Model Selection and Prediction Accuracy

Eltecon Data Science Course by Emarsys

Holler Zsuzsa

November 11, 2020

Homeworks

- Presenters:
  - Alexandrov Dániel - Földesi Attila
  - Nguyen Thai Duong - Szentistványi János
  - Kovács Ádám - Nguyen Nam Son

Goal of the lesson

- Intro to the **theory of model selection**, model complexity, overfitting, etc.
- Understand the concept through real life examples
- Cover most commonly used **practical solutions** to the model selection problem
- Get some hands-on experience

Section 1

Model Selection in Theory

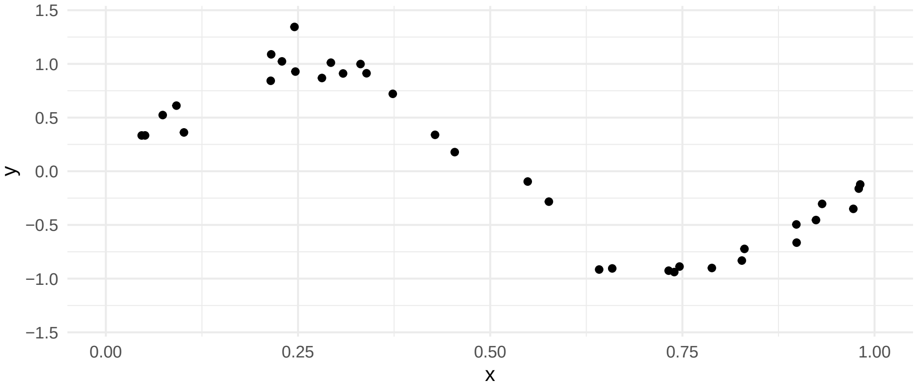
How to Select the Best Model

Goal: Good generalisation i.e.: best predictive performance on new data

What if I choose the one with the lowest error ($RMSE$)/ best fit (R^2)?

How to select the best type of model for our application?

How to Select the Best Model



The Loss Function

Common choice for regression problem is the **squared loss**:

$$L(f(x), y) = (f(x) - y)^2$$

Goal is to choose $f(x)$ that **minimises the expected loss**:

$$E[L(f)] = E[(f(x) - y)^2]$$

The Empirical Loss Minimiser

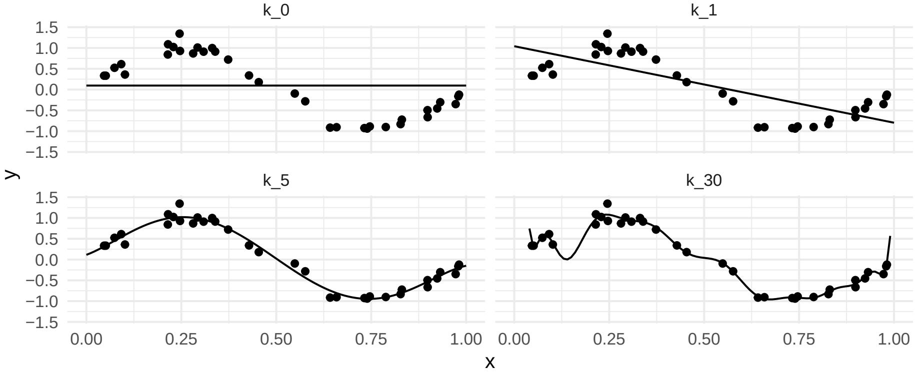
Assume you choose to approximate the relationship with a linear function with k variables.

The **empirical loss** of the fitted model:

