Regularisation

Eltecon Data Science Course by Emarsys

Holler Zsuzsa

November 13, 2019

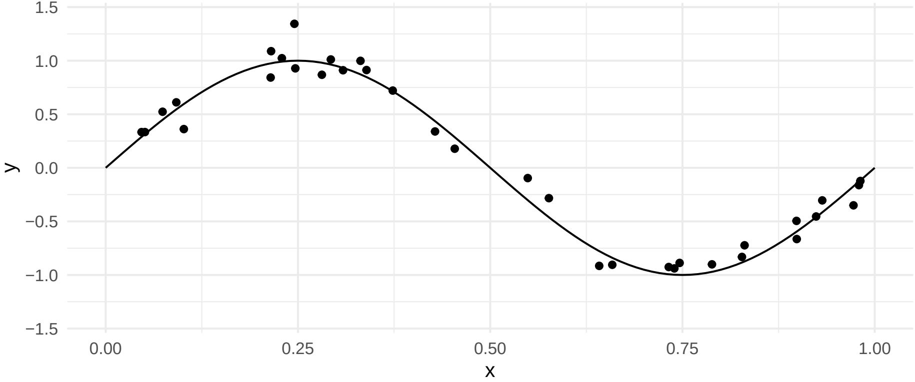
Goal of the lesson

- introduce the concept of **regularisation**
- define and try out **ridge** and **LASSO** regression
- conduct model selection on a real world example

Section 1

Regularisation

Recap



Recap

	train MSE	test MSE	CV MSE	AIC	BIC
k_0	0.54	0.30	0.46	81.65	84.76
k_1	0.21	0.22	0.21	51.08	55.74
k_5	0.01	0.01	0.01	-40.79	-29.90
k_30	0.01	1.32	0.45	-42.63	-11.53

What is Regularisation

Idea: Use a different estimator to estimate the linear regression model. 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.

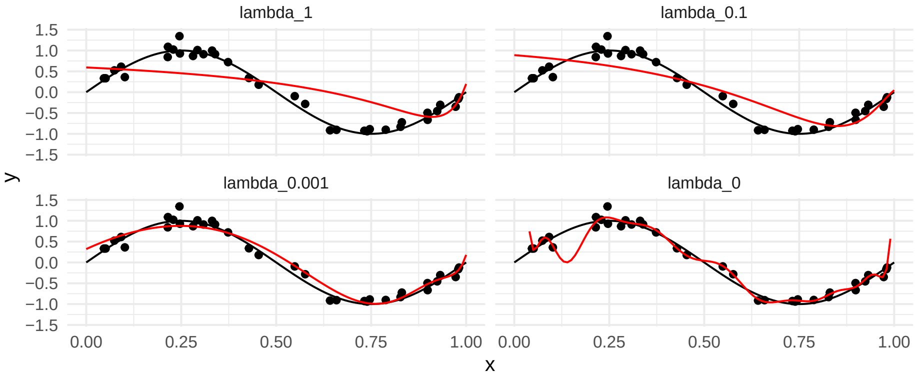
The Ridge

Minimise the following loss function:

$$L(w) = \sum_{i}^{N} (w^T x_i - y_i)^2 + \lambda \sum_{j}^{k} (w_j)^2$$

Luckily, it has a **closed form solution**.

The Ridge


lambda_1
lambda_0.1
lambda_0.001
lambda_0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
y
0.00
0.25
0.50
0.75
1.00
x

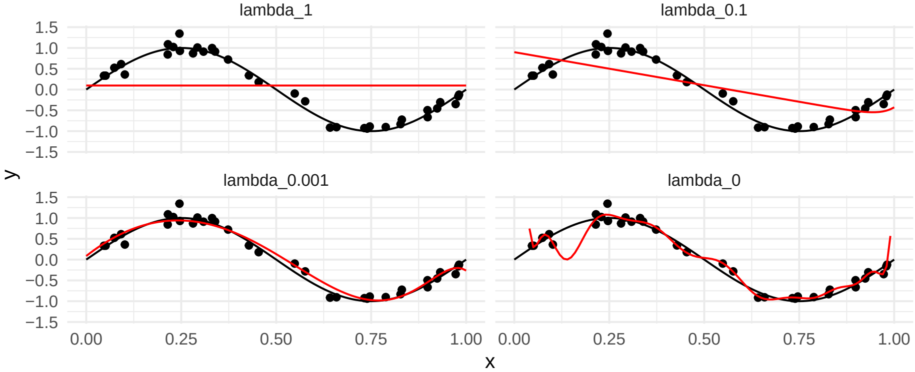
The LASSO

Minimise the following loss function:

