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	2.3857	4.3672	1.0573
Emotion 2	0.6936	2.8501	3.2322
Emotion 3	3.5037	2.7900	0.5018
Emotion 4	4.9059	6.2893	-0.1580
Emotion 5	0.2666	3.6875	3.1753
Emotion 6	3.5283	3.5196	0.2353

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

	DT vs. ANN	DT vs. CBR	ANN vs. CBR
Emotion 1	similar	different	similar
Emotion 2	similar	different	different
Emotion 3	different	similar	similar
Emotion 4	different	different	similar
Emotion 5	similar	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	-5.4336	-3.9836	0.7861
Emotion 2	6.5276	6.3766	0.2338
Emotion 3	2.8913	4.5010	0.2034
Emotion 4	3.2990	4.8491	-1.0284
Emotion 5	0.9481	2.9775	2.4377
Emotion 6	3.8547	4.4856	1.6618

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

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

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 independent, 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

Trade-off between number of folds and examples

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.