Anomaly Detection using Negative Selection Algorithms

Report on Assignment 2

Group 28

February 20, 2026

1 Introduction

Y.

2 Methodology

3 Results

4 Discussion

5 Conclusion

5.1 How to create Sections and Subsections

References

Appendices

Task 1: Using the Negative Selection Algorithm

This section contains the questions and answers from "Your task" on page 2 of the second assignment.

1. Compute the area under the receiver operating characteristic curve (AUC [1][2]) to quantify how well the negative selection algorithm with parameters $n = 10$ and different values of r ($r = 1$ until $r = 9$) discriminates individual English strings from Tagalog strings by using the files english.train for training and english.test as well as tagalog.test for testing. Which value of r belongs to what AUC in figure 1?

   Plot 1 has an r-value of 1, plot 2 has an r-value of 7, and plot 3 has an r-value of 4.

2. How does the AUC change when you modify the parameter r? Specifically, what behaviour do you observe at $r = 1$ and $r = 9$ and how can you explain this behaviour? Which value of r leads to the best discrimination?

   When increasing the r-value pass 3 the AUC curve "flattens", making it slightly worser at discrimination. At r-values 1, 8 and 9, the AUC value is close to 0.5, making it almost no better than random guessing.
   For $r = 3$ you get the best AUC, with a score of 0.8311.

3. The folder 'lang' contains strings from 4 other languages. Which languages can be best discriminated from English using the negative selection algorithm, and for which is this most difficult?
   The best AUC scores for all languages where found with a r-value of 3. The AUC scores of each language compared to english are given below:

   - **Middle-english:** 0.5424.
   - **Plautdietsch:** 0.7747.
   - **Hiligaynon:** 0.8397.
   - **Xhosa:** 0.8893.

   We observe that Middle-English is the hardest language to discriminate from English. This is a pretty logical conclusion since English is a direct descendant of Middle-english and both languages share vocabulary and many patterns. The best AUC value of 0.5424 is close to random guessing and the algorithm produces only a few matches, meaning that it is very difficult to distinguish Middle-endligh from regular English.
   On the other hand, Xhosa, Hiligaynon, and Plautdietsch were much easier to discriminate. Xhosa was the easiest to discriminate with an AUC score of 0.8893, followed by Hiligaynon with an AUC of 0.8397. These languages are very distant from English as they use different consonants and combine characters differently. Plautdietsch is the hardest out of these three to distinguish from English with an AUC score of 0.7747. It could be because is a dialect of German and shares some similarities with the English language.

4. We train the repertoire on strings of a certain length, but we could in theory input strings of other lengths. Explain in your own words the issue with providing strings that are too long/short.

   If we use shorter strings, it will be more difficult to find matches as the window would be much smaller. This would likely result in many 0 or non-matches. Longer strings would have the opposite effect. With longer input strings, we would be able to find more matching patterns. However, the anomalous part of a long string might be overshadowed by the high number of normal matches, making it harder to classify accurately.