Technical University of Denmark

Peter Grønning, s183922
Rasmus Bryld, s183898
Lukas Leindals, s183920

Statistical evaluation of artificial intelligence systems - 02445

Project 1 - Human arm trajectories in obstacle avoidance

January 2020
Technical University of Denmark

Abstract

Looking at data from an arm-trajectory experiment, where participants were to lift a cylinder over another cylinder, this report suggests that it is possible to identify a person based on his/her arm-trajectory when lifting an object over an obstacle. Two different machine learning methods, k nearest neighbor and multinomial logistic regression, has been used to predict participants in 1 of 16 experiments based on their measured arm-trajectories, and both achieved accuracies around 70%. There was no significant difference in performance between the two models, so KNN is preferred due to being less computationally intensive. As each experiment differed in either the height of or distance to the obstacle, we tested whether experiment had a significant influence on the trajectories/curves. The test showed that experiments very likely had an effect on the curves, as the p-values for the null hypothesis were very small on all axes (the highest p-value being $6.2 \cdot 10^{-23}$), when curves were reduced to width, height, and length. The difference between experiments could be directly observed on a boxplot where axes values systematically increased with increased height of the obstacle, and decreased with increased distance to the obstacle.

1 Introduction

Lifting the ketchup over a small obstacle on the tabletop to place it close to the relative in demand may seem a trivial task, and its hard to imagine that people consciously pour their identity into the problem. In this report however, we will investigate if the simple task of moving a cylinder over another cylinder and placing it down, is so unique to different people that these people can be differentiated from one another based on the cylinder-trajectories. We will, more specifically use two machine learning methods to predict participants based on their trajectories/curves in 1 of 16 experiments, and do statistical evaluation of these methods. All 16 different experiments will furthermore be analyzed to test whether different experiments return different curves.

2 Data

The "armdata.RData" file consists of data based on arm trajectories from 16 different experiments. In each experiment 10 participants were recorded doing a movement in 10 repetitions each. Each repetition has a total of 100 points in each three-dimensional direction, resulting in 16 (experiments) x 10 (participants) x 10 (repetitions) x 100 (points) x 3 (axes) = 480.000 data points. The first part of the report looks into experiment 5.

