Saarland University

The Elements of Stastical Learning

Assignement 1

Due Date: 08.11.2017

Thibault Schowing
Sarah Mcleod
November 7, 2017

Problem 1

Statistical learning is the process of learning from data. The majority of statistical learning can be divided into two categories: **Supervised Learning** and **Unsupervised learning**. In Supervised Learning, you have **inputs** as well as the corresponding **outcomes** that serve to direct the learning process. The goal is to estimate some unknown function f, which serves as the information about the relationship between inputs and outputs. In Unsupervised Learning we observe only the predictors; the responses are not available. The goal of unsupervised learning is to discover something about the data, for example how it groups together. For this reason it is often referred to as Clustering.

Supervised Learning is used for **inference** and **prediction**. In Prediction the goal is to estimate some unknown function f so as to accurately predict some output given a new input. This can be done for **quantitative** values, which is known as **regression** (on continuous values), or for **qualitative** (or categorical) values, which is referred to as **classification**.

When talking about supervised learning it's also important to discuss **training data** and **test data**. The training data is used, as the name suggests, to train a statistical model. Test data are data that have not been given to the model during training and are used to test the performance of that model.

The methods used to estimate f can be categorized as **parametric** or **non-parametric**. Parametric methods make the assumption that the unknown function is linear, and use linear methods to estimate it. Non-parametric methods make no assumptions about the underlying form of f, and thus can use methods that are much more complicated and have many more degrees of freedom.

f -0.75

explanation of inference -0.75

8.5

Problem 2

Show that:

$$E(Y) = argmin_c E[(Y-c)^2]$$

We are looking to proof that the value of c for which $E[(Y-c)^2]$ attains its minimum is $E(Y)$.

To show $E(Y) = argmin_c E[(Y-c)^2]$:

- $E[(Y-c)^2] = E[(Y-c)(Y-c)]$
- $= E[y^2 - 2Yc + c^2]$ -0.5
- $= E[y^2] - 2cE[y] + c^2$ ✓
- to minimize solve for where the gradient is zero: $\frac{d}{dc}E[(Y-c)^2] = 0 = 2E[y] + 2c$ -0.5
- the value of c where the gradient is zero is therefore: $c = E[y]$ ✓
- (optional)to verify this is a minimum take $\frac{d^2}{dc^2}$. The result is 2, which is greater than zero, therefore this value is a minimum. ✓

7

Problem 3

Prove the bias-variance trade-off with irreducible error:

$$E[(y_0 - \hat{f}(x_0) - E(\hat{f}(x_0)))^2] + [E(\hat{f}(x_0) - f(x_0)]^2 + Var(\epsilon) =$$

$$Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))] + Var(\epsilon)$$

First, we expand $E[(y_0 - \hat{f}(x_0))^2]$ into $E[y_0^2 + \hat{f}(x_0)^2 - 2y_0\hat{f}(x_0)]$ ✓

Next, assuming that y_0 is deterministic, we can simplify to $E[\hat{f}(x_0)^2] + y_0^2 - 2y_0E[\hat{f}(x_0)]$

f is deterministic and $y = f + \varepsilon$ -2

With the rules of expectation we get $E[\hat{f}(x_0)^2] - 2y_0E[\hat{f}(x_0) + y_0^2$

Adding and Subtracting $E[\hat{f}(x_0)]^2$ we get

$$E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2 + E[\hat{f}(x_0)]^2 - 2y_0E[\hat{f}(x_0)] + y_0^2$$

Next, $E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2 = Var(\hat{f}(x_0))$, so we get

$$Var(\hat{f}(x_0)) + E[\hat{f}(x_0)]^2 - 2y_0E[\hat{f}(x_0)] + y_0^2$$

Then we simplify to $Var(\hat{f}(x_0)) + [E(\hat{f}(x_0) - y_0)]^2$

Then simplify again to $Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$ how? -2

8

$E = [(y - \hat{f})^2]$

assume $E(\varepsilon) = 0$

$E(\varepsilon) = 0$

$Var(\varepsilon) = E(\varepsilon^2) - E(\varepsilon)^2 \Rightarrow E(\varepsilon^2) = Var(\varepsilon)$

$E[\varepsilon(f - \hat{f})] = E(\varepsilon)E(f - \hat{f}) = 0$

independent statement

$E[(y - \hat{f})^2] = E[(f - \hat{f})^2] + Var(\varepsilon)$

Problem 4

b

Data repartition Train and testset.

