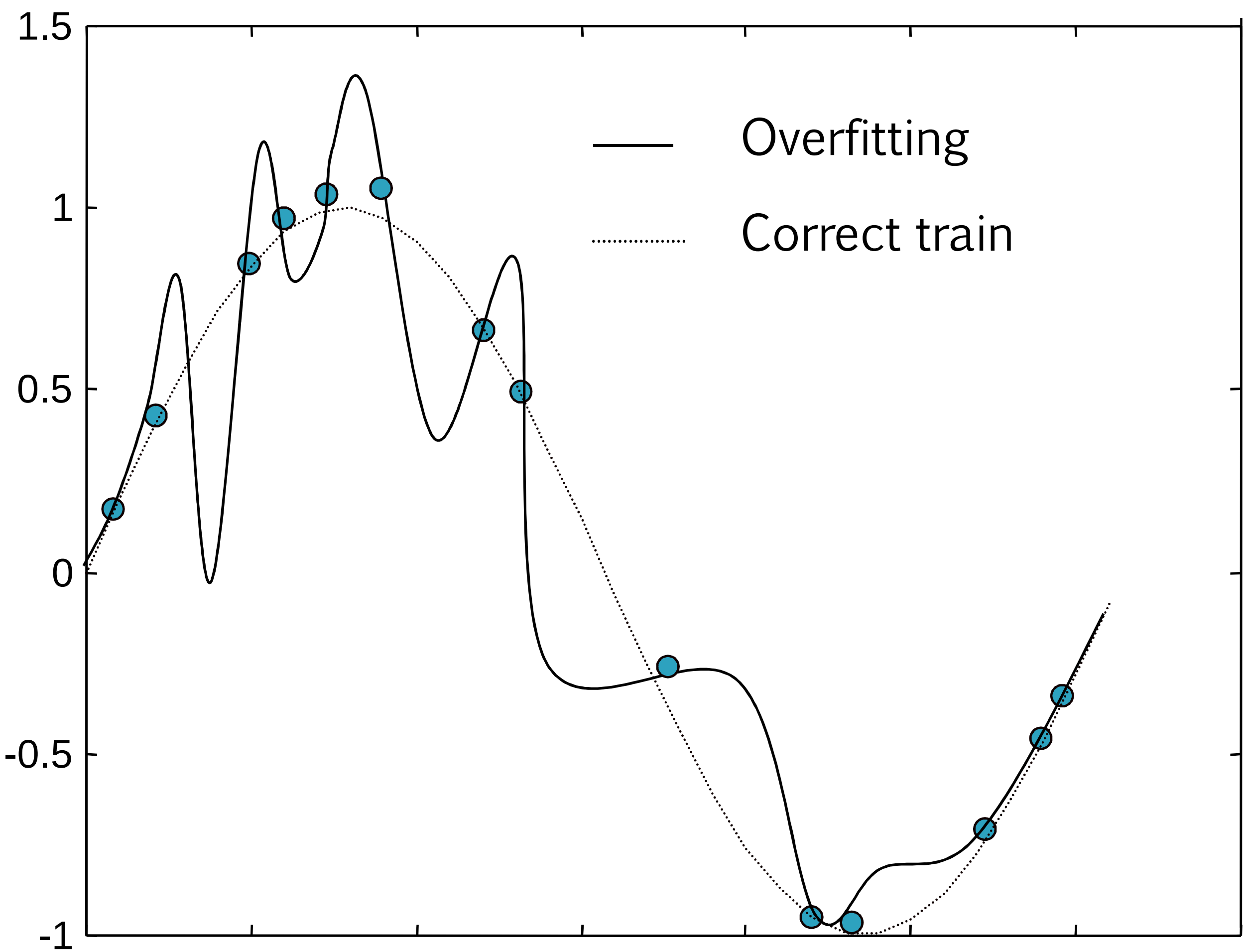
Poor generalization

- Small training set or unrepresentative

- Poor representation of data

- Wrong learning (excessive, possibility of local minima, etc.)

