Technical University of Denmark

02445: Statistical evaluation of artificial intelligence systems

Project 1: Human arm trajectories in obstacle avoidance

Authors:	**Study numbers**
Karl Byberg Ulbæk	s183931
August Semrau Andersen	s183918

Summary

This project seeks to answer two major problems concerning a data set consisting of spatial measurements of arm trajectories.
Firstly, can we classify to which person a given trajectory belong. This classification task was undertaken by 8 simple machine learning models and yielded rather satisfying accuracies in the range of 50 to 84%. Performancewise some of the models were statistically distinguishable from one another.
Secondly, do the experiments have a significant effect on the trajectories. Through box plot inspections of the individual experiments as well as two way ANOVA all evidence pointed to wards the answer: yes the experiments (but also the subjects performing them) indisputably have an effect on the trajectories.

January 21, 2020

Introduction and data examination

Before considering what can be accomplished and inferred from the data, it is natural to explore and understand its capabilities and limits. Prior to addressing the first of two primary task in the project we will briefly introduce and familiarise ourselves with the data set.

The data originates from a simple experiment performed by right handed people who were tasked with moving a cylinder from A to B over an obstacle of varying height, and varying location between the direct path from A to B. The obstacle could be of 3 different heights and located at 5 different spots, amounting to a total of 15 different experiment setups plus one control without an obstacle. Each experiment was performed 10 times by 10 different subjects, resulting in a total of 1500 observations. Observations (executions of the experiment) was tracked with a motion sensor which recorded the spatial trajectory of the hand. Each recording was then post-processed to be exactly 100 coordinates along each of the x, y and z axis, thus resulting in 300 data points per observation.
An example of what the recordings of the trajectory could look like when plotted in 2D is presented below. This is the average of all 10 replications of experiment 9 performed by subject 1. (Only every third point has been plotted to reduce clutter.)

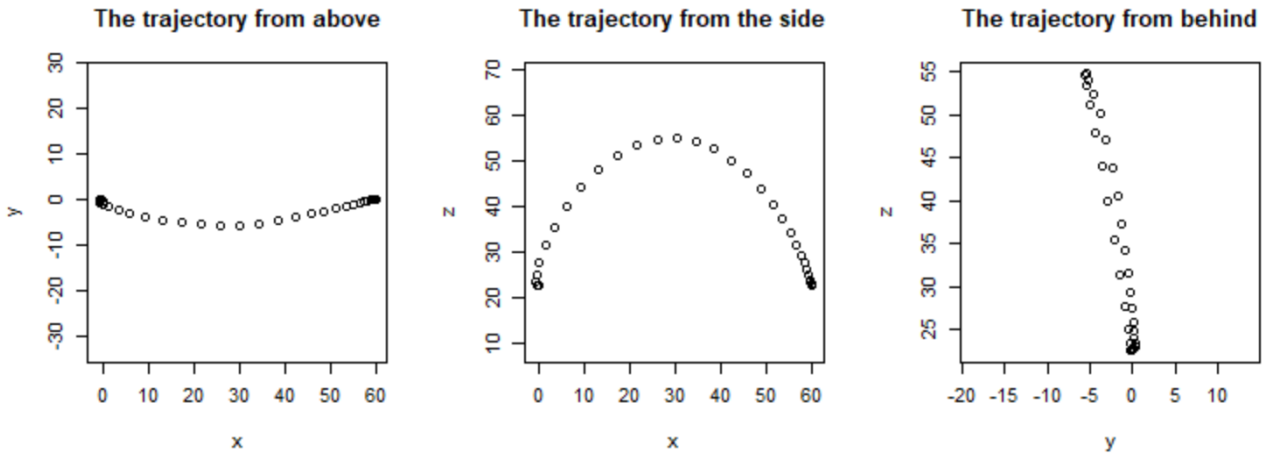

Figure 1: Average movement in experiment 9 by subject 1.

A visual inspection was done on plots across all 10 subjects in experiment 9 and all the following remarks are thus general trends. A few noteworthy conclusions to draw from the visual inspection of the plots are:

- The data is of very high quality as it appears very smooth and continuous, even so when plotting only one repetition and not an average of all 10 as in the figure above.
- As to be expected the movement is very symmetric. From the first and third pane it can be seen that the subject brings the cylinder closer to their body as their trajectory approach the middle of the obstacle. When inspecting the middle pane and directing the attention towards the density of the data points, it clearly shows that the speed of the movement is slowest in either end and fastest midway through.

