Machine Learning and Intelligent Systems

Validation & Model Selection

Maria A. Zuluaga

Dec 8, 2023

EURECOM - Data Science Department

Table of contents

Validation

Generalization and Model Selection

- **Generalization:** Ability of a model to perform well on unseen data

$$\epsilon = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[l(y, h(\mathbf{x}))]$$

Generalization loss

- **Model Selection:** Task of selecting a model from a set of candidate models given the data

Generalization

- We just saw that it is important to look at the generalization/test error
- The training error is not enough to guarantee the good performance of a ML model

Generalization

- We just saw that it is important to look at the generalization/test error
- The training error is not enough to guarantee the good performance of a ML model

- We knew that already since lecture 1...

From Lecture 1: Setup - Slide 19

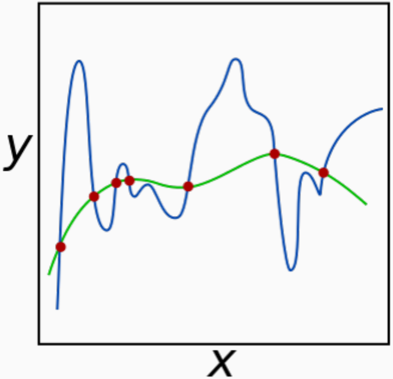
- We split $\mathcal{D}$ into three sets:
  - Training set $\mathcal{D}_{TR}$ - Used to learn h
  - Validation set $\mathcal{D}_{VAL}$ - To check for overfitting
  - Test set $\mathcal{D}_{TEST}$ - Used to evaluate the chosen h and have an estimate of the **generalization error** or loss

- Typical splits are 70/10/20, 80/10/10, 60/20/20.
- If the samples are drawn i.i.d. from the same distribution P, then the testing loss is an unbiased estimator of the true generalization loss.

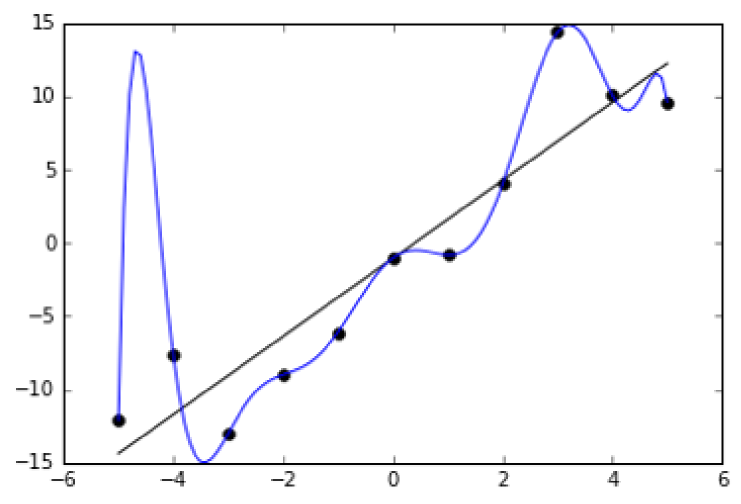
In this part of the lecture: How we use these data splits to perform model selection

Model Selection

- For a set of candidate models we choose that one with the smallest test error
- **Reminder:** We prefer simpler models
- Therefore, we might choose:
  - Slightly higher validation errors
  - Simpler models

