Project 2 – Regression and Classification

02450: Introduction to Machine Learning Course project

Regression Part A and B: Omer Yilmaz s196596@student.dtu.dk

Classification: Sri Vasudha Hemadri Bhotla s196368@student.dtu.dk

Dataset: London Bike Share Data

Regression, part a:

1.We are predicting the cnt variable which is the number of the bikes rented in an hour in London city based on “timestamp”, “t1”, “t2”, “hum”, “wind_speed”, “wheather_code”, “is_holiday”, “is_weekend”, “season” variables.

As mentioned in “Data Manipulation” part of Project 1:

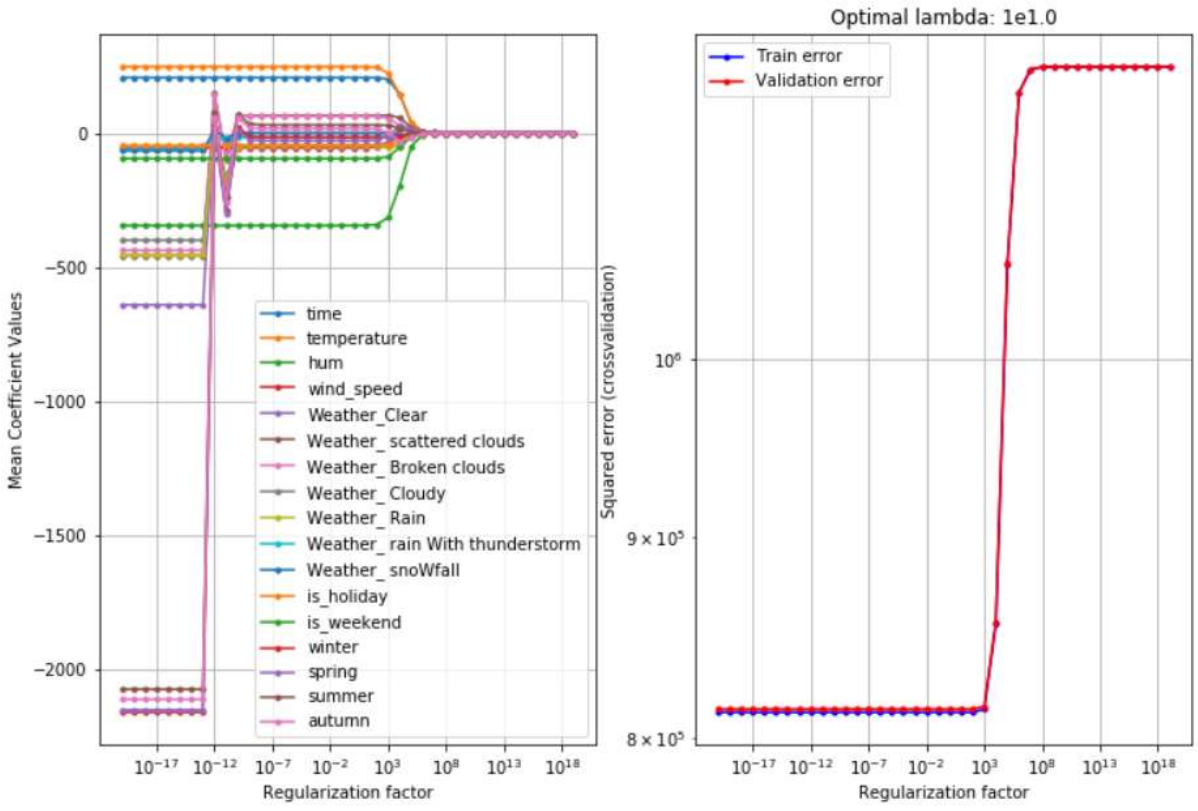
“Firstly, the format of attribute timestamp is like ‘4/01/2015 0:00’ combining the date and time together. For analyzing the influence of different dates and time, here we convert attribute time stamp into two separate columns, namely, attribute dates and attribute time.

Secondly, since the attribute t1 and t2 all indicate the temperature, we use a single representative feature temperature by taking the mean of t1 and t2.