- Lastly, the middle pane shows that the trajectory peaks half way through the movement which is in complete agreement with all subjects in experiment nine, as it was the setup with the tallest obstacle placed in the middle of point A and B.

Problem 1

Problem description

The data set has lot of potential in regards to what can be inferred from it. In this first part of the project, our main objective is to identify the person to which a given trajectory belongs. The point of reference for this classification task is experiment 9. The task involves 10 classes/subjects and 100 observations in total, 10 for each class. The classification is carried out by 8 different machine learning models, and their accuracy is evaluated and compared to one another.

The classification models

The 8 models chosen for classification are the following:

- Linear discriminant analysis (LDA)
- K-nearest neighbours (KNN)
- Gaussian naive bayes (GNB)
- Decision tree classifier (DT)
- Support vector machine (SVM)
- Random forest (RF)
- Regularized multinomial logistic regression (LR)
- Majority voting classifier (MV)

These models are all easily deployable in python with the ScikitLearn library, and we only did model optimisation to the regularized multinomial logistic regression. Each time the other models where trained once, LR was trained 50 times with 50 different lambda values in the range from 0.01 to 100, and the best model was picked. This has most likely led to a slightly biased results in favour of the LR. This could have been circumvented by implementing 2 layer/nested cross validation, but was not done as the focus of this project is statistics rather model tuning.
Lastly it should be noted that the majority voting classifier is not by itself a classifier, but rather an attempt at introducing democracy to machine learning. Jokes aside, it simply works by considering the results of all the other classifiers and then predicting on the label with most votes, hence the name. The fact that it relies on all the other models should be kept in mind when we later get to the comparing their individual performance.

Training and evaluation

Each of the 100 observations consist of 300 data points: 100 along the x, y and axis respectively. For training and testing, we transform our data into a 100 by 300 matrix, where each observation corresponds to a row with the first 100 columns being it's x-coordinates, the second 100 it's y-coordinates and the last 100 its z-coordinates.
The data matrix was standardised along the columns, i.e. subtracting columns by their corresponding mean and dividing by their corresponding standard deviation. The training and testing was performed using leave-one-out cross validation, hereby training all the models on the same 99 observations and then testing on the remaining last observation, repeated 100 times. The ordered testing accuracies are as follows:

- DT: 53 %
- KNN: 60 %
- RF: 69 %
- LDA: 72 %
- SVM: 73 %
- GNB: 76 %
- LR: 84 % (with optimal lambda 2.81)
- MV: 84 %

Comparison of the models

Considering that these are rather simple machine learning models, the results are quite impressive, and all the models are definitely better than baseline/random guessing of 10 % accuracy. The training and testing was designed such that all the models were trained and tested on the same data resulting in a paired design. The prevailing method for comparing two machine learning models in a setting like this is the McNemar test. With an H_0 null-hypothesis stating that two classification models perform the same: $H_0 = 0 = C1_{acc} - C1_{acc}$ (with a given alpha value, in our case the accustomed 0.05) the McNemar outputs a p-value. On a basic level, the McNemar test works by only taking instances where one model is correct and the other is wrong into account when computing the p-value. The reason for this being that instances where both models are correct must have been trivial and instances where both were wrong must have been unreasonably difficult. The McNemar test only works for comparing two models at a time, therefor we had to perform it for each model combination. The result is the following symmetric matrix.

	DT	KNN	RF	LDA	SVM	GNB	LR	MV
DT	NaN	0.31	0.01	0.00	0.00	0.00	0.00	0.00
KNN	0.31	NaN	0.19	0.07	0.01	0.01	0.00	0.00
RF	0.01	0.19	NaN	0.74	0.54	0.23	0.01	0.00
LDA	0.00	0.07	0.74	NaN	1.00	0.60	0.03	0.02
SVM	0.00	0.01	0.54	1.00	NaN	0.66	0.02	0.00
GNB	0.00	0.01	0.23	0.60	0.66	NaN	0.04	0.02
LR	0.00	0.00	0.01	0.03	0.02	0.04	NaN	1.25
MV	0.00	0.00	0.00	0.02	0.00	0.02	1.25	NaN

Table 1: McNemar test compairing all classification models, with black p-values indicating significant difference in performance.

