TECHNICAL UNIVERSITY OF DENMARK

INTRODUCTION TO MACHINE LEARNING AND DATA MINING, 02450

Project 2

Supervised learning: Classification and regression

Name	Student nr.
Luka Avbreht	**s191963**
Weston Jones	**s191380**
Michael Rupprecht	**s191759**

November 12, 2019

1 Regression: Part A

1.1 Question 1

Our data-set contains information about AirBnB property listings in New York City. We determined that a relevant regression problem would be try and predict the nightly price of listings based off the most useful attributes in our data-set: The borough of each listing, the room type, the minimum stay, number of reviews, reviews per month, number of host listings, and the property availability.

Data Attribute Name	Explanation	Type
id	Listing unique identifier	Nominal
name	Name of listing	Nominal
host_id	Host unique identifier	Nominal
host_name	Name of host	Nominal
neighborhood_group	Listing borough (Manhattan, Brooklyn, etc.)	Nominal
neighborhood	Listing specific neighborhood	Nominal
latitude	Approximated latitude	Interval
longitude	Approximated longitude	Interval
room_type	Room type (Private or shared room)	Nominal
price	Price per night	Ratio
minimum_nights	Minimum number of nights in a stay	Discrete
number_of_reviews	Number of guests who've left reviews	Discrete
last_review	Date of the last review	Interval
reviews_per_month	Average number of reviews per month	Ratio
calculated_host_listings_count	Number of listings the host has	Discrete
availability_365	Number of says in the year listing available	Discrete

For the nominal attributes (room type and borough), we used one-out-of-k encoding to extend out data matrix avoid biasing certain values. Additionally, the numerical attributes were standardized by subtracting each attribute's mean and then dividing by its standard deviation.

Next, we narrowed down the optimal features to use in the regression. Eight iterations of cross validation folding with the "feature_selector_lr" routine from the toolbox were run to determine optimal features by minimizing training and test errors. The results of feature extraction can be seen in figure 1.

Listing location in the Bronx, Brooklyn, Manhattan, whether the listing was an entire home, the minimum stay, reviews per month, and apartment availability were determined to be the seven optimal features.The effects of each feature can be observed in the scatter plots in figure 2.

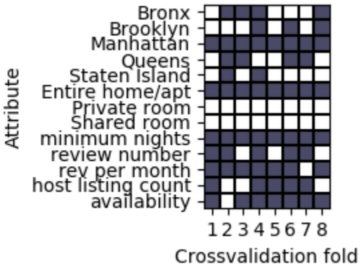

Figure 1: Parameters used in linear regression

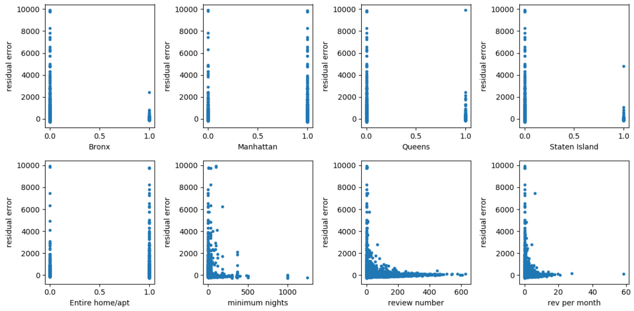

Figure 2: Scatter plot of data points versus their value to be used in regression

1.2 Question 2

Next, we introduced a regularization parameter on the linear regression model using the optimal features described above. The effects of different regularization parameters (λs) on the model can be observed in figure 3. We used a broad logarithmic range of possible λ values as we initially had trouble isolating an optimal choice. From the chart, we see that the accuracy increases at smaller values of λ and that the optimal regularization parameter for our data set is $\lambda = 100$.