Thirdly, the attribute weather_code contains 8 different weather conditions, we utilise one-out-of-k encoding to deal with it. In fact, there is no weather_code =94 in the dataset. In addition, the attribute season is a categorical variable with values from 0 to 3 to indicate spring, summer, autumn and winter, so one-out-of-K encoding also applied in this variable.”

We changed X such that it has 17 attributes and each column has mean 0 and standard deviation 1.

2.Generalization error as a function of λ

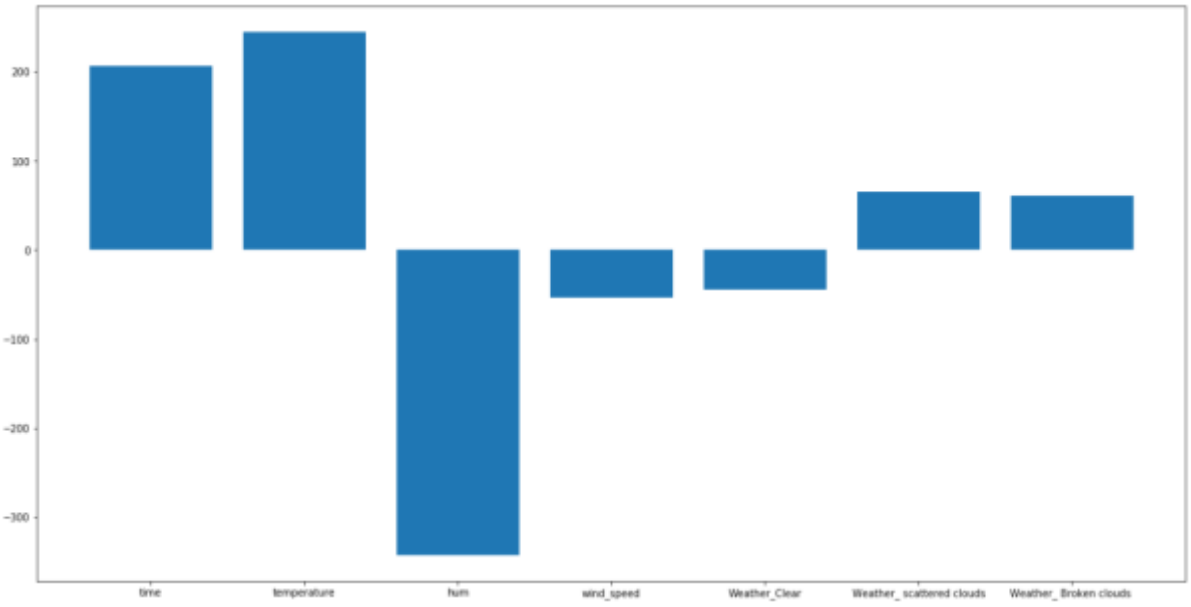


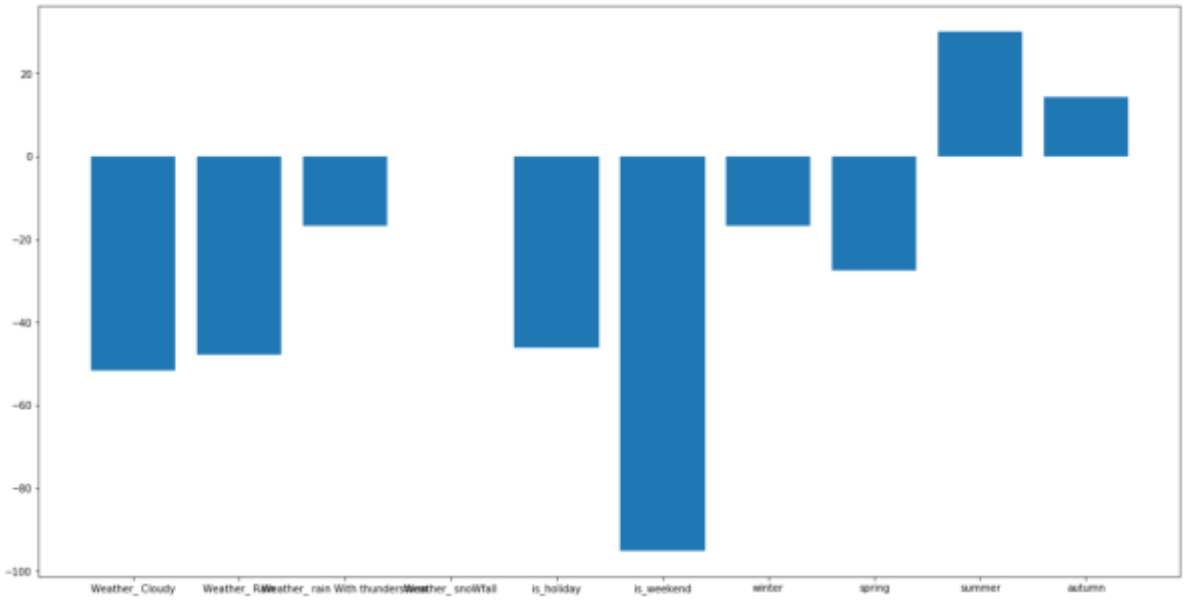
In this part we adjusted the code in Exercise 8.1.1 for our dataset. We set the lambda powers of 10 in range -20 and 20. As you can see in the figures, there is a sharp increase in the error when lambda is bigger than 10^3 because of underfitting. Also, this forces variables to be 0.

When we check the figure with Coefficient values, there is sharp change when the lambda is 10^-12 which is interesting.

Lastly, we found that best lambda value is 10.

3.Effects of selected attributes





In the graphs you can see the coefficients for lambda equal to 10. Time and temperature positively affect the count of the rented bikes. Also, we can see how the weather condition and seasons affect the count. Weekends and holidays cause huge decrease in the count.

Interestingly clear weather decreases the count. Maybe the reason is that linear regression is not predicting perfectly. Also, spring causes decrease in the count. The reason might be the different weather conditions in London.

Regression, part B

1. Models for regression