When inspecting the last two columns, it can be observed that both LR and MV (the two best performing models) are significantly better than all other models except one another. Though, as already discussed, both of these models are biased to some extent and their results should be interpreted carefully.
If we instead direct our attention to the third best performing model, namely GNB found in column 6, we notice that it is only significantly better than the two poorest performing models. This also makes somewhat sense, considering how relatively similar the accuracy of models RF, LDA, SVM, GNB are, ranging from 69-76%. This similarity is further emphasised when examining the confidences intervals for the difference in accuracy between GNB and the other models.

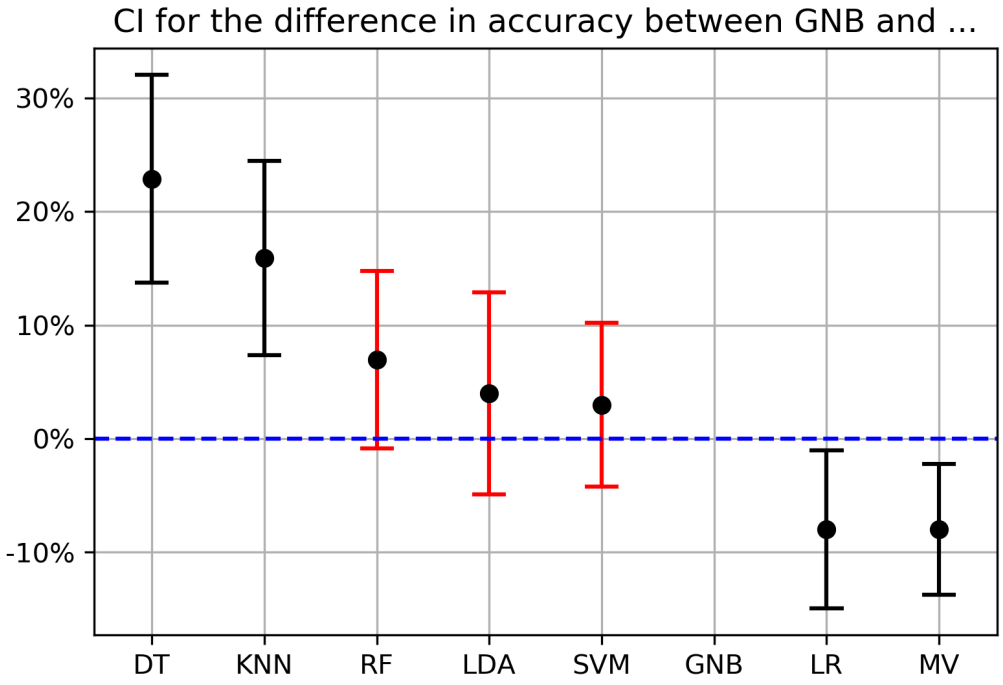

Figure 2: Confidence intervals for difference in accuracy between GNB and all other models.

As expected, the confidence intervals are in accordance with the p-values, as the insignificant p-values have confidence intervals which contain zero.

To sum up: Across the board, all models perform well, and we are able to deem some of the better "non biased" models performance significantly better than the worst.

Problem 2

Problem description

In this section we will take a step back and widen the scope of our data inference from one subset of the data to the whole data set that is all 15+1 experiments.
Instead of trying to distinguish the individual test subjects, we will instead try to distinguish between the experiments. The topic of interest this time around is: Do the different experiment setups have significant effect on the measured trajectories. This task can be interpreted in two ways.
1: Can we determine as few as two experiments to be statistically different? If yes, then we can answer yes to the overall problem statement as well.
We could also take it one step further and ask:
2: Can we find an experiment which statistically differs from all other experiments.