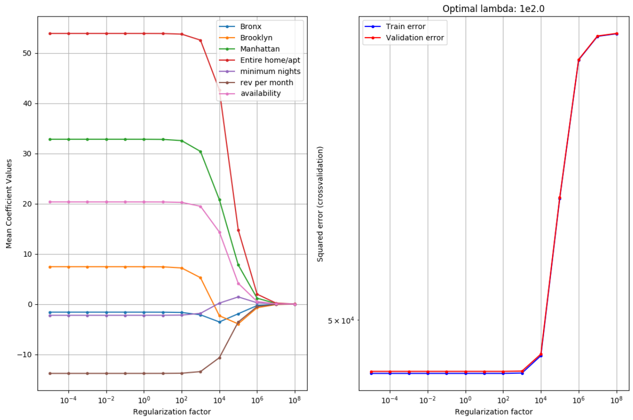

Figure 3: Regularization parameter vs coefficient values and effect of Regularization on test and training errors

1.3 Question 3

Predicting the price of a listing using our model is done with the following linear function.

$$price(x) = w_0 + w_1 * x[0] + w_2 * x[1] + w_3 * x[2] + w_6 * x[5] + w_9 * x[8] + w_{11} * x[10] + w_{13} * x[12]$$

Where x is a vector containing values of parameters for each listing. For example, we take the value of the first column in x (Whether or not the listing is in the Bronx) and multiply it by weight coefficient w_1 (the coefficient corresponding to this feature). This process is repeated for all other features that were selected above.

Weights	
Offset	148.66
Bronx	-3.66
Brooklyn	-4.52
Manhattan	25.68
Entire home	53.92
Minimum nights	-1.96
Reviews per month	-13.98
Availability	20.65

Average total error in dollars across all listings	
Training error:	52391.23286375975
Test error:	52575.74962782959

Rˆ2 Values	
Rˆ2 train:	0.09145253200314354
Rˆ2 test:	0.08634704815299228

The weights assigned to each parameter make sense, but the actual error seems a little off. Whether the property is located in Manhattan and whether the property is an entire apartment / house have the biggest positive contributions to the price, which is logical. Manhattan is the most expensive borough and having an entire property to yourself costs more. Likewise, the reviews per month has a negative contribution to the price, which is also logical as more popular / reviewed properties are probably popular because they are fairly priced.

The actual error amounts seem too high. However, this can be attributed to some of the outlier, luxury apartments in Manhattan that have nightly prices in the thousands of dollars.

2 Regression: Part B

2.1 Question 1

In this section we compared three different regression models designed to accurately predict the prices of the AirBnb listings in our data-set. The three models are as follows:

1. A baseline model that simply uses the mean price for AirBnb listings in New York City as a default.

2. The linear regression model described above. We use the same logarithmic range of lambdas as before.

3. An artificial neural network (ANN) that considers all relevant features and calculates price using a feed forward neural network with one hidden layer and a (Non linear) activation function ReLU. We used PyTorch to implement and train our simple neural network. For the range of h values, we figure extremely high values would result in diminishing returns / over-fitting issues. Thus we used a simple scale starting at 1, increasing by 4, and stopping at 37.

2.2 Question 2

We used KFold to split our data ten times. Each split resulted in a training set and a test set, with training set containing 95% of data points in each fold. For each fold we again calculated the value of the optimal (smallest test error on our test data) regularization parameter for linear regression as well as the optimal number of hidden layers in our ANN. The results for each of 10 folds can be seen in table 2. As before, the error represents the sum of all differences in price between the actual price and the model's prediction across all listings. Also, we don't get the same optimal regularization parameter for the first section.

For the ANN, the error varies a lot, however the more complex neural networks usually preform better.

Comparing the results at a glance, it seems like the ANN performs better than the Linear Regression, which performs better than the baseline. We know that our models are somewhat accurate, as they both perform better than the baseline, and that the underlying relationship between our chosen attributes and the price is most likely non-linear as the neural network outperforms the regression.