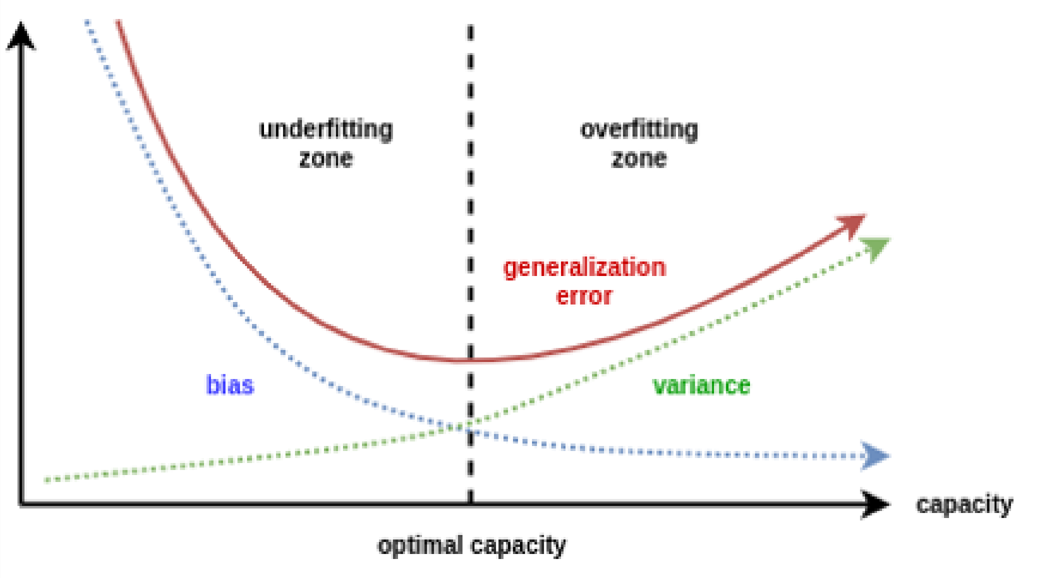


Example



Which model would you choose?

Source: Wikipedia

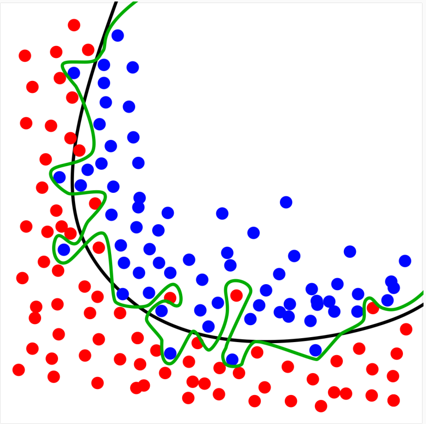


Source: https://djsaunde.wordpress.com/2017/07/17/the-bias-variance-tradeoff/

Overfitting

Reminder...

- **Overfitting** occurs when a model fits the data too well
- It is associated to models of high complexity
- It will lead to failure to generalize
  Training error $<$ Testing error

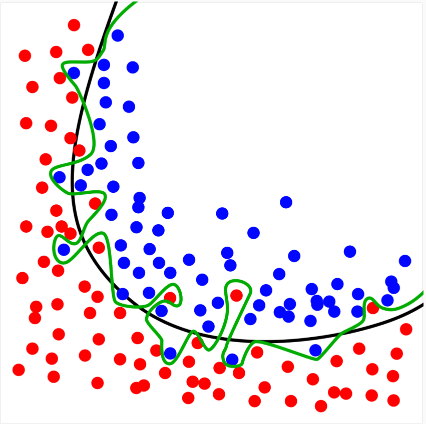


Overfitting

Reminder...

- **Overfitting** occurs when a model fits the data too well
- It is associated to models of high complexity
- It will lead to failure to generalize
  Training error $<$ Testing error

- **Underfitting** occurs when a model cannot adequately capture the underlying structure of the data



Overfitting & Model Selection

- We saw that regularization is a good way to avoid overfitting

Overfitting & Model Selection

- We saw that regularization is a good way to avoid overfitting
- But, how to choose the parameters introduced by regularization?

Overfitting & Model Selection

- We saw that regularization is a good way to avoid overfitting
- But, how to choose the parameters introduced by regularization?
- Validation provides us with a solution

Validation

1. Split $\mathcal{D}$ into $\mathcal{D}_{TR}, \mathcal{D}_{VAL}$ and $\mathcal{D}_{TEST}$
2. Train candidate models using $\mathcal{D}_{TR}$, e.g. different λ for regularization, network hyper-parameters
3. Use $\mathcal{D}_{VAL}$ to evaluate the candidate models
4. Pick the best
5. Retrain the best using $\mathcal{D}_{TR} + \mathcal{D}_{VAL}$
6. Test the generalization capabilities using $\mathcal{D}_{TEST}$

Validation

1. Split $\mathcal{D}$ into $\mathcal{D}_{TR}, \mathcal{D}_{VAL}$ and $\mathcal{D}_{TEST}$
2. Train candidate models using $\mathcal{D}_{TR}$, e.g. different λ for regularization, network hyper-parameters
3. Use $\mathcal{D}_{VAL}$ to evaluate the candidate models
4. Pick the best
5. Retrain the best using $\mathcal{D}_{TR} + \mathcal{D}_{VAL}$
6. Test the generalization capabilities using $\mathcal{D}_{TEST}$

Drawback

- Easy when there is a very large amount of data
- Was the split the good one?