Method and Implementation

If we consider each experiment its own population, then the recorded data when a subject performs said experiment corresponds to a sample from that population. With this perspective on our data, we have 15+1 populations, with which we want to determine whether the experiment setup affects the population.
When facing a problem like this, the first thing that comes to mind is ANOVA, and this is exactly what we are going to do. The very convenient thing about the experiment design is that they used the same 10 subjects across all 15+1 experiments. This allows for a paired design making every person their own control across the experiments, thus opening up for using 2-way ANOVA. Before applying 2-way ANOVA, we condense our data to 1 data point per person per experiment, producing a 10 x 16 matrix, with subject 1-10 in the rows and experiment 1-16 in the columns. The condensation was simply computed as the mean of all the collected data belonging to one subject performing the 10 repetitions of one experiment.
The resulting matrix in its entirety is seen below:

	exp1	exp2	exp3	exp4	exp5	exp6	exp7	exp8	exp9	exp10	exp11	exp12	exp13	exp14	exp15	exp16
sub1	21.0	20.9	21.6	21.3	21.0	21.4	21.0	21.4	22.0	21.2	21.8	22.4	21.4	22.2	23.0	20.8
sub2	20.2	20.6	21.6	20.4	21.0	22.2	20.8	21.6	22.4	20.9	22.2	23.2	21.2	22.7	24.2	19.2
sub3	20.2	20.4	20.9	20.0	20.4	21.4	20.5	21.2	22.0	21.1	21.7	22.3	21.5	22.5	23.0	19.7
sub4	20.8	21.1	22.1	20.9	21.7	22.5	20.9	21.8	22.7	20.6	22.0	23.5	21.4	22.3	24.0	20.5
sub5	21.2	21.4	22.1	21.8	22.0	22.6	21.9	22.2	22.7	22.0	23.1	23.9	22.8	23.8	24.7	20.1
sub6	20.5	20.8	21.7	20.6	21.2	22.0	20.9	21.8	22.5	21.6	22.1	23.1	21.8	22.4	23.9	19.2
sub7	20.3	20.8	21.9	20.8	21.2	22.0	21.0	21.4	22.6	20.5	22.2	22.9	21.4	22.7	23.7	19.2
sub8	19.3	20.0	20.7	19.6	20.4	21.2	20.4	20.8	21.3	20.8	21.5	22.1	21.4	22.1	23.2	19.4
sub9	21.8	22.4	23.0	21.7	22.3	22.8	22.0	23.0	23.5	22.7	22.9	24.0	22.8	23.3	24.4	21.2
sub10	20.5	21.0	21.6	21.1	21.7	22.0	21.2	22.2	22.5	21.5	22.3	23.4	21.6	22.9	24.0	20.4

Table 2: Means of each subject for each experiment.

Not a whole lot can be learned from studying the matrix with the bare eye. On the contrary, if we make box plots of the subjects and the experiments it's an entirely different story.

