Lab 1 - Report

TNM098 - Advanced Visual Data Analysis

Isabelle Rosenquist (isaro242)
Sandra Pettersson (sanpe282)

The task

The given data consist of 60 rows that each represent someone tossing a coin 200 times. Heads or tails is given by zeros or ones. Some of these rows are created by a random number generator and the others are real people trying to act like a random number generator. The task is to determine which of the 60 rows are made by a computer and which are made by humans.

Method

The first approach was to count the number of zeros and ones for each row. The distribution between the two possible outcomes were saved for every row. Then a mean distribution for the whole data set was determined. This mean value was then set to the threshold to determine if a row in the data set was created by a computer or a human. This theory was tested by asking a few people to generate a random sequence and the result were often very close to an even distribution. Since a computer also should provide an even distribution this method though was discarded in the beginning.

The next approach was to count and save the longest sequence of zeros or ones for each row. Since a human is less likely to use longer sequences of the same number because ‘it will not appear random’ a threshold had to be determined. This threshold was set to the mean value of each rows longest sequence. If the longest sequence exceeds the threshold the row is more likely to have been created by a computer.

The third and best approach was to see the distribution of combinations for each row. A sequence of four numbers were investigated. A four number sequence of zeros and ones have sixteen possible outcomes. For each row, the distribution of all possible combinations of four numbers were counted. The optimal distribution for each row is when each of the combinations is present 3.125 times. The difference between the actual number of occurrences for each combination and the optimal distribution was calculated for each row. The mean difference were then determined to see how close it came to the optimal distribution [1].

Result

With the third approach we found out that 25 of the 60 rows were created by humans. These humans are the following:

Number 2, 5, 7, 9, 12, 14, 15, 17, 20, 24, 25, 29, 30, 31, 32, 34, 37, 41, 42, 44, 47, 51, 52, 54 and 59.

Conclusion

When using the chosen method 16 humans were found. 40 percent of these also corresponds to the second approach, when counting how long the sequences of the same number were. This would indicate that the humans discovered are plausible.

The threshold used was 1.4. It was difficult to determine the way to get the best threshold but it was then decided that a mean value would be used. This value is the mean of each rows difference from the optimal distribution. Since the threshold could have been chosen by another approach the outcome could be significantly different. With a higher threshold less rows would be considered to have been created by a human.

References

[1] Khan Academy. 2012. *Frequency stability property short film*. [ONLINE] Available at: https://www.khanacademy.org/computing/computer-science/cryptography/crypt/v/frequency-stability. [Accessed 24 April 2018].