Outer fold	ANN		Linear Regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
0	5	80076.139	1	81855.882	87051.661
1	17	57040.390	10	58139.619	60814.024
2	5	33812.731	10	34477.168	37637.058
3	21	46032.876	10	47044.400	51440.011
4	29	25172.991	10	26923.950	31167.697
5	25	51945.718	10	53591.767	58451.096
6	15	52019.941	10	53809.682	59007.374
7	27	37818.446	10	39397.919	46516.920
8	23	66036.048	1	68603.581	75908.643
9	1	60697.019	10	61014.411	69229.674

Table 1: Regression comparison results

2.3 Question 3

	ANN vs. lin. reg.	ANN vs. baseline	lin. reg. vs. baseline
lower bound	597.12	587.62	-405.27799836610774
upper bound	598.41	589.07	-405.27799836610760
p-value	3.4299×10^{-27}	1.0631×10^{-26}	2.7170×10^{-144}

Based on these results coupled with the results from the previous question, we see that the ANN is the best model, the linear regression is the second-best, and the baseline is the worst, and that these results hold to a degree of statistical significance. Thus, it would make sense to use an ANN with this data.

3 Classification

3.1 Question 1

The classification problem we have chosen for our data set involves classifying AirBnB listings as either private rooms, shared rooms, or entire homes. Since there are three possible output classes, this makes our problem a multi-class classification problem. Aside from the obvious and trivial problem of using coordinates to classify neighborhoods, our data-set is not very well suited for classification. There are not very many nominal attributes available to use and those that do exist mostly have to do with location (something that is a direct function of latitude and longitude). We chose to classify on room types because it is not a direct function of location and because it could potentially lead to some interesting insight on New York properties. For example, in which neighborhood is one most likely to have to share a room? Does paying a higher nightly price increase your odds of having a room all to yourself? We felt like these were interesting enough questions to examine for this exercise.

3.2 Question 2

In approaching this problem, we considered the borough of each listing, as well as its price and its "days available" attribute (i.e. the number of days out of the year in which the property is available for booking).

Just as before, the boroughs attribute was transformed using a one-out-of-K transformation while the price and availability attributes were regularized by centering the mean at 0 such that the standard deviation also equalled $|1|^2$.

For the ANN, we used roughly the same range of h values as before: A lower bound of 1 and an upper bound of 40 with increments of 4.

For the logistic regression, we experimented with various ranges of λ values. We found another logarithmic range from 10^-8 to 10^2.

3.3 Question 3

Outer fold	ANN		Logistic Regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
0	17.000	0.149	15.264	0.180	0.509
1	1.000	0.182	15.264	0.196	0.510
2	5.000	0.188	24.421	0.215	0.508
3	9.000	0.166	0.000	0.202	0.508
4	5.000	0.192	0.000	0.192	0.509
5	5.000	0.200	9.541	0.215	0.512
6	25.000	0.160	0.000	0.176	0.511
7	25.000	0.178	9.541	0.215	0.511
8	25.000	0.202	0.910	0.233	0.509
9	37.000	0.170	0.000	0.186	0.509

Table 2: Regression comparison results

Our classification models (excluding the baseline) all perform fairly well. We get error percentages around 15-25% which seems to indicate our models are functioning. Additionally the neural network, the most complicated of the models, consistently performs the best. This shows that the clustering of the attributes requires the more flexible decision boundaries of the neural networks to get the most accurate classification.

3.4 Question 4

	ANN vs. lin. reg.	ANN vs. baseline	lin. reg. vs. baseline
lower bound	$-.15649$	$-.19383$	$-.33164$
upper bound	$-.13928$	$-.15022$	$-.32871$
p-value	1.2237×10^{-11}	1.2385×10^{-8}	1.0868×10^{-21}

Based on these results coupled with the results from the previous question, we see that the ANN is the best model, the linear regression is the second-best, and the baseline is the worst, and that these results hold to a degree of statistical significance. Thus, it would make sense to use an ANN with this data. Note that these are essentially the same results we got in question 3 of part b of the regression section.