Since our dataset contains 17414 observations and 17 attributes, we decided to take K1 and K2 equal to 5. We used the same range for lambda for linear regression. For number of hidden layers, we used numbers "1, 4, 16, 64" which are powers of 4. We made max_iter variable 10000 to have some good results and transfer function ReLU to make it faster.

Let's take a deep look in our code! We copied X and Y for the ANN. After that we created some lists for errors and r1, r2, r3 for the statistical evaluation. Then, the cross validation fors!

First, we trained the linear regression model and baseline model as in the part A with K2 equal to 5. And stored the best lambda and errors in the arrays we created before.

Secondly, we split the training set for the second level of the cross-validation for ANN. Then we trained our ANN for different hidden units and choose the best number of hidden layers.

Thirdly, we calculated our error with the test set for the model we choose and stored in the array.

Lastly, we calculated Confidential Intervals and p-values for the differences between errors for each model.

2. Two level cross- validation

Outer Fold	**ANN**		**Linear Regression**		**Baseline**
i	h	Error_test	lambda	Error_test	Error_test
1	64	550169.	10	771267.93434241	1125754.37558499
2	64	625225.8125	10	832766.58738098	1208513.94104254
3	64	627325.3125	10	847691.85457206	1216590.64416057
4	64	598898.9375	10	784228.95098348	1192921.48432306
5	64	588202.25	10	769953.27484253	1142640.91383833

3.

	p-value	**Confidence Interval**
Regression-baseline	6.1226350614819826e-06	(-409347.4018314921, -342857.70089971623)
Baseline-ANN	4.0015849712286166e-07	(553460.8254058437, 605179.1931739486)
ANN-Regression	5.576917251377137e-05	(-234530.63855770568, -171904.27729087835)