$$L(w) = \sum_{i}^{N}(w^T x_i - y_i)^2 + \lambda \sum_{j}^{k}|w_j|$$

Unfortunately, it has **no closed form solution**. One has to use a clever algorithm to find the solution (shooting algorithm).

The LASSO

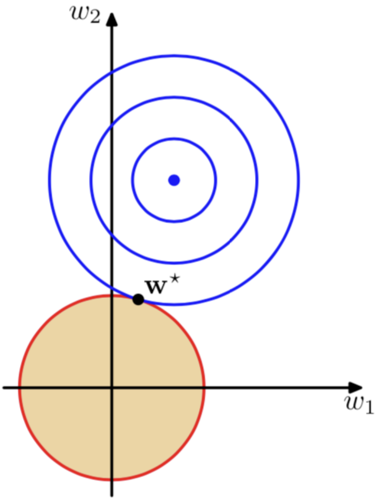

lambda_1
lambda_0.1
lambda_0.001
lambda_0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
y
0.00
0.25
0.50
0.75
1.00
x

The Bias-Variance trade-off

$$MSE(\hat{w}) = E[(\hat{w} - w)^2] = \underbrace{E[(\hat{w} - w)]^2}_{\text{bias}} + \underbrace{E[(\hat{w} - E\hat{w})^2]}_{\text{variance}}$$

- OLS is unbiased estimator
- ridge and LASSO are **biased but have a smaller variance** than least squares
- by optimally choosing λ it is possible to obtain an estimator with smaller MSE

Ridge vs. Lasso

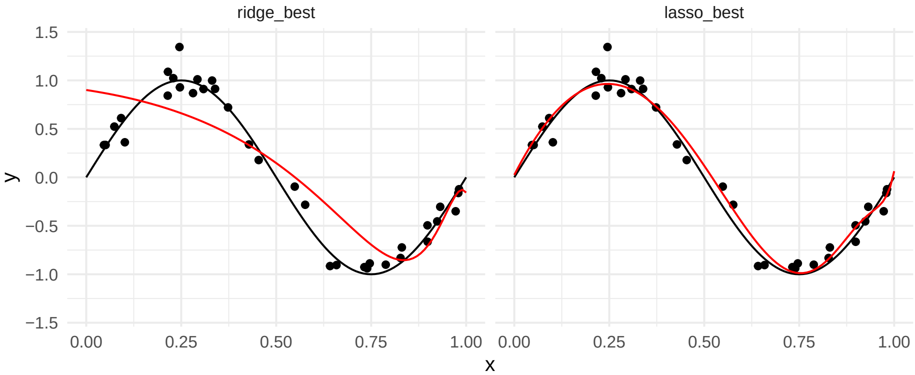


Ridge vs. Lasso

- both are useful when k is large relative to N
- ridge is useful when regressors are highly collinear
- LASSO when true regression parameter vector is sparse and regressors are not highly collinear
- one can use LASSO as variable selection method

How to choose lambda?

Cross-validate!



How to choose lambda?

	train MSE	test MSE	CV MSE
k_0	0.54	0.30	0.46
k_1	0.21	0.22	0.21
k_5	0.01	0.01	0.01
k_30	0.01	1.32	0.45
ridge_best	0.10	0.05	0.13
lasso_best	0.01	0.01	0.02

Section 2

Model Selection Example

The Data

- Articles published from January 7 2013 to January 7 2015 on Mashable: http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity
- **Target:** number of shares in social networks
- **Predictors:** different summary measures of article content (e.g.: links, images, videos, keywords)

Resources

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