# CO395 Machine Learning
# CBC #4
# t-test


**Group 1**

Yong Wen Chua, ywc110

Thomas Morrison, tm1810

Marcin Baginski, mgb10

Marcin Kadziela, mk4910

# Contents

# Results of the t-test

## Clean dataset

| | DT vs. ANN | DT vs. CBR | ANN vs. CBR |
|---|---|---|---|
| Emotion 1 | 3.0498 | 8.9429 | 3.9477 |
| Emotion 2 | 4.2663 | 3.9580 | 1.0657 |
| Emotion 3 | 1.0668 | 6.0045 | 2.2500 |
| Emotion 4 | 4.6751 | 6.2453 | -1.2377 |
| Emotion 5 | 3.8522 | 6.9003 | 5.5578 |
| Emotion 6 | 2.9902 | 3.5254 | 0.5006 |

Table 1: t-values for every emotion and algorithm on the *clean* dataset

| | DT vs. ANN | DT vs. CBR | ANN vs. CBR |
|---|---|---|---|
| Emotion 1 | different | different | different |
| Emotion 2 | different | different | similar |
| Emotion 3 | similar | different | similar |
| Emotion 4 | different | different | similar |
| Emotion 5 | different | different | different |
| Emotion 6 | different | different | similar |

Table 2: t-values for every emotion and algorithm on the *clean* dataset

## Noisy dataset

| | DT vs. ANN | DT vs. CBR | ANN vs. CBR |
|---|---|---|---|
| Emotion 1 | 7.7947 | 8.4406 | 1.0476 |
| Emotion 2 | 6.2755 | 7.9588 | 2.2119 |
| Emotion 3 | 2.0966 | 1.9214 | 0.2191 |
| Emotion 4 | 2.5572 | 5.3489 | 0.5100 |
| Emotion 5 | 1.0530 | 1.8700 | 1.3814 |
| Emotion 6 | 3.9404 | 3.0950 | -0.7212 |

Table 3: t-values for every emotion and algorithm on the *noisy* dataset

| | DT vs. ANN | DT vs. CBR | ANN vs. CBR |
|---|---|---|---|
| Emotion 1 | different | different | similar |
| Emotion 2 | different | different | similar |
| Emotion 3 | similar | similar | similar |
| Emotion 4 | similar | different | similar |
| Emotion 5 | similar | similar | similar |
| Emotion 6 | different | different | similar |

Table 4: Interpretation of the t-values for every algorithm for the *noisy* dataset

# Questions

## Performance of the algorithms

How can we say that just by looking at the table? ANN has the most similarities, hence the best??
We cannot make general assumptions based on the results that we obtained. As mentioned in the previous report, neaural networks weren't giving us the same results after retraining them. That could be explained by the fact that initial weights and biases were randomised by MATLAB. There are also concerns regarding the data itself. We haven't taken a full advantage of similarity measure functions for CBR, as the data was one dimensional and binary.

## Adjustment of the significance level

Since we have 3 algorithms and we want to compare each with every other, we need $\frac{3\times 2}{2} = 3$ multiple comparisons. Our initially chose significance level was $\alpha = 0.05$ which, after applying the Bonferroni correction, is equal to $\alpha = \frac{0.05}{3} \approx 0.02$. The t-value for 9 degrees of freedom and $\alpha = 0.02$, which we used to determine whether the samples are statistically different, is $t = 2.821$.

## Type of the t-test

In each fold for each algorithm the test set consisted of the same examples. For this reasion, we used the paired t-test because the samples which we are comparing are clearly not independent.

## Classification error vs. $F_1$ measure

The inituitive explanation of that was is fact that for each emotion we wanted to take into account the whole result set. F1 measure would only focus on true positives, false positives and false negatives, while error rate also considers true negatives. In our experiment the data set is quite balanced therefore we can assume that this method will give reliable results. VERY BAD

## Trade-off between number of folds and examples

Increasing the number of folds will result in a better approximation of the distribution function (more samples) at the expense of correctness of each sample. As the fold takes into account less examples it is no longer general and becomes sensitive to untypical examples. In the extreme case (one fold) we only have one sample in each distrubtions and we can only compare their means. In another scenario (number is folds is equal to number of examples) we end up having a distribution that only consists of 0 and 1 samples, which indisposes us to conclude about the similarity of distributions.

## Additional emotions

If we wanted to add new emotions to the dataset, the Case Based Reasoning algorithm would require the fewest changes. We would simply add the new emotions to the existing Case Base which would be the only required change.

The Decision Trees and Neural Networks however would require a complete re-training. We would need to partition the new set of emotions together with the old ones into training and validation set and then run the training algorithm. This is obviously a more laborious task than for the CBR.