Project 5 Q11-14 Report

**Question 11**

The MSE obtained from various neural network architectures we test on are summarized below, with hidden layer sizes listed, grouped by the number of hidden layers. For each architecture, we test on both logistic and tanh activation functions. We did not use relu, because the resulting MSEs are become very high.

***One hidden layer:*** [2,5,10,50,100,200,300,400,500]

logistic :

[28754.204105085813, 28753.855694454414, 28788.59475232106, 28740.597246997637, 28711.89829202658, 28719.499150106578, 28641.6375867408, 28649.84564102266, 28540.37562984672]

Minimum MSE = 28540.37562984672

tanh :

[28775.593376809276, 28759.156362876012, 28788.4528120922, 28753.442763458068, 28706.560488627467, 28753.052086402895, 28750.588593469223, 28752.683491968422, 28749.134405383822]

Minimum MSE = 28706.560488627467

***Two hidden layers:*** [(10,10), (20, 10), (20,20), (50,50), (100,100), (200,200)]

logistic :

[28443.981591592787, 28444.09875841569, 28380.38820623823, 28378.480785872398, 28417.989732970007, 28384.012590969585]

Minimum MSE = 28378.480785872398

tanh :

[28788.103795861836, 28443.812738672033, 28380.381475064314, 28575.403698808364, 28457.392518130186, 28367.00906982105]

Minimum MSE = 28367.00906982105

We see that the architecture that gives the lowest MSE is using tanh activation function, with hidden layer sizes of (200, 200).

**Question 12**

We use StandardScaler to scale the data before feeding it to MLPRegressor with tanh activation function and two hidden layers with sizes (200, 200). The resulting MSE is 13684.990, which is much less than the MSE obtained without preprocessing the input by scaling it.

To understand this, we know that the idea behind StandardScaler is that it transforms the data so that the distribution will have zero mean and unit variance. This helps to train the model better in the sense that, scaling avoids the situation when one or several features dominate others in magnitude.

**Question 13**

We use gridSearchCV to perform grid search on MLPRegressor to find the parameters of the architecture that yields the best performance, for each window length described in Question 6. We use StandardScaler to scale the data before feeding it to the grid search function. The parameters we fine-tuned are shown in the table below:

|  |  |
| --- | --- |
| **Parameter** | **Range** |
| hidden layer sizes | (100,100,100), (20,20,20), (200,200), (100,100), (20,20), 20, 50, 100, 200, 500 |
| activation functions | "logistic", "relu", "tanh" |
| alpha | 0.01, 0.001, 0.0001 |
| maximum number of iterations | 1000, 2000, 3000, 5000, 10000 |

The optimal parameters for MLP Regressor for each period and the performance of the optimized MLP Regressor models are shown below:

|  |  |  |
| --- | --- | --- |
| **Period** | **Optimal Parameters** | **MSE** |
| 1  (before Feb 1, 8 AM, 1-hour window) | Optimal parameters for MLP Regressor: {'activation': 'tanh', 'alpha': 0.0001, 'hidden\_layer\_sizes': 100, 'max\_iter': 1000, 'random\_state': 42} | 1700.923 |
| 2  (between Feb 1, 8 AM and Feb 1, 8 PM, 5-min window) | Optimal parameters for MLP Regressor: {'activation': 'tanh', 'alpha': 0.0001, 'hidden\_layer\_sizes': (20, 20), 'max\_iter': 1000, 'random\_state': 42} | 4181.166 |
| 3  (after Feb 1, 8 PM, 1-hour window) | Optimal parameters for MLP Regressor: {'activation': 'tanh', 'alpha': 0.01, 'hidden\_layer\_sizes': (100, 100), 'max\_iter': 1000, 'random\_state': 42} | 643.155 |

**Question 14**

In this question, we use random forest regressor to predict the number of tweets on the 6-th window, using the data from the previous 5 windows to train the model. The time span of the dataset for the all samples across different periods are shown below.

Sample 0 start and end date for period 1:

PST dataset start date: 2015-01-31 05:00:37-08:00

PST dataset end date: 2015-01-31 10:30:38-08:00

Sample 1 start and end date for period 1:

PST dataset start date: 2015-02-01 01:00:10-08:00

PST dataset end date: 2015-02-01 05:40:26-08:00

Sample 2 start and end date for period 1:

PST dataset start date: 2015-01-26 17:05:59-08:00

PST dataset end date: 2015-01-26 20:49:01-08:00

Sample 0 start and end date for period 2:

PST dataset start date: 2015-02-01 19:00:05-08:00

PST dataset end date: 2015-02-01 19:29:48-08:00

Sample 1 start and end date for period 2:

PST dataset start date: 2015-02-01 12:30:18-08:00

PST dataset end date: 2015-02-01 12:59:51-08:00

Sample 2 start and end date for period 2:

PST dataset start date: 2015-02-01 08:30:12-08:00

PST dataset end date: 2015-02-01 08:59:15-08:00

Sample 0 start and end date for period 3:

PST dataset start date: 2015-02-04 00:00:16-08:00

PST dataset end date: 2015-02-04 05:46:07-08:00

Sample 1 start and end date for period 3:

PST dataset start date: 2015-02-05 20:04:00-08:00

PST dataset end date: 2015-02-06 01:50:04-08:00

Sample 2 start and end date for period 3:

PST dataset start date: 2015-02-05 17:00:04-08:00

PST dataset end date: 2015-02-05 22:35:12-08:00

We see that the time span for each sample is not identical to each other, and not exactly 6 hours for periods 1 and 3, and not exactly 30 minutes for period 2. Therefore, we would expect to see that the predicted value deviates from the true value.

We yield the following result:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Period # | Period 1 (1-hour window) | | | | Period 2 (5-min window) | | | | Period 3 (1-hour window) | | | |
| Sample # | 0 | 1 | 2 | Mean | 0 | 1 | 2 | Mean | 0 | 1 | 2 | Mean |
| Prediction | 182.1 | 486.2 | 199.4 | 289.2 | 3846.6 | 3725.6 | 967.3 | 2846.5 | 3387.7 | 3387.7 | 3387.7 | 3387.7 |
| Normalized Prediction | 3.0 | 8.1 | 3.3 | 4.8 | 3846.6 | 3725.6 | 967.3 | 2846.5 | 56.5 | 56.5 | 56.5 | 56.5 |

In the last row, we normalized our predictions to a rate of number of tweets per 5 min. We observe from the results that for the period where the event of Super Bowl is included, the model predicts a significantly higher number of tweets in the 5-minute window, compared to the 1-hour windows for the other periods. This is reasonable, since people tend to discuss about the news, and it is likely that more people tweet about the Super bowl.