3.5 Question 5

We used a model with λ of 0 as this value of λ generated the smallest error. This resulted in the following weights:

Logistic Regression Model Coefficient Weights	
Feature	Weight
In Bronx	-0.04155114
In Brooklyn	0.03322442
In Manhattan	0.0653307
In Queens	-0.07687987
In Staten Island	0.01955367
Price	1.22619261
Availability	-0.12166959

Our logistic regression formula takes the form of the following with X values taken from our data-set and weights taken from the table above:

$$\hat{y} = \sigma(w_0 + w_1x[1] + w_2x[2] + w_3x[3] + w_4x[4] + w_5x[5] + w_6x[6] + w_7x[7])$$

Our classification parameters are slightly different than our regression parameters as they are two separate problems (Calculating price, a continuous variable using a set of features vs calculating room type, a variable that can have one of three values using a different set of features), but there are some similarities between the models. The "In Manhattan property" has the highest positive weight which parallels the regression, as its one of the dominant features in the data-set.

4 Discussion

4.1 Question 1

There are a few important takeaways from the regression portion of this exercise.

1. Though a somewhat intuitive conclusion, we found that the borough of a listing (specifically whether or not it's in Manhattan) and whether or not the listing is an entire property (rather than a shared room), seem to be the greatest determinants of price. In our linear regression, those attributes had the highest weights associated with them.

2. The relationship between attributes and price may not be as clear cut however. We found that our artificial neural network performed better than our linear regression, which indicates that the determinants of price may be more complicated than a simple linear relationship.

3. The actual error of our regression models still seem very high. We believe this has something to do with higher number of outlier, luxury apartments available in Manhattan that can cost upwards of a thousand dollars per night. In the future, it may be worth removing these expensive properties from our analysis or perhaps even building separate models for each borough.

Additionally, we were surprised by how much better our classification models performed over our regression models. We didn't think that room type would have such a strong relationship to other attributes and instead believed that the location of the listing would matter more. But given the difference in accuracy between our regression models (Essentially using location and room type to guess price) and our classification models (Essentially using location and price to guess room type), I think we can conclude that the room type is a more important consideration.

4.2 Question 2

Since our data comes from Kaggle, which is a collaborative data-set sharing platform focused on machine learning, we were able to find some discussion and preliminary analysis of our data-set. One such user also experimented with regressions using various attributes to try and predict the price of each listing. They found that the neighborhood of Manhattan was a major barrier for any sort of useful analysis. Because Manhattan is one of the most expensive places to live in the world, any model that tries to predict price will usually default to assigning heavy weight using the listing's borough which isn't always useful.

The user tries a couple of methods to solve this issue. First, they tried omitting all Manhattan properties from the analysis completely, but this removed a large number of entries (not only is Manhattan very expensive, but it is also very densely populated) and biased cheaper properties.

Like us, they concluded that examining each borough separately might be helpful in getting a better understanding of the less obvious variables that can effect the price of a listing.

5 Collaboration

Section	Luka	Weston	Michael
Regression A	60% — Wrote the majority of the code and worked on the write-up.	30% — Worked on the write-up and made a few additions and tweaks to the code to better document our process.	10% — Made a few additions and tweaks to the write-up.
Regression B	40% — Wrote the majority of the code and the first few sections of the write-up.	20% — Made a few additions and tweaks to the write-up.	40% — Assisted with writing the code for the stats focused problems and made a few additions and tweaks to the write-up.
Classification	50% — Wrote the majority of the code for this section.	40% — Wrote up the section, documented our process, and answered the questions.	10% — Repurposed code from the regression section to do the statistical analysis and made a few additions and tweaks to the write-up.
Discussion	33% — We discussed and wrote up the conclusions as a group	33% — We discussed and wrote up the conclusions as a group	33% — We discussed and wrote up the conclusions as a group