EBA35303 Machine Learning and Forecasting
Term Paper - 40%

Instructions

Read carefully: Create and upload to WISEflow a **single** pdf file containing your answers and appendix. The pdf file must contain: i) questions as header and its corresponding answer, ii) numbers on all pages, iii) student's id generated by wiseflow. The pdf file cannot be longer than **10 pages**. Therefore, your answers should be short and to the point. Your Python code must be included in pdf format as an appendix and should be placed after your answers in the same pdf file you upload to WISEflow. Note that the Python code does not count as part of the 10 page limit. Make sure the code is readable! Do not include screenshots of Jupyter notebooks or the output of functions, for example, `print(df.head())`, `print(model.summary())`, `print(x)`, etc., as this could negatively affect the grade.

Honor Code: By answering this project, I confirm that I will not give or receive, from any person, group, or any AI software, any help in this project, as this is considered cheating. **Any suspected cheating or use of AI tools will be reported to the exam administration immediately and students will be called for an oral consultation as an additional verification before getting a final grade.**

Multinomial Classification

You have just been hired by Pulpo Data Analytics as a Junior Data Scientist. Pulpo Data Analytics performs consulting work for a major cell phone manufacturer, which is currently developing a new pricing model and a classifier model. Mr. Arau, Chief Analytics Officer (CAO) at Pulpo Data Analytics, was tasked by the cell phone manufacturer to understand the characteristics of cell phones that determine the price range of cell phones in the market and predict the price category of a new cell phone in the market given its attributes.

Mr. Arau calls your team for a meeting and explains all the details about the task and the data set that he has obtained from the cell phone manufacturer. In total, there are 20 features and one dependent variable, which corresponds to one of the low, medium, high, and very high cost price categories. See the appendix for details on the features and dependent variable. Mr. Arau suggests that your team build a classifier model with the data in the file `mobile_data.csv`.

1 Model selection, data visualization, and pre-processing (40 pts.)

1. Look carefully at the 20 features listed in the appendix and based on your knowledge about cell phones, select 5 features that you think can be important drivers to determine the price of the phone. Then,

   a) Argue why you select those 5 features and write down the equation describing a logistic regression based on your 5 features and the dependent variables.

   b) Plot densities, or normalized histograms, along with scatter plots for all 5 features. Both densities and scatter plots must show the four different price categories in the dependent variable. Comment on your plots; what do they tell you about the data?

   c) Take the logarithm of base 10 of the following variables: *battery_power*, *int_memory*, *px_height*, *px_width*, *mobile_wt*, *ram*, *sc_h*, *sc_w*, *talk_time*. Explain what is the purpose of it and why is it a good idea.

   d) Using 80% of the data, fit a multinomial logistic regression[1] using the 5 features you selected in

[1] The library `statsmodels` has a class called `MNLogit` that performs multinomial logistic regressions.

Exercise a) plus an intercept term. Note, if you didn't select any of the features from Exercise c), make sure to include one of those features in your logistic regression model, i.e. 6 features in total and the intercept. Report the accuracy of the model in the remaining 20% data. Note: make sure to use the seed value 12345 when creating the training and test sets.

e) Write down all model coefficient estimates together with their interpretation, i.e. what is the relationship between the dependent variable and independent variables according to the estimates?

f) Use a 5-fold cross-validation approach using the same variables as in Exercise d) and report the average and standard deviation of the accuracy of the model.

g) Discuss the main difference between the approaches in Exercises d) and f).

2 Neural networks and dimensionality reduction (40 pts.)

You want to impress Mr. Arau and remember that you learned how to use neural networks for (multinomial) classification problems in one of your bachelor courses. You decide to explore the following architectures

No. of hidden layers	No. of neurons
1	5
1	20
2	5
2	20
3	5
3	20

Table 1: Architecture of different neural networks.

Recall that the number of epochs required to train a neural network is not a trivial choice. Therefore, you split the data set as 80% training, 20% testing, and convert the dependent variable to one-hot-encoders. You can use 15% of the training set for validation. Note: make sure to use the same seed value 12345 as before. Then,

a) Use the same features you selected in the previous exercise and train[2] all the architectures in Table 1 for a large number of epochs. Then show the plots of the (categorical) accuracy of the validation and training data sets[3]. Comment on the graphs and report the number of epochs you choose to train each of the 6 architectures in Table 1. Use the categorical cross-entropy loss function to train all neural networks. You should choose the rest of hyperparameters appropriately.

b) Report the (categorical) accuracy in the test set for all architectures in Table 1.

c) Use principal component analysis (PCA) and find the transformations that preserve 20, 10, and 5 components. Make sure to use the original data, not the data where you use the log of base 10 and selected only some features. How much variability of the original data do 20, 10, and 5 components preserve?

d) With each of the 3 data transformations of exercise c), train the 6 architectures in Table 1. Report the (categorical) accuracy in the test set for the 18 different runs.

e) List the total number of weights for each of the 18 different trained neural networks.

[2]If you use Google Colab change your runtime to T4 GPU if available, as it will speed up model training.
[3]You don't have to show all 6 plots if they are similar and/or your report is close to the page limit.

3 Executive summary (20 pts.)

Your team receives a meeting request to present your findings to the top management of Pulpo Data Analytics. You have only 5-10 minutes to present your findings and conclusions. Therefore,

a) Create a table that summarizes the model performance of the 26 different models[4] you have tested using the test set.

b) Discuss pros and cons of the different approaches you have taken and give a recommendation. Which model should the cell phone manufacturer use as decision support system for a pricing model and which model as a classifier?

[4] 1 logistic regression with 5/6 features, 1 logistic regression with 5/6 features + crossvalidation, 6 neural networks with 5/6 features, 18 neural networks with 3 different PCA components.

Appendix

Features names of the data `mobile_data.csv` together with brief explanations.

data.txt

```
battery_power: Total energy a battery can store in one time measured in mAh
blue:          Has bluetooth or not
clock_speed:   Speed at which microprocessor executes instructions
dual_sim:      Has dual sim support or not
fc:            Front Camera mega pixels
four_g:        Has 4G or not
int_memory:    Internal Memory in Gigabytes
m_dep:         Mobile Depth in cm
mobile_wt:     Weight of mobile phone
n_cores:       Number of cores of processor
pc:            Primary Camera mega pixels
px_height:     Pixel Resolution Height
px_width:      Pixel Resolution Width
ram:           Random Access Memory in Mega Bytes
sc_h:          Screen Height of mobile in cm
sc_w:          Screen Width of mobile in cm
talk_time:     Longest time that a single battery charge will last when you are
three_g:       Has 3G or not
touch_screen:  Has touch screen or not
wifi:          Has wifi or not
price_range:   This is the target variable with value of 0 low cost, 1 medium cost,
               2 high cost and 3 very high cost.
```

Tips and Tricks

- Split the work! don't let people just look at you working.
- If you get accuracy values lower than 0.7 something is wrong.
- If you work with Google Colab, create one file and share it with all members of your team. The same file, not the same cell, can be modified by different persons simultaneously.
- Code loops and functions that can be (re)use.
- Take care of the format of your term paper, it matters! Use tables and/or figures to summarize your results.
- Comment and justify all your findings, even if it isn't asked!
- If you have difficulty with something, you can come to my office for help (send an email to set it up). However, depending on the question and the help you get, it could subtract points from your final grade on the term paper.