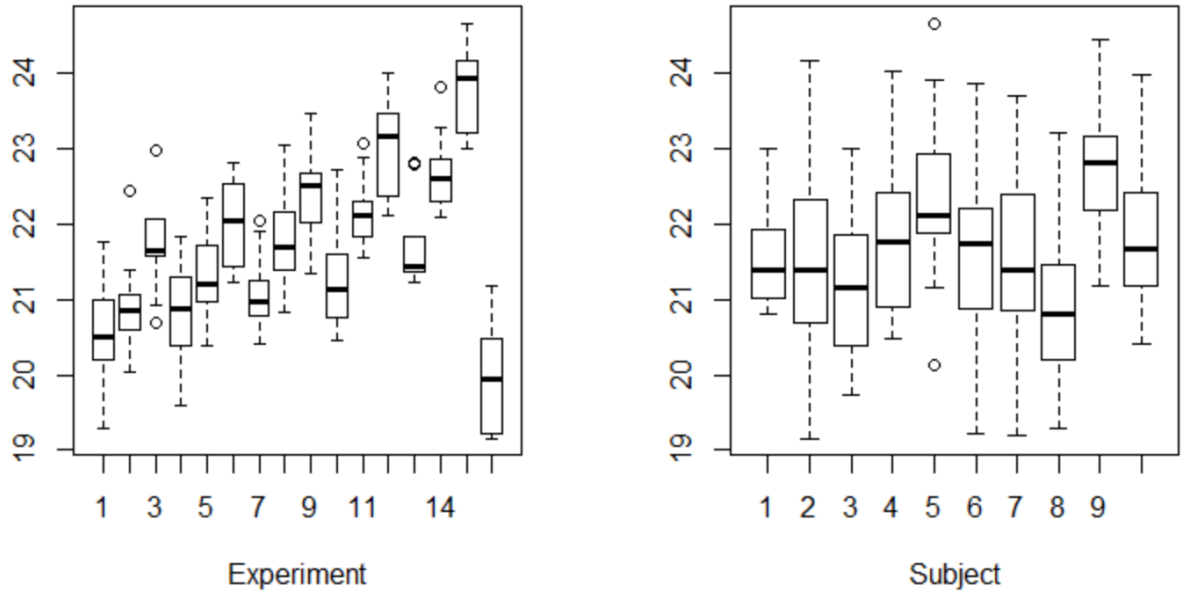

Figure 3: Box plots showing distribution of mean of individual experiments and subjects.

At a glance, it can easily be concluded that the experiments have a clear effect on the trajectories, and that the subjects too hava a noticeable effect on the trajectories. If we take a more thorough look at the experiment box plots, a pattern is apparent. It becomes evident that the box plots consistently follow in groups of 3. This is a direct consequence of the experiment design; as experiment 1 has the lowest obstacle, 2 the one with medium height, 3 the tallest, 4 the lowest etc. This makes sense as a taller obstacles lead to larger z values and thus an overall lager mean.
Furthermore it is distinct, that the means of the box plots increase as we move up in experiment number. This too can be explained from the study design; the density of the data points are consistently largest at the end of the movement. This is due to the subjects slowing down their movement as they carefully place the object onto the desired destination. A slower moment means more recorded data as the measuring equipment is time dependant. When the obstacle becomes increasingly displaced to the right, so does the trajectory, which causes z-coordinates of greater magnitude to enter the high density zone, resulting in an overall larger mean. This is best explained with a figure.
The following is an instance of the tallest obstacle displaced all the way to the left (experiment 3) and to the right (experiment 15) and how it affects the trajectory. The high density areas are highlighted by the blue boxes.

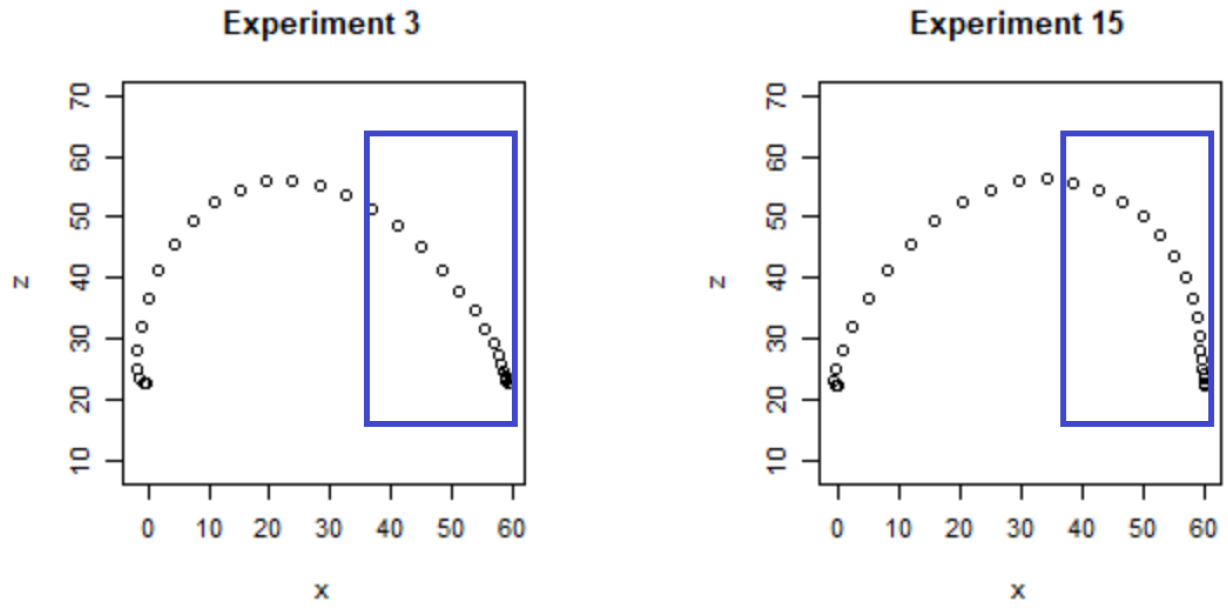

Figure 4: High density areas of movement are highlighted by the blue boxes.

