Assignment 2 - Regression and classification

Nathan HAUDOT, Hugo MATH

November 16, 2021

1 Regression for villas in Landvetter

1.a Find a linear regression model

From the dataset, we cleaned two data point which we find were not relevant for the linear regression (maybe not actual houses). We applied this filter to the panda dataframe :

```
df_clean = df.drop(df[(df.Living_area > 160) & (df.Selling_price < 3000000)].index)
```

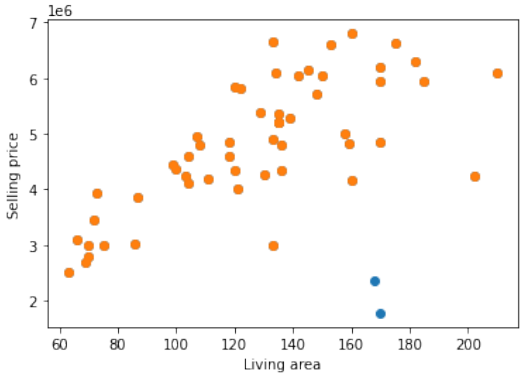

Figure 1: Living area vs. selling price

1.b What are the values of the slope and intercept of the regression line?

We can clearly see a correlation between the living area and the selling price. While using a linear regression model using **sklearn**, we get a slope of **1 809 821** and an intercept of **23 597**.

1.c Predict the selling price of 100m2, 150m2, and 200m2

For a prediction of 100, 150 and 200 meter squared living areas with our model, we respectively get : **4 169 601**, **5 349 490**, and **6 529 380** SEK

1.d Residual Plot

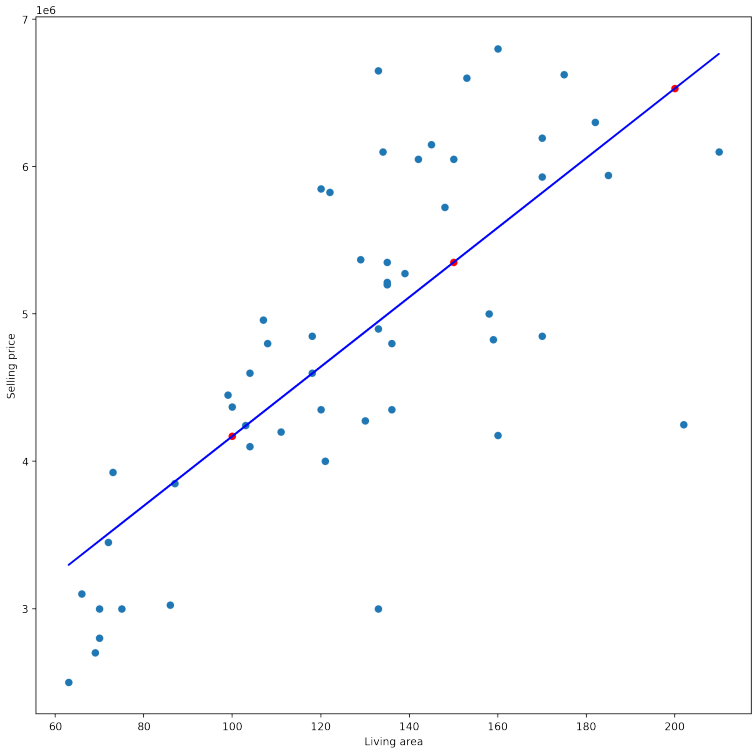

Figure 2: Linear regression model (blue) & predictions markers (red)

2 Iris dataset

2.a Confusion matrix

If we apply logistic regression to the iris dataset and try to predict 3 classes, we thus get a 3x3 confusion matrix :

$$C_m = \begin{bmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 1 & 49 \end{bmatrix}$$

The resulting matrix means that our model classify correctly Setosa, but detects 3 Versicolour as 3 Virginica, and 1 Virginica as 1 Versicolour.

