Problemset 3

Zarah Aigner

July 28, 2025

Abstract

The following document contains the answers of problemset 3, the details of the tasks can be found in the book "An Introduction to Statistical Learning". All the code can be found in my Git Repository, in the folder Problemset_3, as well as a description of the structure of the Problemset.

1 Problem

In this exercise we want to estimate the regression coefficients in a linear regression model by minimizing:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s \tag{1}$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

a)

Correct answer: iv. Steadily decrease.
Justification:
So the training RSS cannot increase, due to the fact that each solution for a smaller s also applies on a higher s, although it might not be the optimal solution. This means the optimization problem gets easier, not more difficult.

b)

Correct answer: ii.
Justification:
For a small s, the model is strongly stricted, which could lead to underfitting, however we may obtain a high test error. For a large s, the model has enough flexibility, we obtain a smaller test error. However if s is too large, the model is too flexible, it leads to overfitting, which also leads to a higher test error.
So for the justification: The test error will be large at first, then decreases, till the error reaches a minimum (optimal s), after the minimum the test error starts rising again because the model gets to complex.

c)

Correct answer: iii.
Justification:
If we want to know how the variance behaves, for a small s the model is strictly restricted, which means we have small coefficients, which leads to a stable not very adaptable model and leads to a small variance.
For a large s, the model is really flexible, which means the model reacts strongly on the trainingsdata, which leads to a higher variance.
So to justify the answer: If we have a larger flexibility (larger s), the model adapts better on the training set and reacts more on changes in the data, which means the variance increases.

d)

Correct answer: iv.
Justification:
For a small s, the model cannot really adjust to the data, which leads to a high bias. For a large s, we have more flexibility, which leads to a better model and to a smaller bias.
So the justification is that if we have more degrees of freedom, which means we can approximate the data better, so the bias is shrinking.

e)

Correct answer: v.
Justification:
The irreducible error is independent of the flexibility of the model, the irreducible cannot be reduced, no matter how we choose s.

2 Problem

Suppose we estimate the regression coefficients in a linear regression model by minimizing:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2}$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

a)

Correct answer: iii.
Justification:
For $\lambda = 0$, we do not have a penalty, we have a normal linear regression, which leads to a optimal adaption to the training data, which leads to the smallest training error. So it can be said that for a larger λ, the training RSS is increasing steadily.

b)

Correct answer: ii.
Justification:
For a small λ the model overfits the data, which leads to a high test error due to overfitting. For a medium λ we have a pretty good balance between fitting and regularization, which leads to a low test error. However if we have a large λ, the model will be too simple, which leads to underfitting and to a higher error. This leads to the typical U curve of the test error.

c)

Correct answer: iv.
Justification:
For a small λ, we have a very flexible network, which leads to a high variance. For a large λ, we have a strict model, which leads to a small variance. So it can be said that the variance is steadily decreasing, because the regularization stabilizes the model, which leads to a small variance.

d)

Correct answer: iii.
Justification:
For a small λ we have a complex model, which leads to a small bias, for a large λ, a have a high bias, because we have a lack of flexibility. The model will increase with an increasing λ.

e)

Correct answer: v.
Justification:
The irreducible error is independent on the flexibility of the model, so the irreducible cannot be reduced.

3 Problem

In this exercise, we will predict the number of applications received using the other variables in the College data set.

(a) Split the data into a training set and a test set.

(b) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) Fit a ridge regression model on the training set, with a λ chosen by cross-validation. Report the test error obtained.

(d) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates

The main part of the exercise was implementing the subexercises, therefore I refer to my provided code. For the first exercise I chose a train and test split of 70-30, which means 70 percent of the data was used in the training set and the other 30 percent was used in the test set.
For the following exercises, the tabular 1 below is given, whereas in the tabular all results can be found:

model	**test-MSE**	$\boldsymbol{\lambda}_{\text{best}}$	**# non-zero coefficients**
OLS (Least Squares)	1,931,803.19	–	17
ridge regression	1,907,146.91	0.8111	17
lasso regression	1,928,496.02	3.9456	16

Table 1: Comparision of the test-MSEs and regularization parameters for OLS, ridge and lasso.

4 Problem

It was mentioned in the chapter that a cubic regression spline with one knot ξ can be obtained using a basis of the form $x, x^2, x^3, (x-\xi)^3_+$, where $(x-\xi)^3_+ = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise. We will now show that a function of the form:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+, \tag{3}$$

is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

(a) Find a cubic polynomial:

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3, \tag{4}$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

(b) Find a cubic polynomial:

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3, \tag{5}$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(x)$ is a piecewise polynomial.

(c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continous at ξ.