Over fitting effect



Training time

- When do you stop training the network?
  - If the error reaches a minimum
  - If the error reduction rate is below a certain value
  - After a certain number of epochs

- These methods do not guarantee a good generalization

- Validation tries to solve this problem

VALIDATION METHODS

Validation methods

- Replacement
- Simple validation
- Cross validation (Hold – K - out)
  - Hold – one – out
- Bootstrapping
- Stacked generalization

Replacement method

- Stop training when the error in the training data satisfies a certain condition

- Produce overtraining

- An "optimistic" measure the actual error occurs

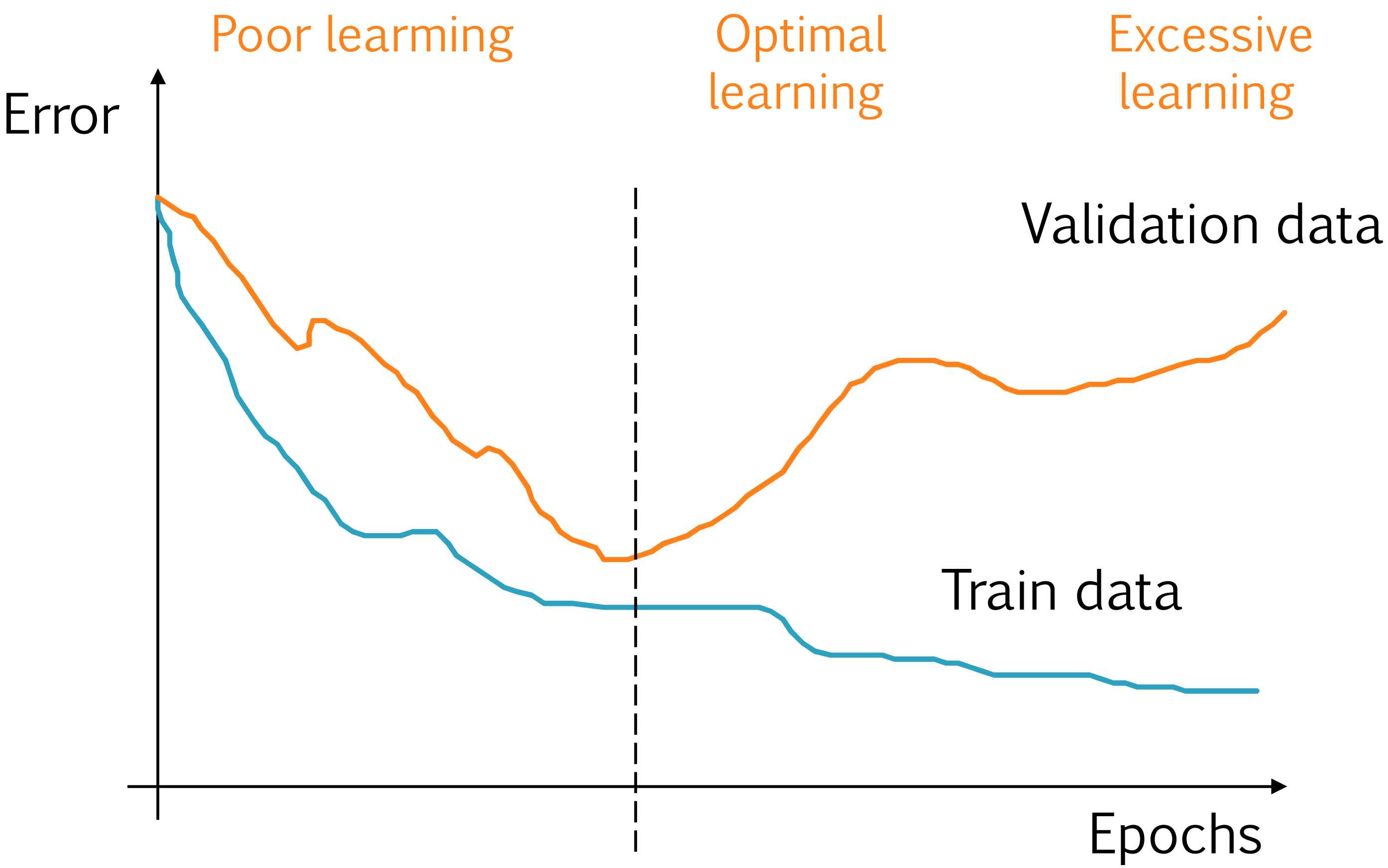
Simple validation method

- It is an integral part of the training process

- Divide the available data into three groups:
  - Training dataset (≈ 65%)
  - Validation data set (≈ 15%)
  - Test dataset (≈ 20%)