2.b K-nearest neightbours

We seek the perfect parameter for k and weights using heuristic methods. Therefore we can test several parameters on the iris dataset. We also know that we have 3 classes so we are not faced to tied votes. Increasing K would mean reducing the noise of the classification, and a better generalization.

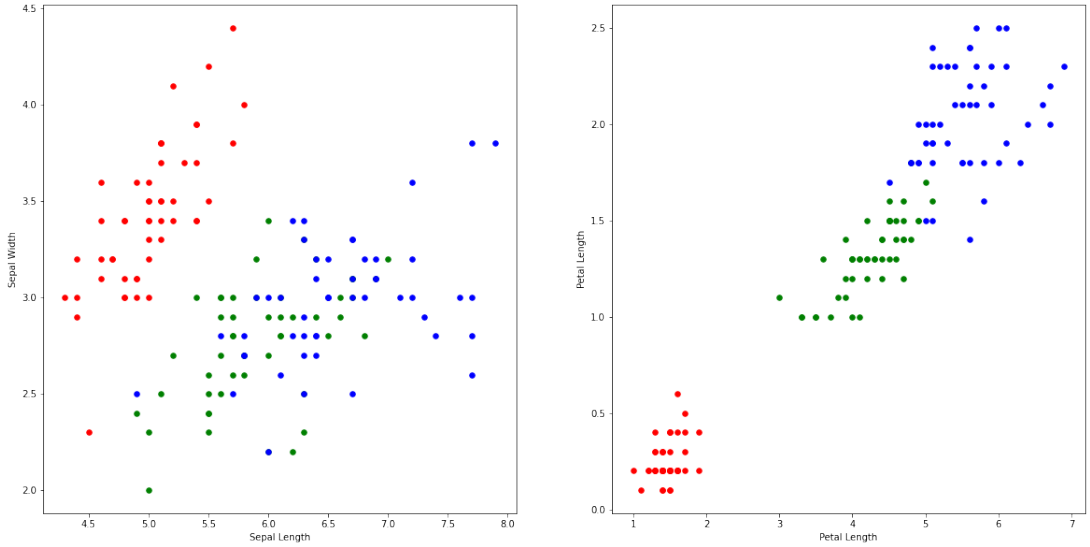

Figure 3: Iris dataset with Setosa(red), Virginica(blue) Versicolor(green)

We can discuss about uniform and distance weights. As we see in the two plots, the distance weight seems more relevant. In fact we can clearly see separable clusters. Let's run several confusion matrix to see if our assumptions are corrects. After running our experiments, we get these results :

K	Weight	Score
1	uniform	1.0
1	distance	1.0
3	uniform	0.96
3	distance	1.0
5	uniform	0.966
5	distance	1.0
8	uniform	0.98
8	distance	1.0
10	uniform	0.98
10	distance	1.0
15	uniform	0.986
15	distance	1.0
20	uniform	0.98
20	distance	1.0
25	uniform	0.98
25	distance	1.0
30	uniform	0.953
30	distance	1.0

We thus choose the highest K to generalize and the 'distance' weight which seems more relevant.

2.c Compare the classification models for the iris data set that are generated by k-nearest neighbours

We can use a confusion matrix between the both k-nearest neighbours & logistic regression models :

$$C_m log = \begin{bmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 1 & 49 \end{bmatrix}$$

$$C_m k = \begin{bmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{bmatrix}$$

Obviously, we can see that our k-nearest model performs better on this dataset (score of 1.0). However, it does not necessarily mean that our model performs better overall. In fact, we do not have an other iris dataset to test the two models on it. If we run theses two models directly in the "field", we might have some problems regarding the generalization and the overall scores. That is why we should have split the dataset into two separate datasets, one for the train and one for the test.

3 Why it is important to use a separate test (and sometimes validation) set?

The purpose of using two different datasets (for training and testing) is to ensure model performance and generalization, giving an unbiased estimate of the final model skill. The testing dataset can also be used to tune the trained model when comparing the two datasets (which then requires the use of a validation dataset).