(e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ.

a)

We start by investigating for $x \leq \xi$. We know that if $x \leq \xi$, we have given $(x-\xi)^3_+ = 0$, therefore (3) reduces to:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{6}$$

We compare the equation above to the following form of the cubic polynomial:

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3. \tag{7}$$

We compare the coefficients:

- $a_1 = \beta_0$
- $b_1 = \beta_1$
- $c_1 = \beta_2$
- $d_1 = \beta_3$

So to summarize it we obtain for all $x \leq \xi$:

$$f(x) = f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{8}$$

So it can be said that the term with $(x-\xi)^3_+$ does not contribute to the equation, therefore $f_1(x)$ reduces to a regular cubic polynomial, whereas the coefficients can be directly compute by using the β values.

b)

We have given that for all $x > \xi$ we have:

$$f_2(x)_{=}\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3, \tag{9}$$

we also have that for $x > \xi$: $(x-\xi)^3_+ = (x-\xi)^3$.
Our goal is now to find a polynomial with the form of (5), whereas $f(x) = f_2(x)$ for all $x > \xi$.

We can rewrite the term $(x-\xi)^3$:

$$(x-\xi)^3 = x^3 - 3\xi x^2 + 3\xi^2 x - \xi^3 \tag{10}$$

We multiplicate with β_4:

$$\beta_4(x-\xi)^3 = \beta_4 x^3 - 3\beta_4 \xi x^2 + 3\beta_4 \xi^2 x - \beta_4 \xi^3. \tag{11}$$

We insert the equation above in (3):

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^3 - 3\beta_4 \xi x^2 + 3\beta_4 \xi^2 x - \beta_4 \xi^3. \tag{12}$$

We compare the terms:

- $a_2 = \beta_0 - \beta_4 \xi^3$
- $b_2 = \beta_1 + 3\beta_4 \xi^2$
- $c_2 = \beta_2 - 3\beta_4 \xi$
- $d_2 = \beta_3 + \beta_4$

So as a result we obtain for $x > \xi$:

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3 \tag{13}$$

with:

- $a_2 = \beta_0 - \beta_4 \xi^3$
- $b_2 = \beta_1 + 3\beta_4 \xi^2$
- $c_2 = \beta_2 - 3\beta_4 \xi$
- $d_2 = \beta_3 + \beta_4$

So it can be said that the function $f(x)$ is a stepwise defined cubic polynomial, which means the function is a regression spline with a knot at ξ.

c)

We should investigate if the function $f(x)$ is continuous at $x = \xi$, therefore we need to check if it satisfies the following condition:

$$f_1(\xi) = f_2(\xi) \tag{14}$$

To show that the function satisfies the condition we use the following functions (8) and (9). We calculate both functions for $x = \xi$:

$$f_1(\xi) = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3 \tag{15}$$

$$f_2(\xi) = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3 + \beta_4 \cdot 0 \tag{16}$$

Now it can be easily seen that the functions fulfill the condition:

$$f_1(\xi) = f_2(\xi) \tag{17}$$

and it can be said that the function $f(x)$ is continuous at $x = \xi$.

d)

To show that the first derivative of the function $f(x)$ is also continuous at $x = \xi$, we start by computing the derivatives:

$$f_1'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 \tag{18}$$

$$f_2'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + 3\beta_4 (x - \xi)^2 \tag{19}$$

here we just used the normal rules for computing the derivative. We know set $x = \xi$:

$$f_1'(\xi) = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 \tag{20}$$

$$f_2'(\xi) = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 + 3\beta_4 (\xi - \xi)^2 = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 \tag{21}$$

Again it can be easily seen:

$$f_1'(\xi) = f_2'(\xi) \tag{22}$$

e)

For proving that the second derivative of the function is continuous, we do mostly the same as in the exercise before. We start by computing the second derivative, which means we take the derivative of the first derivative of the function (18) and (19):

$$f_1''(x) = 2\beta_2 + 6\beta_3 x \tag{23}$$

$$f_2''(x) = 2\beta_2 + 6\beta_3 x + 6\beta_4 (x - \xi) \tag{24}$$

We set $x = \xi$:

$$f_1''(\xi) = 2\beta_2 + 6\beta_3 \xi \tag{25}$$

$$f_2''(\xi) = 2\beta_2 + 6\beta_3 \xi + 6\beta_4 (\xi - \xi) = 2\beta_2 + 6\beta_3 \xi \tag{26}$$

Again it can easily be seen that the function fulfills the following condition:

$$f_1''(\xi) = f_2''(\xi) \tag{27}$$

5 Problem

Fit some of the non-linear models investigated in this chapter to the Auto data set. Is there evidence for non-linear relationships in this data set? Create some informative plots to justify your answers. Find at least one non-linear estimate which does better than linear regression, and justify this using a t-test or by showing an improvement in the cross-validation error with respect to a linear model. You must also produce a plot of the predictor X vs. the non-linear estimate $\hat{f}(X)$.

