Regularisation

Eltecon Data Science Course by Emarsys

Holler Zsuzsa

November 18, 2020

Homework

- Presenters:
  - Bakirov, Aslan - Yatsenko, Anzhelika
  - Both Márton - Kamenár Gyöngyvér
  - Emerson, Ian - Ralbovszki Judit

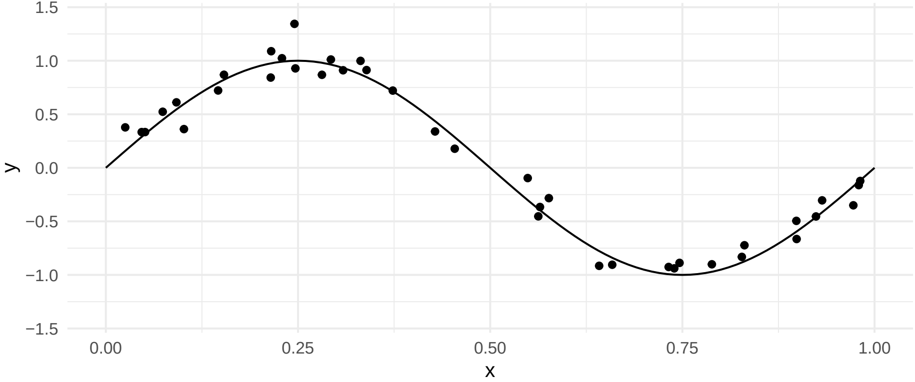
Goal of the lesson

- 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 RMSE	test RMSE	CV RMSE
k_0	0.71	0.54	0.68
k_1	0.45	0.51	0.45
k_5	0.11	0.08	0.11
k_30	0.09	1.49	0.34

Quiz

1. The test error is always larger than the train error.
2. The overfitting error tends to increase with model complexity.
3. The smaller the overfitting error the better the fitted model.
4. If the fitted model parameters vary a lot across different samples of the same data one should check for overfitting.

Share your results in Socrative!

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.