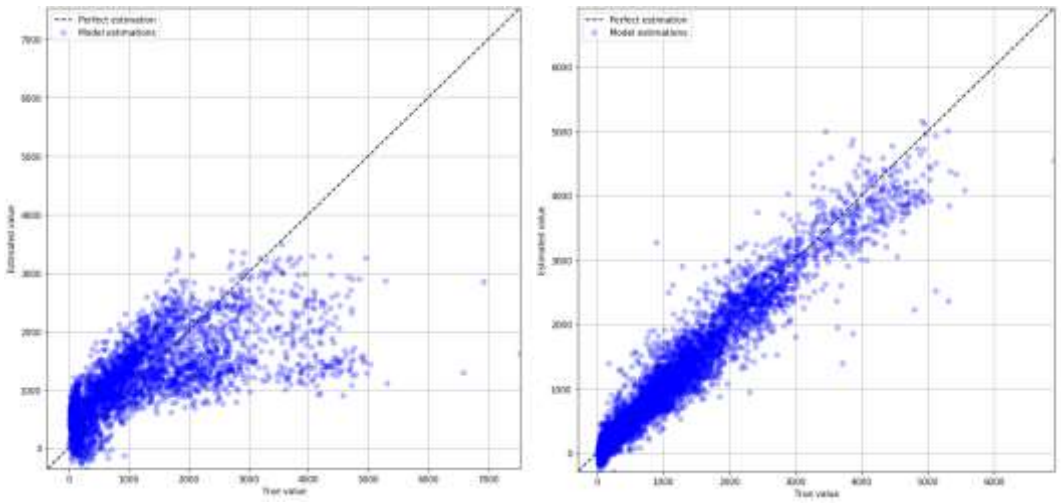
As you can see our p-values are much less than 0.05 which rejects the null hypothesis of any model is similar to each other. Also, as you can see the Confidence Intervals, there is huge error

differences between our models with %95 probability. Therefore, ANN with 64 hidden units is much better than regression which is better than baseline.

PS:

First one is our output when we run exercise 8.2.6 with 64 hidden units and ReLU transfer function.

The second one is the output with 3 hidden layers with 16 units for each and with ReLU transfer function. As you can see more layers gives better results



Classification

The classification problem chosen was to predict which season the bicycle was borrowed from, based on the rest of the attributes(with the time stamp removed). This is a multiclass classification problem.

Baseline:

The baseline model basically returns the season which has the greatest number of cycles borrowed. We use this model to compare the other models with, to test for accuracy of the models. Obviously, we would expect a lot of bikes to be borrowed in the summer.

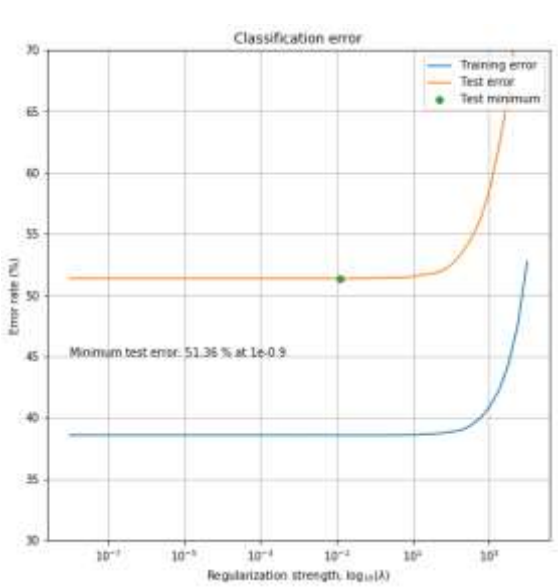
Multinomial Regression:

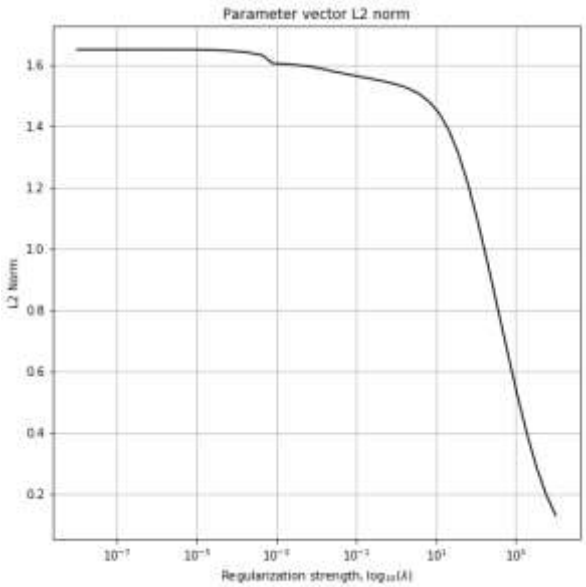
We should expect our model to work better than the baseline model, since the data has been analyzed more deeply.

