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

Measuring the Quality of Fit

Regression:

- $$RSE = \hat{\sigma} = \sqrt{RSS/(n-2)}$$
- $$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Classification:

- Confusion matrix - accuracy
- ROC curve - AUC

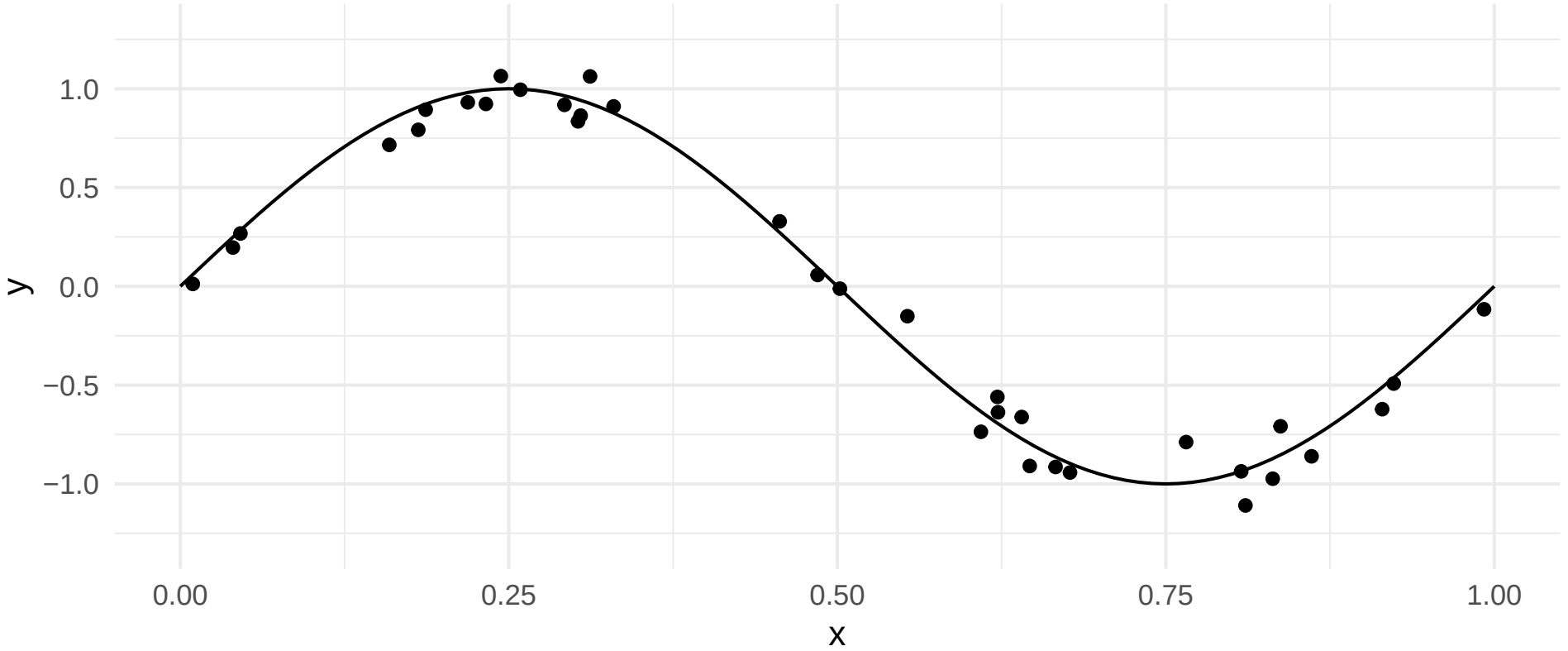
How to Select the Best Model

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

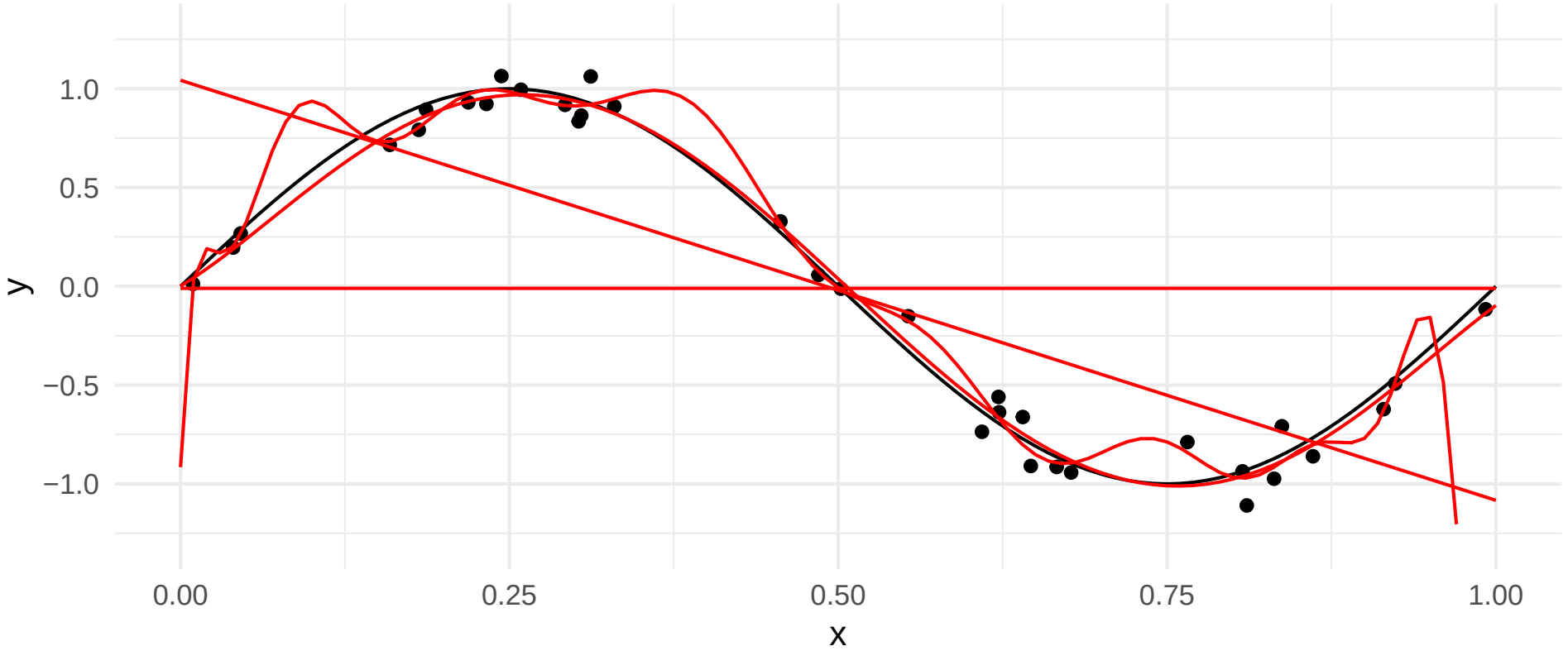
Let's say you fit a linear regression model with k parameters.

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



What is overfitting

Among a set of possible models we choose one with poor generalisation properties.

Why? Because we have a biased 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.

Train vs. Test Error

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

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.

Train vs. Test Error

Advantages:

- Simple approch

Disadvantages:

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

Train vs. Test Error

```
##           train MSE   test MSE
## pred0  0.573647893 0.64993203
## pred1  0.213537141 0.17864762
## pred5  0.007663754 0.02172085
## pred30 0.005243365 0.03760893
```

Information criteria

Idea: Correct for the bias in the estimation of prediction error in complex models by adding a penalty term.

Definition:

- **BIC** (Bayesian approach):

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

- **AIC** (Information theory):

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

where M is the number of parameters, N is the number of data points and $\hat{L}$ is the maxinal 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       AIC       BIC
## pred0  0.573647893 0.64993203  83.87481  86.98551
## pred1  0.213537141 0.17864762  51.28764  55.95368
## pred5  0.007663754 0.02172085 -57.16817 -46.28073
## pred30 0.005243365 0.03760893 -46.45204 -16.90043
```

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

Disadvantages:

- not clear how to choose λ

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

Cross validation

Leave-one-out: **K-fold:**

Advantages:

- Utilizes all the data

Disadvantages:

- computationally expensive