- It is known as the hold out method

Simple Validation. Optimal stopping point

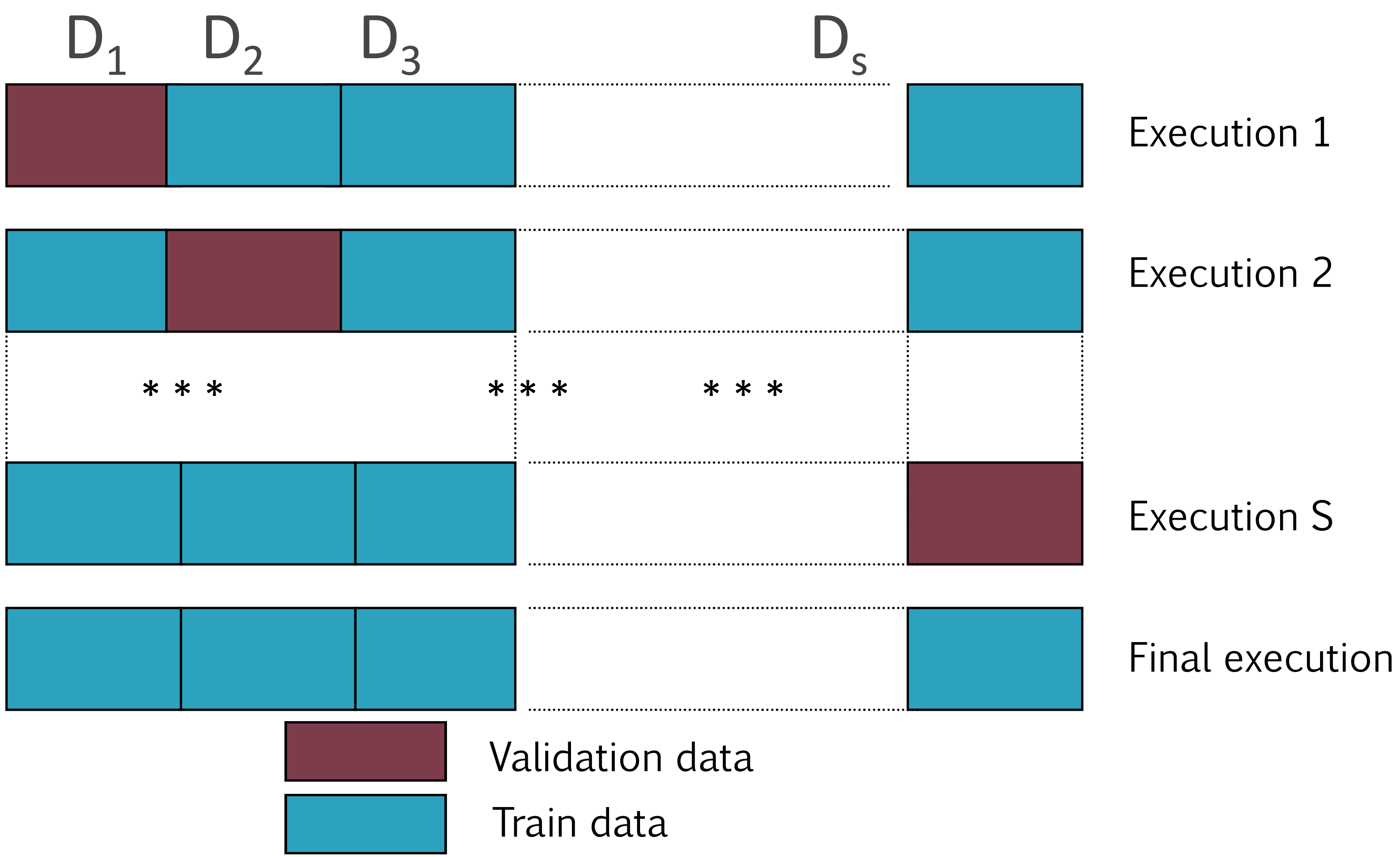


Cross validation(Leave – k – out / Hold – K – out)

Universidad de Cádiz

- Procedure
  - Divides the data into S groups of similar size.
  - A typical value for S is about 10
  - Train the network S times, each time leaving one set for validation.
- Can be used for:
  - Calculate the mean generalization error
  - Train the network, in which case, it is the end a workout with all available data, using the validation data obtained to stop the training prematurely

Cross validation. Graphical representation



Cross validation. Advantages and disadvantages

UCA
Universidad
de Cádiz

- Advantages:
  - Is superior to simple validation with small datasets.
  - Use all available data for training
- Disadvantages:
  - Accurate S + 1 training (increasing time)
  - The final training does not use validation directly

Hold – one - out / Leave – one - out

- It is a particular case of the cross-validation

- It consists of cross-validation, considering S = number of training patterns

- It works well for continuous error functions as the mean square error.

Bootstrapping

- It works with subsamples data instead of data subsets (as cross-validation)

- Each subsamples is a random example that can be replaced in the complete dataset

- In many cases works better than cross-validation.