The main part of the exercise was writing the code Problem_5.py, for a programming language Python was used, the dataset was downloaded from GitHub and can be found in the file Data_Problem_5.csv.
We tried the following models:

- Linear regression
- Quadratic regression (polynomial degree 2)
- Cubic regression (polynomial degree 3)

Then we produced the plot with the 3 different models, the plot can be found below in figure 1. We also produced

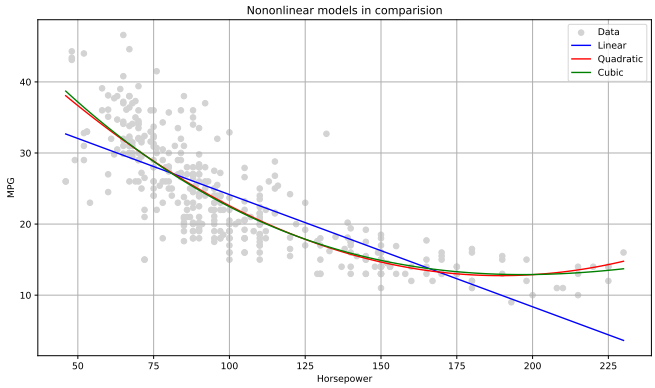

Figure 1: Comparision of the linear, quadratic and cubic regression models, between the horsepower and MPG.

the results of the cross validation, the results can be found in the table below.

Model	**10-fold CV-RMSE**
Linear	27.44
Quadratisch	21.24
Kubisch	21.34

Table 2: Comparision of the models with RMSE (Root Mean Squared Error) by 10 fold cross-validation.

Due to the table it can be seen that the quadratic and cubic model produce a lower error, therefore we can assume a nonlinear relationship between the horsepower and MPG.

6 Problem

The problem involves hyperplanes in two dimensions.

(a) Sketch the hyperplane $1 + 3X_1 - X_2 = 0$. Indicate the set of points for which $1 + 3X_1 - X_2 > 0$, as well as the set of points for which $1 + 3X_1 - X_2 < 0$.

(b) On the same plot, sketch the hyperplane $-2 + X_1 + 2X_2 = 0$. Indicate the set of points for which $-2 + X_1 + 2X_2 > 0$, as well as the set of points for which $-2 + X_1 + 2X_2 < 0$.

a)

We start by rewriting the equation:

$$1 + 3X_1 - X_2 = 0 \Leftrightarrow X_2 = 3X_1 + 1 \tag{28}$$

This is just the function of a linear equation. We can see the result in figure 2, it can also be seen that for the purple part $1 + 3X_1 - X_2 > 0$ and for the blue part $1 + 3X_1 - X_2 < 0$.

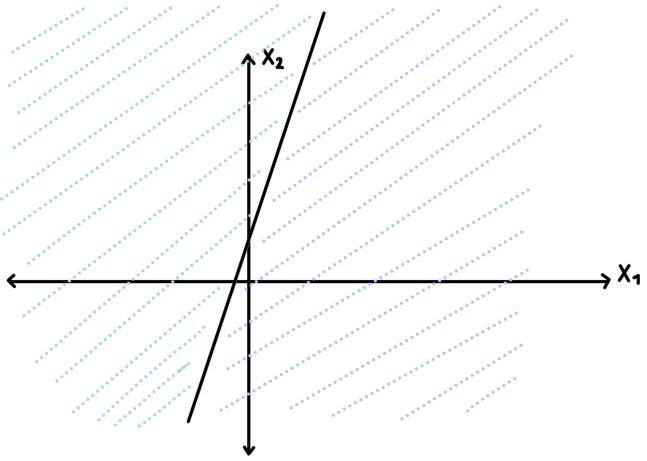

Figure 2: Sketch of the hyperplane. For the purple part $1 + 3X_1 - X_2 > 0$ and for the blue part $1 + 3X_1 - X_2 < 0$.

b)

We now need to insert a second hyperplane in the sketch. We have the given equation for the hyperplane:

$$-2 + X_1 + 2X_2 = 0 \Leftrightarrow X_2 = 1 - \frac{X_1}{2} \tag{29}$$

We did the same procedure as in exercise (a), the obtained sketch can be seen in the figure