	Observations
Testset	31
Trainset	80
Total	111

2 ✓

Data types There is four columns, in this work "ozone" is considered to be the output.

Data	Data Type
ozone	Numerical
radiation	Integer
temperature	Integer
wind	Numerical

✓

c

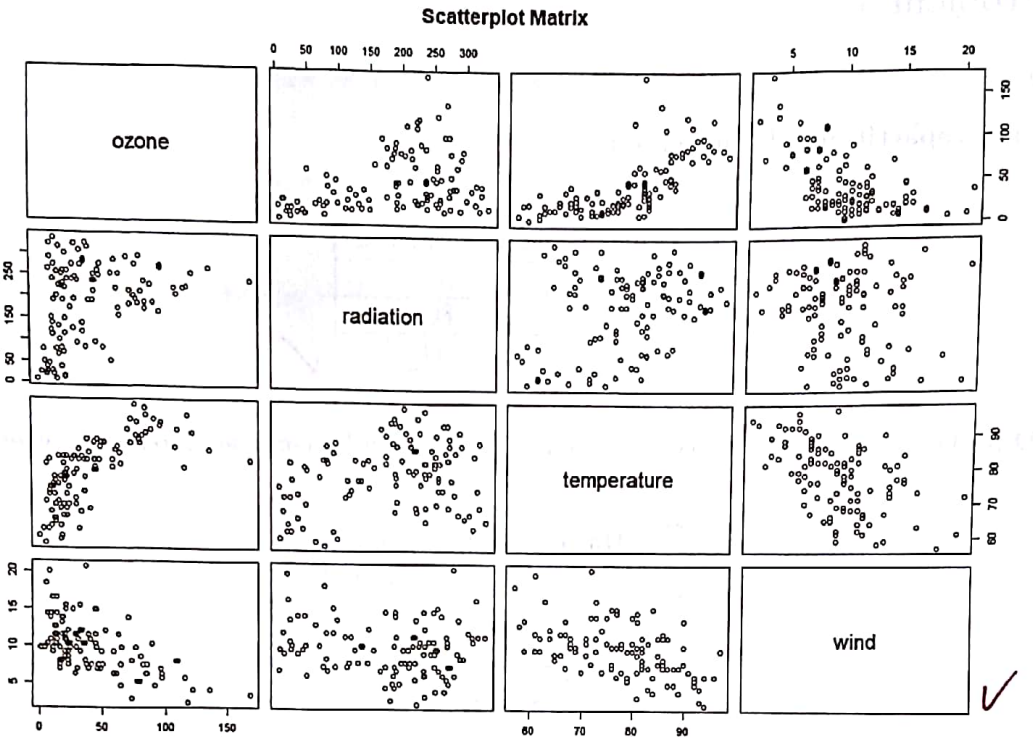
	Range	Mean	SD
Ozone	1:168	42.099099	33.275969
Radiation	7:334	184.801802	91.152302
Temperature	57:97	77.792793	9.529969
Wind	2.3:20.7	9.938739	3.559218

2 ✓

d

The range of the Pearson correlation coefficient is between -1 and 1. ✓ A coefficient of 1 means that the variables have a total positive linear correlation, a 0 means that there is no linear correlation and a -1 means that there is a total negative linear correlation. ✓ One can see in table ?? that each variable has a total positive linear correlation with itself (obviously).

When we look at the graphs below and the Pearson correlations, we can see a relation between wind and ozone, that gives the impression of a negative sloped line and between temperature and ozone but positively this time. The coefficients confirm what we see on the graph. The two related pairs present high coefficients. We have -0.61 for the wind-ozone pair and 0.69 for the ozone-temperature pair. ✓



3 ✓

Figure 1: Scatter-plots between each variables.

	ozone	radiation	temperature	wind
ozone	1.0000000	0.3483417	0.6985414	-0.6129508
radiation	0.3483417	1.0000000	0.2940876	-0.1273656
temperature	0.6985414	0.2940876	1.0000000	-0.4971459
wind	-0.6129508	-0.1273656	-0.4971459	1.0000000 ✓

Table 1: Pearson Correlation between the features

1 e)✓

f

Ozone prediction using a linear regression model.