From only visual inspection of the box plots, it is evident that the experiments have indisputable effect on the trajectories. For good measure we will of course still apply 2-way ANOVA on the 10x16 matrix, as we set out to do in the first place. Before getting to the results, lets confirm that our data is in agreement with the ANOVA assumptions of normal and independently distributed data, as well as approximately equal variance among the grounds. Normality can be confirmed from a qq-plot and an homogeneity of variance from a residual plot.

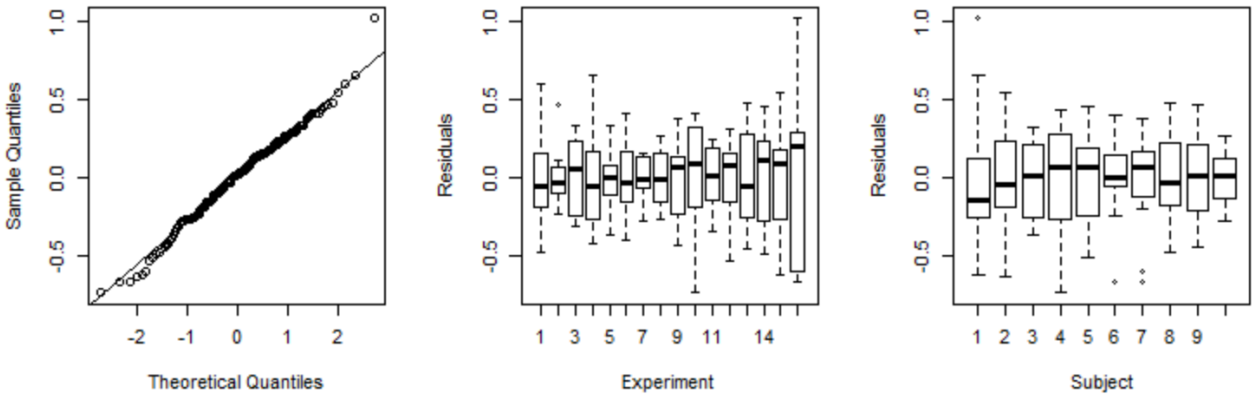

Figure 5: qq- and residual plots

It is safe to say that the normality assumption is very much satisfied. The homogeneity of variance on the contrary leaves a lot to be desired especially when it comes to experiments. We will do the ANOVA nevertheless. Both the ANOVA p-value for the experiments and the subjects turned out to $2.2 * 10^{-16}$: an extremely significant result. ($2.2 * 10^{-16}$ might not even be the actual value, but just the lowest R will show). Once and for all we can confirm both the experiment as well as the subjects affect the trajectory.

We will take our investigation one step further in an effort to try and determine an outlier experiment, or at least identity the one that differentiates the most from the rest. This will be undertaken by computing a paired t-test between every combination of experiments, resulting in a symmetric 16x16 matrix similar to the one from the classification problem.

