Introduction to Statistical Learning

Eltecon Data Science Course by Emarsys

Holler Zsuzsa

November 13, 2019

Goal of the lesson

- cover the basics of theory of model selection
- train and assess the quality of linear/logistic regression models in `R`

Section 1

Model Selection in Theory

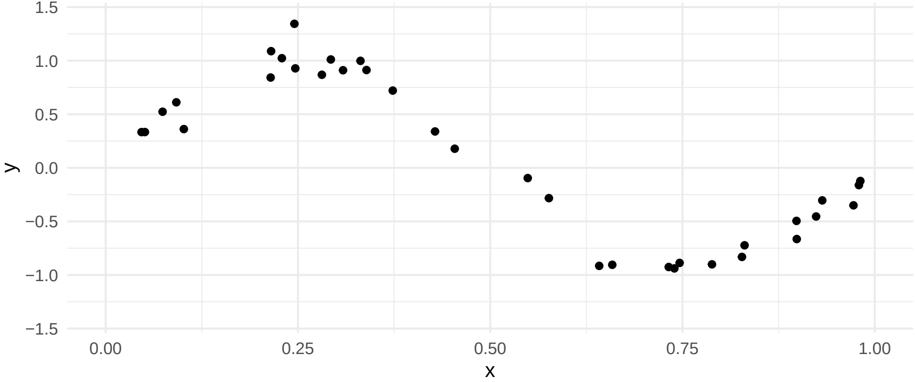
How to Select the Best Model

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

What if I choose the one with the lowest error (RSE)/ 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**:

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

On can show that the:

$$f^*(x) = argmin_{f(x)} E[L(f(x), y)] = E[y|x]$$

The Empirical Loss Minimiser

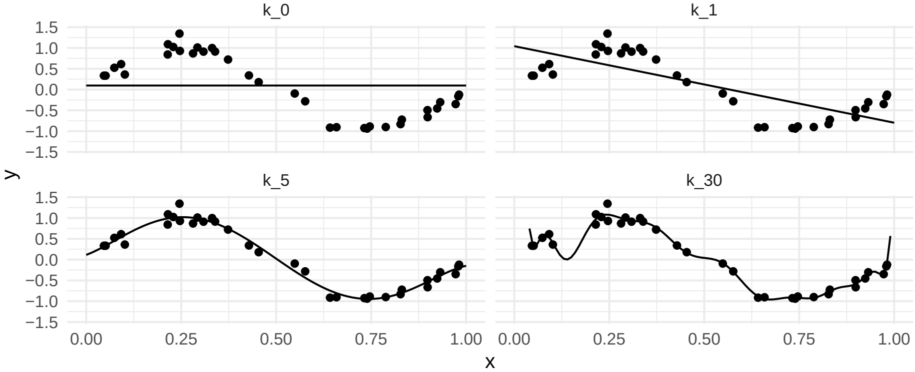
Assume you choose to approximate the relationship with a linear function with k variables.

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

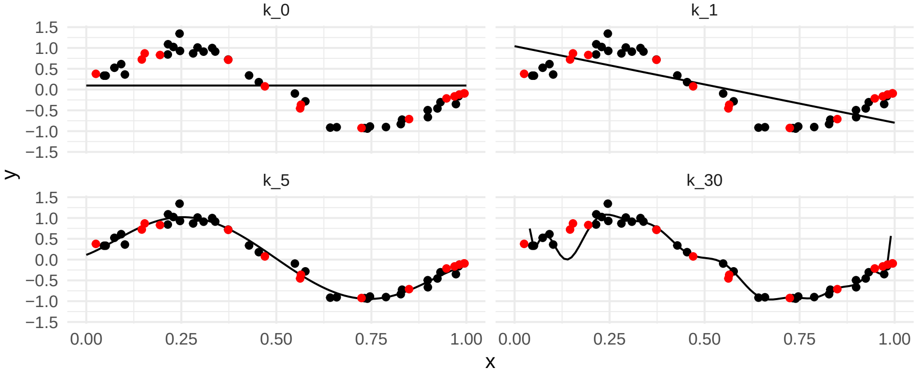
$$\hat{L}(f_k) = \frac{1}{n} \sum (f_k(x) - y)^2$$

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 with 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

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

Train vs. Test Error

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

	train MSE	test MSE
pred0	0.54	0.30
pred1	0.21	0.22
pred5	0.01	0.01
pred30	0.01	1.32

Train vs. Test Error

Advantages:

- Simple approch

Disadvantages:

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

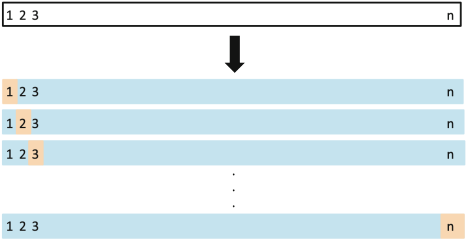
Overfitting to the validation set??? Possible!

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

Cross validation

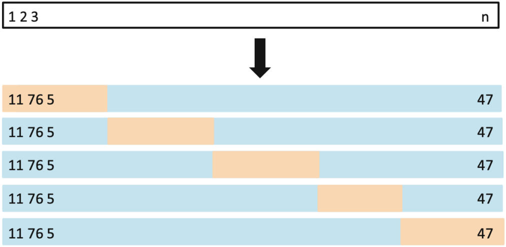
Idea: Instead of having a single validation set spli the data multiple times to estimate the performance of the fitted model

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	0.54	0.30	0.46
pred1	0.21	0.22	0.21
pred5	0.01	0.01	0.01
pred30	0.01	1.32	0.45

Cross validation

Advantages:

- utilizes all the data

Disadvantages:

- computationally expensive

Information criteria

Idea: Penalize model complexity by adding a penalty term.

Definition:

- **BIC** (Bayesian approach):

$$-\ln(\hat{L}) + \frac{1}{2} k \ln(N)$$

- **AIC** (Information theory):

$$-2\ln(\hat{L}) + 2k$$

where k is the number of parameters, N is the number of data points and $\hat{L}$ is the maximal value of the likelihood function.

Information criteria

Advantages:

- No need to set aside data for validation
- No need to train models multiple times

Disadvantages:

- Rely on assumptions that are often invalid in practice
- In practice, they tend to favor overly simple models

Information criteria

	train MSE	test MSE	CV MSE	AIC	BIC
pred0	0.54	0.30	0.46	81.65	84.76
pred1	0.21	0.22	0.21	51.08	55.74
pred5	0.01	0.01	0.01	-40.79	-29.90
pred30	0.01	1.32	0.45	-42.63	-11.53

Regularisation

Idea: Add a penalty term to the error function to discourage the coefficients from reaching large values.

$$E(w) = E_D(w) + \lambda E_W(w)$$

where $E_D(w)$ is the data-dependent error, $E_W(w)$ regularisation term and λ is the regularisation parameter that controls the relative importance of these two terms.

Regularisation

Advantages:

- allows to train complex models on limited size data
- computationally cheap (not always true)

Disadvantages:

- not clear how to choose λ

More on ridge, LASSO, the Bias-Variance trade-off later...