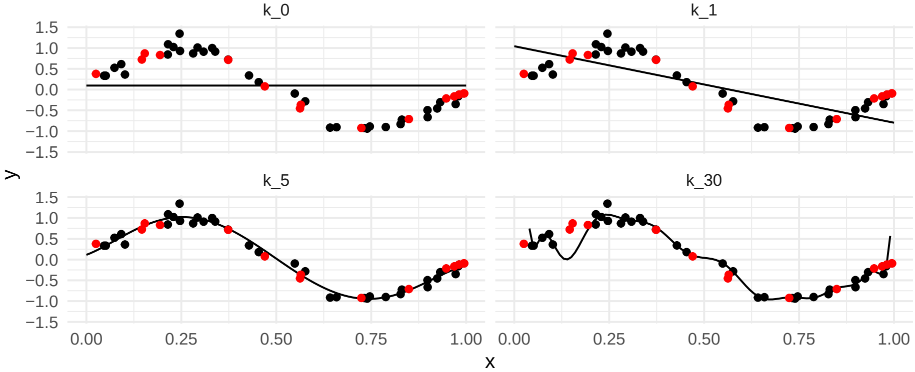
$$\hat{L}(f_k) = \frac{1}{n}\sum(f_k(x) - y)^2$$

Is this a good estimate of the expected loss of $f_k(x)$? Beware of overfitting!

The Empirical Loss Minimiser



The Empirical Loss Minimiser



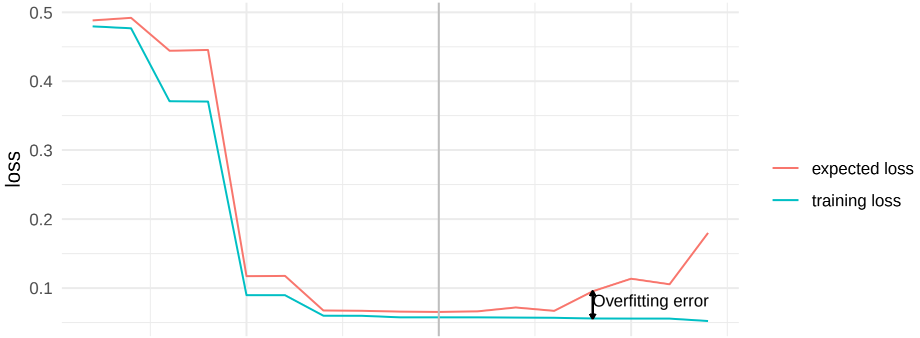
What is overfitting

Among a set of possible models we choose one that is too complex and has poor generalisation properties.

Why? Because we have an incorrect estimate of its expected loss.

Overfitting error:

$$E[L(f_k)] - \hat{L}(f_k)$$



Model Complexity in Practice

- "Classic" variable selection: Which explanatory variables should I include?
- Functional form selection: In what form should I include my variables?
- Tree models: How complex tree structure should I allow?
- Deep learning: How complex neural network should I train?

Model Complexity in Practice

Take the bike rental example from last time.

How should we incorporate the information on the time of the day?

- include “hour” variable as it is
- create a dummy variable for each value of hour
- include “hour” as a third degree polynomial

Task: Order the listed options by model complexity. Share your results in Socrative!

Model complexity

How to avoid overfitting?

Find the ideal level of **model complexity** within a given model type (e.g.: choose k for linear regression) for a **given set of data**.

$$E[L(f_k)] - E[L(f^*)] = \underbrace{[E[L(f_k)] - E[L(f_k^*)]]}_{\text{estimation error}} + \underbrace{[E[L(f_k^*)] - E[L(f^*)]]}_{\text{approximation error}}$$

where f_k^* is the best estimator among models with complexity k.

Section 2

Model Selection in Practice

Train vs. Test Error

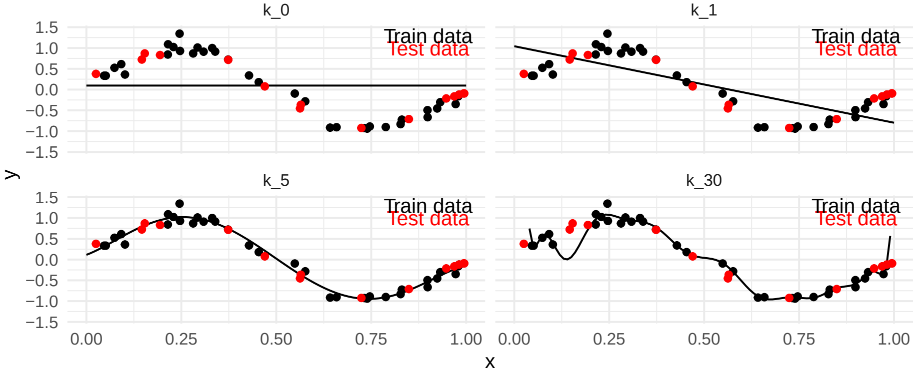
Idea: have an independent sample to estimate the performance of the fitted model

Training set: N observations of labeled data used to tune the parameters of the model (e.g.: estimate coefficients of linear regression)

Validation set/Test set: M observations of data used to optimize model complexity and/or choose between different types of models

Watch out for use-cases where random assignment does not work!