	exp1	exp2	exp3	exp4	exp5	exp6	exp7	exp8	exp9	exp10	exp11	exp12	exp13	exp14	exp15	exp16
exp1	NaN	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
exp2	0.00	NaN	0.00	0.37	0.00	0.00	0.15	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
exp3	0.00	0.00	NaN	0.00	0.00	0.01	0.00	0.69	0.00	0.05	0.00	0.00	0.91	0.00	0.00	0.00
exp4	0.02	0.37	0.00	NaN	0.00	0.00	0.03	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
exp5	0.00	0.00	0.00	0.00	NaN	0.00	0.03	0.00	0.00	0.98	0.00	0.00	0.02	0.00	0.00	0.00
exp6	0.00	0.00	0.01	0.00	0.00	NaN	0.00	0.03	0.00	0.00	0.12	0.00	0.09	0.00	0.00	0.00
exp7	0.00	0.15	0.00	0.03	0.03	0.00	NaN	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00
exp8	0.00	0.00	0.69	0.00	0.00	0.03	0.00	NaN	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00
exp9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.04	0.00	0.00	0.11	0.00	0.00
exp10	0.00	0.03	0.05	0.04	0.98	0.00	0.11	0.00	0.00	NaN	0.00	0.00	0.00	0.00	0.00	0.00
exp11	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.04	0.00	NaN	0.00	0.00	0.00	0.00	0.00
exp12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.01	0.00	0.00
exp13	0.00	0.00	0.91	0.00	0.02	0.09	0.00	0.88	0.00	0.00	0.00	0.00	NaN	0.00	0.00	0.00
exp14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.01	0.00	NaN	0.00	0.00
exp15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NaN	0.00
exp16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NaN

Table 3: Paired t-test p-values, with black numbers being accepted under the null-hypothesis.

By reading down through the columns, it is revealed that 3 of the experiments are statistically different from all the others. These are: exp1, exp12, exp15 and of course exp16. Exp16 was expected as it was the control experiment.

Lastly we shift our attention back to the familiar exp9 from the classification problem. Here we would like to examine an instance of what the mean difference confidence interval could look like. The confidence interval is calculated from a paired t-test between experiment 9 and all the other experiments (given an alpha value of 0.05).

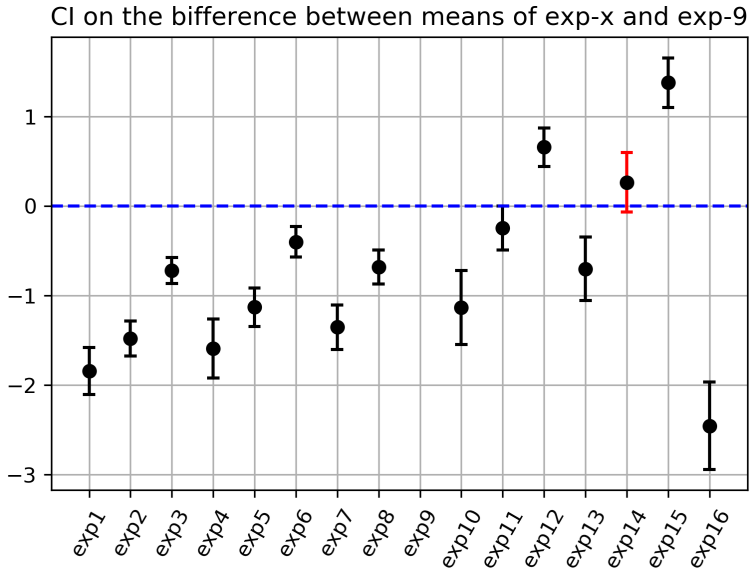

Figure 6: Confidence intervals of the mean difference between experiment 9 and all other experiments.

The results are of little surprise, as they were already spoiled by examining column 9 in the t-test matrix. Nonetheless it gives a better visual representation of how dissimilar experiment nine is from the rest, and how tight the confidence intervals are in general.

Conclusion

Regarding the classification problem on experiment 9 we found that all 8 classifiers were able to distinguish surprisingly well between the different classes: the worst having an accuracy of 50% and the best 84%.
With and alpha of 0.05, we can furthermore conclude that the two best performing models each were statistically better than the rest. This result should though be taken carefully, as these two models might have a slightly favourable bias. When comparing the third best model we were only able to conclude it was significantly better than the worst two.

When it came to evaluating whether the experiment had a significant effect on the trajectory, it became evident from the very beginning that this was definitely the case. From simple examination of the box plots distinct patterns, directly related to the experiment setup, were easily uncovered. We then went on to compute 2-way ANOVA which further emphasised the previous findings, with an extremely significant p-value of $2 * 10^{-16}$ for both the subjects and the experiments.
Lastly we were even able to conclude that experiment 1, 12 and 15 each were statistically significant from all other experiments given an alpha of 0.05.

Link to code in Github:
https://github.com/augu0093/02445-StatisticalEvaluation$_{E}xamProject1$