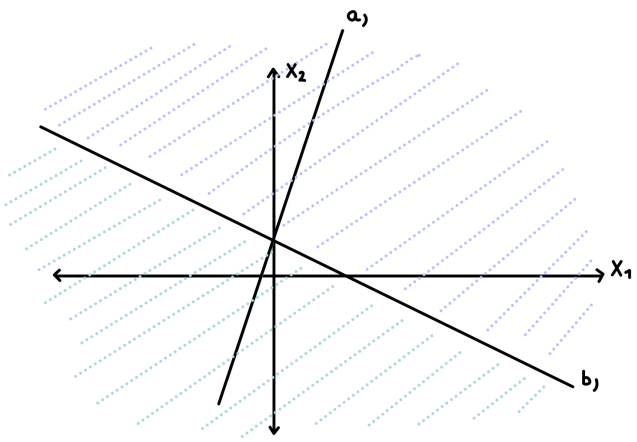

Figure 3: Sketch of the hyperplane. For the line (a), we used the equation $1 + 3X_1 - X_2 = 0$. For the line (b) we used the equation $-2 + X_1 + 2X_2 = 0$, it can be seen that for the blue part we have $-2 + X_1 + 2X_2 < 0$, for the purple part we have $-2 + X_1 + 2X_2 > 0$.

7 Problem

This problem involves the 0J data set which is part of the ISLR2 package.

(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

(b) Fit a support vector classifier to the training data using cost = 0.01, with Purchase as the response and the other variables as predictors. Use the summary() function to produce summary statistics, and describe the result obtained.

(c) What are the training and test error rates?

(d) Use the tune() function to select an optimal cost. Consider values in the range 0.01 to 10.

(e) Compute the training and test error rates using this new value for cost.

(f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for gamma.

(g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel. Set degree = 2.

(h) Overall, which approach seems to give the best results on this data?

The main part of the exercises consisted of implementing the code, therefore the programming language Python was used. The main file of the code is Problem_7.py, also the data can be found in the file Data_Problem_7.csv. For the first exercise we needed to create a training set of 800 random observations, the rest was used for the test set. For the rest of the exercises I refer to the following table, where each result of the exercises is listed.

In this exercise we have investigated different SVM-models, with a linear, a radial and a polynomial kernel. We

Model	**Cost (C)**	**Train Error**	**Test Error**
Linear SVM (initial)	0.01	0.251	0.252
Linear SVM (best)	**1**	**0.172**	**0.144**
RBF SVM (initial)	0.01	0.390	0.389
RBF SVM (best)	**0.01**	**0.390**	**0.389**
Polynomial SVM (initial)	0.01	0.390	0.389
Polynomial SVM (best)	**0.01**	**0.390**	**0.389**

Table 3: Comparision of the errors of divers SVM-models on the OJ dataset. The best model is marked in a bold type.

used all three on the OJ dataset and compared the test errors.
The smallest test error (14.4%), we obtained with the support vector classifier with a linear kernel, whereas we used the cost parameter $C = 1$.
In comparision to the linear model, with the polynomial and the radial kernel we obtained worse results.
So overall it can be said, that for this specific dataset a linear model fits the data the best, which may be due to the fact that the classes can be seperated linearly.

8 Problem (Bonus)

In class, we reviewed that the variance from bagging is:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{30}$$

Derive this formula. Furthermore, this appears to fail if ρ is negative; diagnose the problem in this case.

To solve the problem, we start by investigating equation (30), we know that σ^2 is the variance of the individual model, and ρ is the correlation between two base learners also B is the number of bootstrapped models. We now let $f_1(x), f_2(x), ..., f_B(x)$ be the B base models, we can define:

$$\bar{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{31}$$

We need to take the variance:

$$\text{Var}(\bar{f}(x)) = \text{Var}\left(\frac{1}{B}\sum_{b=1}^{B} f_b(x)\right) \tag{32}$$

We use the formula for the variance of the mean of correlated random variables:

$$\text{Var}(\bar{f}) = \frac{1}{B^2}\left(\sum_{b=1}^{B}\text{Var}(f_b) + \sum_{i\neq j}\text{Cov}(f_i, f_j)\right) \tag{33}$$

If we now assume that each $f_b(x)$ has the same variance as σ^2 and each pair (f_i, f_j) has the same covariance $\text{Cov}(f_i, f_j) = \rho\sigma^2$. We can express (33):

$$\text{Var}(\bar{f}) = \frac{1}{B^2}(B\sigma^2 + B(B-1)\rho\sigma^2) = \frac{\sigma^2}{B^2}(B + \rho B(B-1)) \tag{34}$$

Simplifying:

$$\text{Var}(\bar{f}) = \sigma^2\left(\frac{1}{B} + \rho\frac{B-1}{B}\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{35}$$