Bootstrapping

- It is not easy to implement.
- The error estimator is statistically equivalent of leave-one-out
- The results are not entirely reliable.
- It does not work well for some methodologies (eg empirical decision trees)

Stacked generalization

- Combines network training with partitioning the dataset

- Modular system network

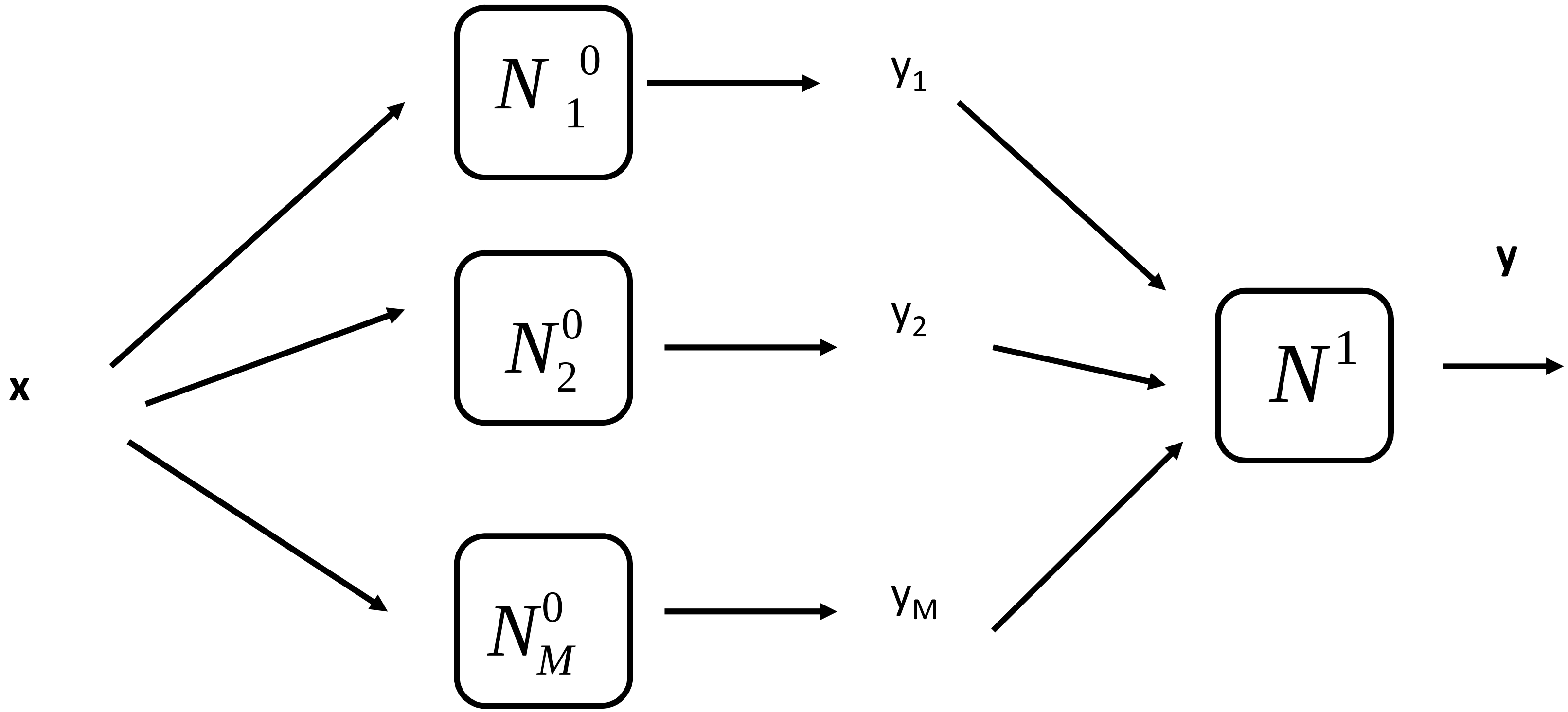
- It has M networks "level-0“ from N_1^0 to N_M^0 whose outputs are combined using a network of "level-1"

Stacked generalization

- The idea is to train the networks of level-0 first and see how it behaves in generalization

- This provides a new training set is used to train the network of level-1.

Stacked generalization. Graphic representa



SUMMARY

Summary

- It is essential to validate a model on data not used in training
- Ensure that the models and data sets are representative of the problem
- Use good estimators of the error
- The results should be averaged over many executions
- Finally, does the model approximate real situations?