Train vs. Test Error



Train vs. Test Error

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{f}(x) - y)^2}$$

	train RMSE	test RMSE
pred0	0.86	0.74
pred1	0.68	0.68
pred5	0.33	0.29
pred30	0.27	1.07

SMS Spam Prediction Dataset

- Source: Kaggle
- Goal: Predict if SMS was a spam using text of the SMS

Pre-cleaned the data (removed stopwords, special characters etc.) and created word count variables: **spam_clean.csv**

is_spam	message	nchar	nwords
0	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...	111	12
0	Ok lar... Joking wif u oni...	29	4
1	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's	155	20
0	U dun say so early hor... U c already then say...	49	6
0	Nah I don't think he goes to usf, he lives around here though	61	8
1	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv	148	16

plus top 400 most frequent words.

SMS Spam Prediction

Let's see some prediction models! **spam_pred_train_test.R**

Practice Time

- Task: include all the available variables and compute train and test accuracy!
- Share your results in Socrative!
- You have 15 minutes - feel free to take a break if needed.

Train vs. Test Error

Advantages:

- Simple approach

Disadvantages:

- Loss of valuable training data
- Small validation set gives noisy estimate of predictive performance

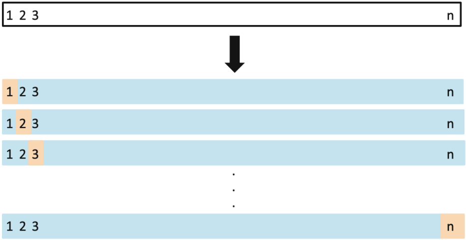
Overfitting to the validation set??? Possible!

One may want to set aside a third set of data to assess the performance of the final model.

Cross validation

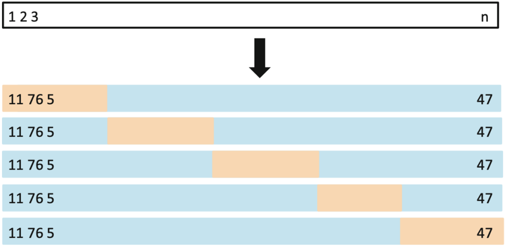
Idea: Instead of having a single validation set split the data multiple times to estimate the performance of the fitted model

Leave-one-out: split tha data N times, always leave one observation out for testing



Cross validation

K-fold: split the data into k sub-samples of equal size and leave one out for testing



How to choose k? Larger k results in larger variance in the error estimation but provides nearly unbiased estimate of the performance of the fitted model. ($k = 5$ is a common choice)

Cross validation

$$CV_k = \frac{1}{k} \sum MSE_i$$

	train MSE	test MSE	CV MSE
pred0	0.73	0.55	0.46
pred1	0.46	0.47	0.22
pred5	0.11	0.08	0.01
pred30	0.07	1.15	0.93

SMS Spam Prediction

Let's do cross-validation for our spam prediction models!
spam_pred_cv.R

Practice Time

- Task: Compute CV accuracy for all models we tested and compare their performance!
- Share your results in Socrative!
- You have 10 minutes - feel free to take a break if needed.

Cross validation

Advantages:

- utilizes all the data
- suitable for parameter tuning
- can decrease variance of the error estimation

Disadvantages:

- computationally expensive

Homework

- If you haven't finished computing CV accuracies, do so.
    -
- Presenters:
    - Bakirov, Aslan - Yatsenko, Anzhelika
    - Both Márton - Kamenár Gyöngyvér
    - Emerson, Ian - Ralbovszki Judit

Resources

- Bishop, Christopher: Pattern Recognition and Machine Learning
- Gareth J., Witten D., Hastie T. and Tibshirani R.: An Introduction to Statistical Learning