color coding is wrong -0.5

3.5

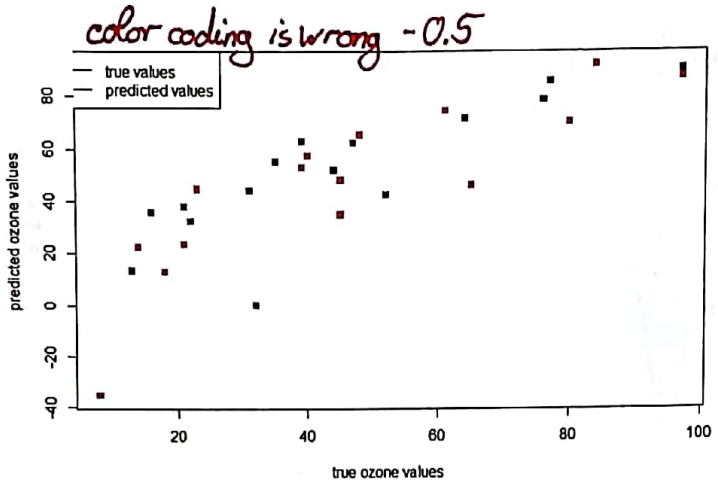
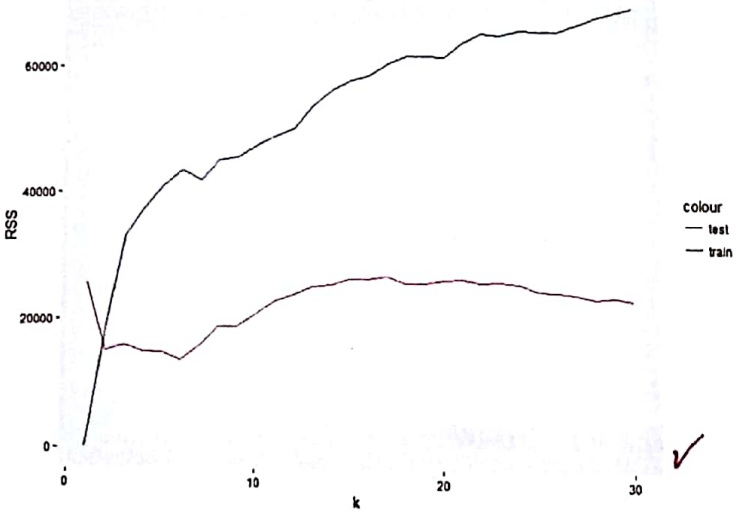


Figure 2: Predicted and true values of ozone for the test set

RSS: 8208.509 ✓

Pearson Correlation between predictions and true responses: 0.8268958. ✓

g



✓

Figure 3: RSS for training and test set and for each k.

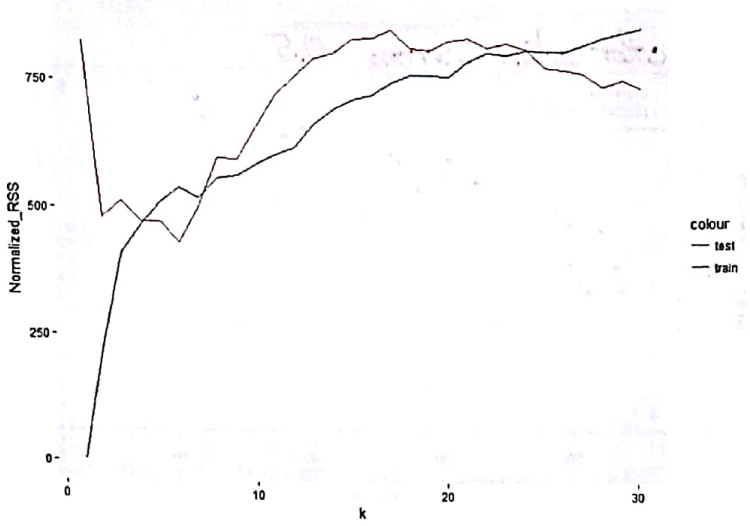


Figure 4: Normalized RSS for training and test sets.

4 left -0.5

The more complex models will be on the right hand side of the graph. This is due to the fact that the model will fit all the tiny changes in the data, each time a new point is added. Consequently it will lead to a high variance and a low bias. For this data, we would choose a k value around 6, minimising the RSS, as we can see on the graphs (normalized and not). ↑ on the test set -0.5

KNN doesn't make any assumption on the underlying data distribution. It is a non-parametric learning algorithm. ✓

0 ↑k -3

15.5 / 3.9

Knn → locally constant function.

Liner better → complexity lower

model assumptions → in favor of KNN, but is linear relationship, so LR better.