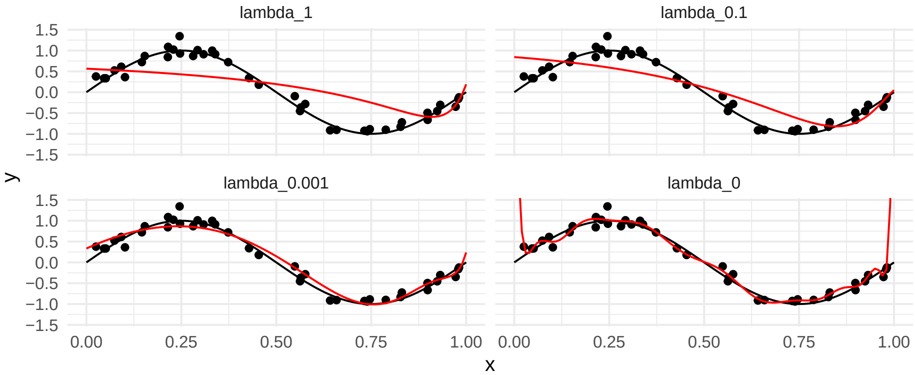
The Ridge

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

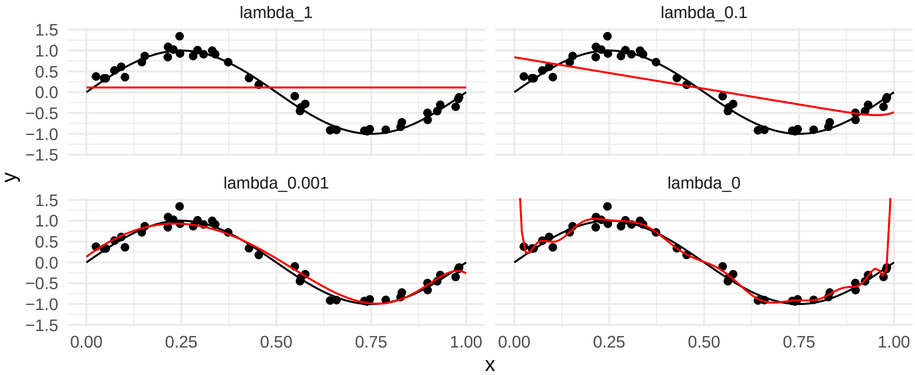
The LASSO

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

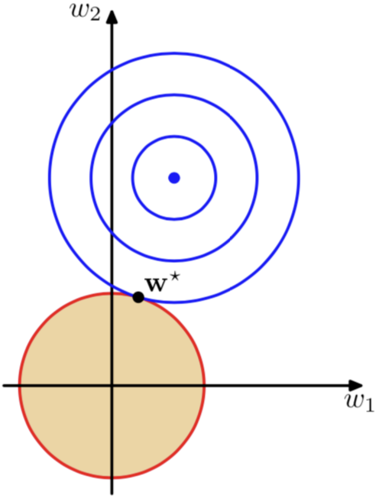

lambda_1
lambda_0.1
lambda_0.001
lambda_0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
y
0.00
0.25
0.50
0.75
1.00
x

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



SMS Spam Prediction

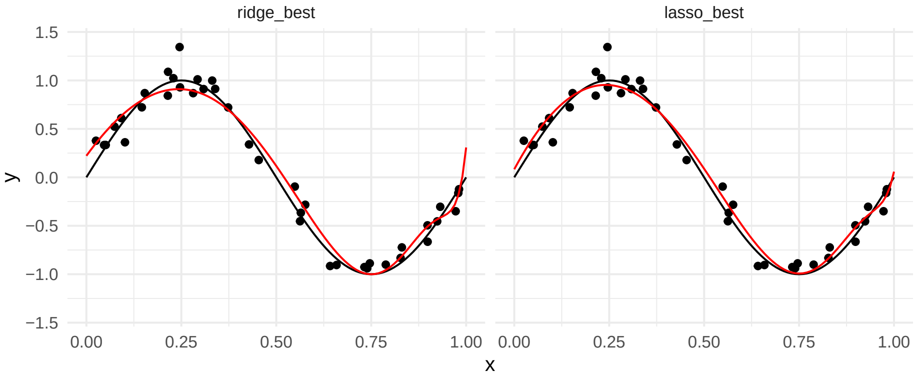
Let's see how it works in practice! **spam_pred_reg.R**

Practice Time

- Task: Implement the same model with Lasso penalty! Use the function documentation!
- Share your results in Socrative!
- You have 10 minutes.

How to choose lambda?

Cross-validate!



How to choose lambda?

	train MSE	test MSE	CV MSE
k_0	0.51	0.30	0.46
k_1	0.20	0.26	0.20
k_5	0.01	0.01	0.01
k_30	0.01	2.23	0.11
ridge_best	0.02	0.02	0.02
lasso_best	0.01	0.01	0.02

SMS Spam Prediction

Let's see how it works in practice! **spam_pred_reg.R**

Practice Time

- Task: Find the optimal lambda value with Lasso penalty!
- Share your results in Socrative!
- You have 10 minutes.

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

Regularisation

Advantages:

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

Disadvantages:

- not clear how to choose λ

Section 2

Model Selection Example

Online News Popularity

- 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)

Homework

- Implement cross-validation to find the optimal lambda parameter for the ridge regression on the Spam prediction example.
- Use any library/function for cross-validation except the one used in the class (you might check the caret package or use your own implementation)
- Choose at least 2 potential lamdba values and report their cross-validated accuracy.
- Presenters:
  - Bat-Erdene, Boldmaa - Kashirin, Andrey
  - Im Seongwon - Kim Yeonggyeong
  - Szőnyi Máté - Tran, Dung

Resources

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