Cross-Validation

Better known as K-fold cross-validation

Algorithm

1. Split the data into $\mathcal{D}_{TR}, \mathcal{D}_{TEST}$
2. Split $\mathcal{D}_{TR}$ into K-folds
3. For each fold $k \in \{1, \ldots, K\}$, a candidate model is trained in all but the k^{th} fold
4. Test on the k^{th} fold
5. Average the error across folds
6. Use the resulting average error of each candidate model to select one
7. Retrain the chosen one using $\mathcal{D}_{TR}$
8. Test the generalization capabilities using $\mathcal{D}_{TEST}$

Cross-Validation

Better known as K-fold cross-validation

Algorithm

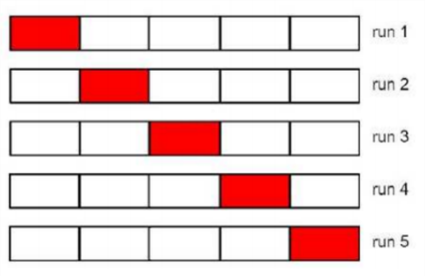
1. Split the data into $\mathcal{D}_{TR}, \mathcal{D}_{TEST}$
2. Split $\mathcal{D}_{TR}$ into K-folds
3. For each fold $k \in \{1, \ldots, K\}$, a candidate model is trained in all but the k^{th} fold
4. Test on the k^{th} fold
5. Average the error across folds
6. Use the resulting average error of each candidate model to select one
7. Retrain the chosen one using $\mathcal{D}_{TR}$
8. Test the generalization capabilities using $\mathcal{D}_{TEST}$

Note

If $K = N$, it is denoted leave-one-out CV (LOOCV)

K-fold Cross-validation

- CV gives an idea of the variability of the test error
- It can assess stability of the method by looking at the models parameter obtained for each fold
- A common value for K is 5



Using CV Properly

- Checking generalization and doing model selection should be two different tasks
- **Model selection:** Estimates the performance of different models in order to choose the best one (validation set via CV)
- **Model assessment:** Having chosen a final model, estimates its prediction error (generalization) on new data (test set)

How to Look for Hyper-parameters?

Coarse-to-fine

- First find the best order magnitude (e.g $\lambda = 0.01, 0.1, 1, 10, \ldots$)
- Once the good order is identified, do a fine search around that value

How to Look for Hyper-parameters?

Coarse-to-fine

- First find the best order magnitude (e.g $\lambda = 0.01, 0.1, 1, 10, \ldots$)
- Once the good order is identified, do a fine search around that value

Grid search

- Useful when there are multiple hyper-parameters to set
- Fix a set of values for each of them and try every combination
- Drawback: Computationally expensive

How to Look for Hyper-parameters?

Coarse-to-fine

- First find the best order magnitude (e.g $\lambda = 0.01, 0.1, 1, 10, \ldots$)
- Once the good order is identified, do a fine search around that value

Grid search

- Useful when there are multiple hyper-parameters to set
- Fix a set of values for each of them and try every combination
- Drawback: Computationally expensive

Random search

- Alternative to grid search
- Parameters are selected randomly within pre-defined intervals

Feature Selection

Feature Selection

- Given D features, feature selection can be seen as a special case of model selection where there are 2^D models to choose from
- **Question:** Why you might be interested in reducing the number of features?
- For large values of D this task can be computationally expensive
- Typically heuristic search procedures are used to find the best subset.

Algorithm

```
𝓕 = {}
for i=1..D
   if i ∉ 𝓕
      𝓕_i = 𝓕 ∪ {i}
      Train model using 𝓕_i
      Estimate generalization error using CV
      If generalization error improves
         𝓕 = 𝓕_i
return 𝓕
```

Other Algorithms

Wrapper model feature selection: Wraps around the learning algorithm

- Forward search
- Backward search

Other Algorithms

Wrapper model feature selection: Wraps around the learning algorithm

- Forward search
- Backward search

Filter feature selection: Computes a score that measures how informative is a feature

- Information theory approaches, e.g. mutual information

Other Algorithms

Wrapper model feature selection: Wraps around the learning algorithm

- Forward search
- Backward search

Filter feature selection: Computes a score that measures how informative is a feature

- Information theory approaches, e.g. mutual information

Other?

Wrap-up

Wrap-up

- We presented the problem of model selection
- We presented cross-validation as a way to perform model selection and assess a model's generalization
- We introduced feature selection

Key Concepts

- Generalization
- Model Selection
- Cross-Validation
- K-fold
- Coarse-to-fine, grid and random search
- Feature selection

References

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Ch 3, 4, 7
The Elements of Statistical Learning	Sec. 11.5 - Training of Neural Networks
Sci-kit Learn	Model Selection and Evaluation
Selection bias in the reported performances of AD classification pipelines	(link)