1. title

Hi everyone, I am XX and this my teammate XX and XX. Our project is about “Appliances energy prediction Data Set”. In short, the data set is mainly about humidity, temperature data collected by wireless sensors and the power consumption which is dependent on these air conditions. The entries are recorded every 10 minutes. Our goal is to construct a model for predicting the power consumption based on the real-time air condition.

2. Data Overview

There are nearly 20000 records in the data set, and they have 29 attributes which could support the prediction. All the variables are quantitative data, but we still need to preprocess first.

3. Data Overview

The graph on the left is a summary of the dataset. Limited by the size of the page, we can not show all the 29 columns here. But the basic situation is all these temperatures and humidity are similar. In addition, there are columns of wind speed and visibility, which are numeric and relevant to the prediction. It is notable that the data are collected in a house, generally we can choose to exclude the “date” column from our model based on the assumption that influence of sunlight is represented by the air condition in the house.

4. Data Overview

From the result of the method “describe”, we can see that the standard deviation of the column “Appliances” is much higher than the rest columns. All these variables are in different scale, we need to deal with that before any further analysis.

5. Data Overview

Before any further analysis, we have to make sure all null values are replaced with median or mean. We use the pandas built-in function “isnull” to find these null values, the result shows that there is no null value in the whole dataset. So we can continue our task.

6. Data Visualization

This graph demonstrates 5 columns of temperature, the range is approximately 15 to 30.

7. Data Visualization

This graph is about 5 columns of humidity. The values are in a wider range comparing to the previous slide. One conventional approach for dealing with it is to transform these columns to the same range: 0 to 1, no matter which model is utilized here.

8. Data Visualization

One objective is to predict the column “Appliances”. The value varies from 10 to 1080. There are 92 unique values and most of the occurrences are below 200.

9. Data Visualization

Another objective is the column “lights”. There are only 8 unique values and the range is smaller. We decided to work on this goal first since the situation is simpler. Because the number of different values is only 8, we supposed we can consider this column to be categorical.

10. Data Preprocessing

First, we transformed the column “lights” in to one-hot code, in other words, one column is extended into 8 columns and each of these indicates one value. And all the other numeric values are in the same range between 0 and 1. Because this is one kind of multivariable time series, we should not shuffle the order, otherwise the model would be problematic. We chose to create the training set with the first 80% records, and test set with the last 20%.

11. Data Preprocessing

For this preprocessing task, we utilize the functions from the “sklearn” package. First, we transform the values of column “lights” to labels with the function “LabelEncoder”. Then we transform these labels to one-hot code with the function “OneHotEncoder”. For these numeric columns, we use the function “MinMaxScaler” to put them between 0 and 1.

12. Model Selection

For this dataset, we chose RNN for the predicting model. Specifically, the LSTM model. The reason for this decision is that the records consists a time-related series, the sequence of the entries is significant for predicting and the status of any time point will impact the upcoming status. In short, the model should have memories. In this slide, the LSTM sell represents one single layer, all these addition and multiplication signs are performing element-wise calculations. All these parameters are trained to judge what to remember and what to forget. h(t) is the output to the next layer and C(t) is the memory for the current LSTM layer.

13. Model Selection(“lights”)

Assuming the inner correlations are not very complex, we created the neural network with 3 layers first. The first layer is the input layer with 26 neurons which has the same number with the input columns. We set the second layer with 16 neurons since 16 is in the middle of 26 and 8. Plus, add a dropout layer with the rate 20% to avoid overfitting. For categorical prediction, we use “cross-entropy” as the loss function and “softmax” for the activation function. We chose the “rmsprop” optimizer for gradient descent which will reduce the fluctuation during the gradient descending.

14. Model Training(“lights”)

We ran 4 iterations, and with “epochs=200”, “batch\_size=2048” for each. In the last iteration, we found that that the accuracy of categorical prediction reaches 89%, almost 90%, but from the graph we can see that most of the red lines, which indicate the real energy, is not overlain by the blue lines, which indicate the predicted value. It seems the accuracies are not consistent with the graphs.

15. Model Training(“lights”)

We thought there may be some kinds of overfitting. So, we turned down the step of iteration with “epochs=100”, “batch\_size=512” for each. But the result didn’t change too much. We checked the code for a while, still found nothing wrong.

16. Model Diagnosis(“lights”)

Finally, after checked the data again, we realized that the data is not balanced, it is biased. In short, about 77% of the values in column “lights” is zero. That means in most cases, when we build the model with this dataset, the value in that column can be easily predicted when true values are zeros, while provide less guarantee for other values. On the other hand, considering the previous graphs, there are more than 3000 records in a single graph, and the red vertical lines we saw is just a small portion of the values. Although the images are misleading, from the accuracy and mean squared error, we can see that the result is still acceptable.

17. Model Selection(“Appliances”)

Then we started to work on the column “Appliance”. In order to simplify the problem, we assume the values in this column are also categorical, so we use the similar commands to construct the neural network, except specifying different number of neurons in the hidden layer. It is notable that we can hardly draw 64 nodes or 92 nodes in the graph. If you find it misguiding just ignore it…

18. Model Training(“Appliances”)

But the result is not ideal, moreover, the accuracy is below 20% and the mean squared error is more than 10000. After 4 iterations it’s still underfitting. So, we conclude this model is not suitable for this scenario.

19. Model Selection(“Appliances”)

Then,the only option is to change the output layer to numeric, With optimizer of “adam” and loss function of “mean\_squared\_error”.

20. Model Training(“Appliances”)

From the results we can see that, at the end of the 2nd iteration, the mean squared error reaches 7629, much lower than the previous results, then it starts to overfit. Considering the range of this variable, prediction with root mean squared error below 90 is acceptable. For many values in the graph, our prediction only catches the trend, there are large gaps between the magnitudes. If we need to get a more precise model, more layer may be needed for extracting more deeper information. That is time consuming and we are planning to accomplish it in the future when we become more experienced.

21. end

After all these tasks, visualization, preprocessing and analyzing, we made some conclusions about this project. First, if the dataset is biased, the model may be biased too. And the most efficient way to solve this is to acquire more data which would make the dataset more balanced. Second, even if working with a single dataset, with different objectives, the optimal models maybe quite different, there is no guarantee of “one model for all purposes”. Third, when the records to display have a big amount, one single graph may not reveal the truth, probably, it may introduce misunderstanding and lead further mistakes. These are most of our jobs and we are looking forward to your comments and suggestions. Thank you!