10-fold cross-validation has been used to classify the data and to find the best value of lambda to be used. The range of values of lambda taken are 10^{-8} to 10^{4}

The trend of the error with change of lambda can be observed in the following figures. The minimum error was obtained when the value of lambda was $10^{-0.9}$

The L2 norm is plotted against lambda, and we can see how it decreases with increase in lambda. This model gave less error than the baseline model.





ANN:

A neural network has been used to predict the class and to see how good the neural network is.

We might get a better model when we keep training, but since my computer gets hanged several times between the process, we are not able to. To use 10-fold cross-validation gets expensive.

The complexity controlling parameter is the number of units which was chosen from the array of [1,4,16,64]. The error value was least when 1 was used.

The number of hidden layers used was not high, and the error seemed to increase with increase in the number of hidden layers. However, for classification usually one layer works well to predict the output and hence we have gone ahead with that.

The error values for ANN model were less than the error from the multinomial regression model.

Two-level cross-validation:

Outer Fold	**ANN**		**Linear Regression**		**Baseline**
	Hidden layers	E^{test}	λ	E^{test}	E^{test}
1	1	0.42980189	1.e-06	0.53714933	0.74734424
2	1	0.36606373	0.00044984	0.54260536	0.7556704
3	1	0.64599483	1.e-06	0.53901665	0.75538329
4	1	0.52081538	0.82864277	0.53750873	0.74849268
5	1	0.61556129	1.e-06	0.5394201	0.74956921

We divided the dataset into 5 splits, because it was taking a lot of time when we used 10-split. We then used the training data to be split further into 5 parts to perform the classification again and select the complexity parameter.

The results show that the error for the ANN model is less than that of the other models. When the dataset is large (17414 entries, each with 15 attributes), we have broken down our analysis to focus on smaller subsets. We felt that the models might be performing similarly on the dataset as a whole, instead of getting a single model which works best. Hence, we need to do the statistical analysis next.

Statistical Evaluation:

	p-value	Confidence Interval
Regression,Baseline	0.41992453146359277	(-0.15366351997157132, 0.07007397244801142)
ANN,Baseline	0.41992453146359277	(-0.07065179740535749, 0.1549306183502039)
ANN,Regression	0.41992453146359277	(-0.0012670983786326052, 0.0005778249573460761)

Setup II was used to get the p values and the confidence intervals.

The value of p is not less than 0.05, and this deems our method statistically not significant. We cannot reject the null hypothesis in this case. The p values lead us to say that there isn't a statistically significant difference between our models.

The confidence intervals on the other hand, imply that Regression and ANN are the most trustworthy. Their CI is much smaller than the other intervals. Therefore, the baseline is not as good as the other models and ANN gives us the least error, and is also very trustworthy.

Discussion:

Regression: This dataset is suitable to do regression. Out of the three models used for regression, We have learnt about which attributes are more important to predict the count of bicycles. ANN is a better model to use than regression in this case.

Classification: We have learnt that the regression model is better than the baseline model since it gives us lesser error value. We also learnt that predicting the season the bike was borrowed in was probably not the best use of this dataset. Regression makes more sense to predict the number of bikes. Also, we cannot expect the machine to learn more than what a human could possibly understand. The relation between the attributes and the season the cycle was borrowed in is not really high. Hence, we couldn't make the best predictions with less error.

Previous Analysis:

This has been used to make predictions before, and it has been used for regression and not classification. As discussed, classification on this dataset doesn't really take advantage of what the dataset has to offer to us.

Reference: https://www.kaggle.com/hmavrodiev/bike-sharing-prediction-rf-xgboost

This is the only proper use of our data in regression. In our project we have implemented 2 level cross validation and compared different models' error by value and statistically. In the previous analyze, they compared different machine learning methods to predict the number of bikes rented. Then they also compared the errors and scores. They also looked at feature importance. Forexample: