Homework

Iris Robyn Talgre
iris.robyn@hotmail.com
Student
University of Tartu

ChatGPT
chat.openai.com/chat
LLM
OpenAI

Abstract

In this homework I attempt to clean up a Zulip chat data and understand embeddings, linear and logistic regressions by analysing said data.

1 Introduction

With my lack of expertise in the tools we use for this homework and the topics covered, I was quite frustrated and confused. For days I couldn't prompt ChatGPT well enough to help me, only confuse me even more. So for the first plot I reviewed the lecture where homework was covered and traced those steps. Later after gathering some courage I turned back to ChatGPT which this time was very insightful and helped me plot the probability of a message's sentiment.

2 Linear regression

For linear regression analysis I looked at different words' occurrence. Since one of the variables is a boolean and there are a lot of data points the graph wasn't very intuitive. To better understand the graph dynamics, I tried different words. I think this analysis might have been more insightful if I had used logistic regression, but I did not think of anything else to use for linear regression. Altough I do think it might be interesting to swap the logistic regression I did with linear and vice versa. I will try that once ChatGPT is up and running again.
I also plotted message length in relation to sender, but that wasn't very interesting.
The formula for linear regression:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

Y - the dependent variable
β_0 - the y-intercept of the line, which is the value of y when x is equal to 0
β_1 - the slope of the regression line, represents the change in y for each unit change in x
X - independent variable
ϵ - the error term, represents the difference between the predicted and actual value of y

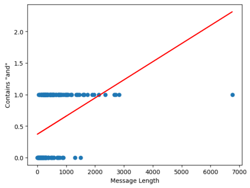


Figure 1: **Word "and" occurrence in relation to message length**

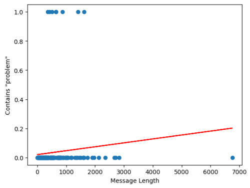


Figure 2: **Word "problem" occurrence in relation to message length**

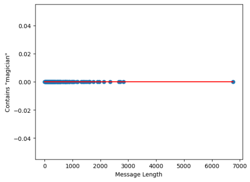


Figure 3: **Word "magician" occurrence in relation to message length**

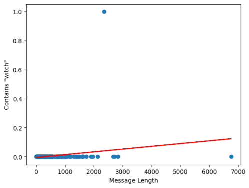


Figure 4: **Word ”witch” occurrence in relation to message length**

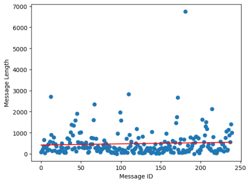
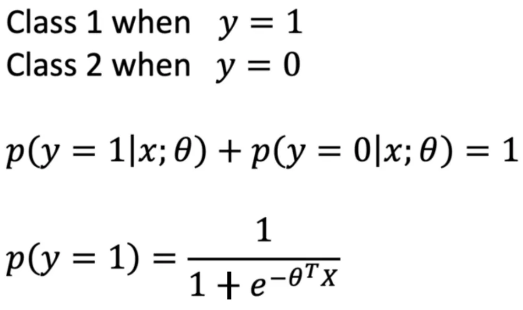


Figure 5: **Message length in relation to sender**

3 Logistic Regression

For logistic regression I used TextBlob to get the sentiments of messages and plotted them against message lengths. The plot shows the decision boundary. I chose this because ChatGPT recommended it, it was interesting and actually worked out quite well.
The formula for logistic regression:



$$\text{Class 1 when} \quad y = 1$$
$$\text{Class 2 when} \quad y = 0$$

$$p(y = 1|x;\theta) + p(y = 0|x;\theta) = 1$$

$$p(y = 1) = \frac{1}{1 + e^{-\theta^T X}}$$

Figure 6: θ - the estimated parameter vector and X - the vector of variables considered.

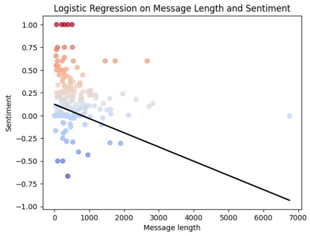


Figure 7: **Message sentiment in relation to message length**

4 Embeddings

For embeddings I traced the steps done in lecture and then asked ChatGPT for an alternative plot which was based on messages and subjects. I am not too sure about the quality of it, but I'm also still struggling to understand the visualizations. That didn't stop me making a 3D visualization as well though.

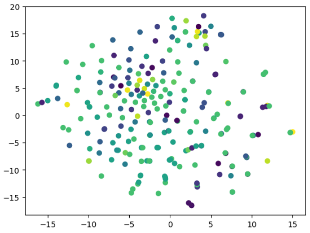


Figure 8: **2D embeddings visualization**

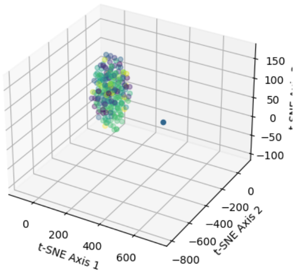


Figure 9: **3D embeddings visualization**

5 Conclusion

In conclusion this homework is subpar, but I will keep working on it to better understand linear and logistic regression and embeddings. I did also learn a lot and cannot wait to be proficient in using these tools.

References

Logistic regression and decision boundary
Embeddings