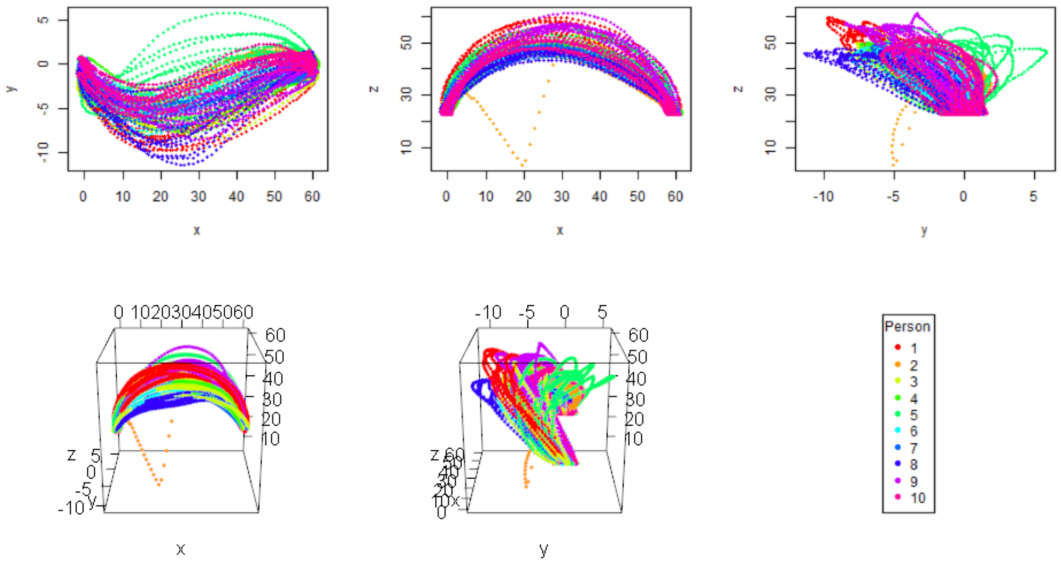


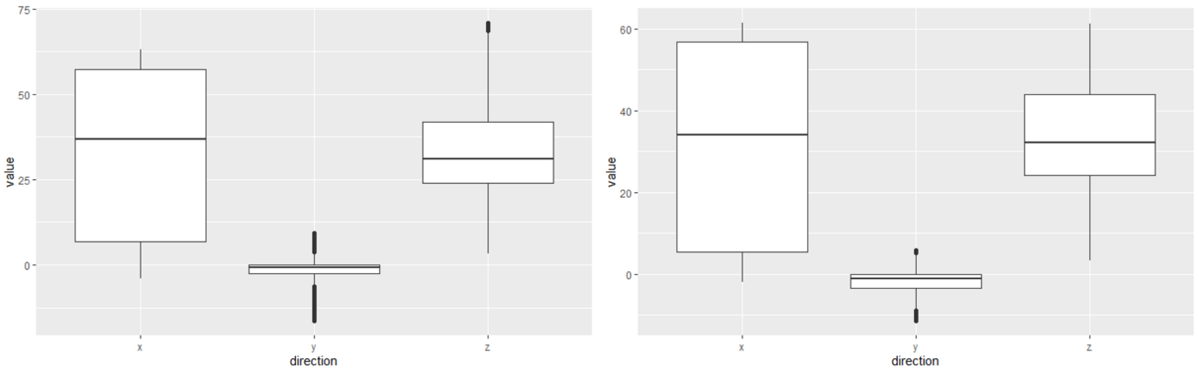
Figure 1: Plots of the movements in experiment 5, where each color represents a person

In figure 1 various plots of the movements in experiment 5 is plotted. From the plot we see that the z vs. x plot shows the intended direction of a movement, as this forms a parabular. Furthermore, we notice that person 2 seems to cause an outlier in one of the repetitions as one of the movements drops to a point near 0 in the z direction.

Summary statistics of the x, y and z coordinates is listed below in form of both a table and boxplots to give an overview of the data.

axis / Summary stat.	x	y	z
Min.:	-1.96	-11.42	3.34
1st Qu.:	5.38	-3.57	24.17
Median :	34.11	-1.13	32.21
mean :	31.47	-1.93	34.33
3rd Qu. :	56.84	-0.052	43.9
Max. :	61.44	5.83832	61.33

axis / Summary stat.	x	y	z
Min.:	-3.9	-16	3.3
1st Qu.:	6.9	-2.5	24.0
Median :	36.7	-0.85	31.1
Mean :	32.6	-1.45	33.9
3rd Qu.:	57.3	0.0074	41.9
Max. :	63.1	9.3	70.9



(a)

(b)

Figure 2: Boxplot and summary statistics for x, y and z axis. Left: Data from experiment 5. Right: Data from all experiments.

3 Methods and analysis

3.1 Preprocessing the data

The data from experiment 5 has been rearranged into a wide 100 x 302 data frame. The first 2 columns are factors describing the participant number and the repetition. The next 300 columns, are numeric and constructed by flattening x, y and z coordinates from a repetition. There are 100 rows as there are 10 participants with each 10 repetitions.
All 16 experiments are of interest in the second part of the report, so here the data is arranged in a 160.000 x 7 data frame. The first 4 columns describe the experiment, participant, repetition and point, which are all treated as factors. The next 3 are numeric x,y and z coordinates.

3.2 Classification

For the classification task the two supervised machine learning models, K Nearest Neighbors (KNN) and Multinomial Logistic Regression (MLR), are chosen - both well suited for classification with KNN being the most simple of the two. Before any evaluation of performance of the models can be made, the optimal parameter K (the number of nearest neighbors) in KNN has to be determined by parameter tuning. This is done using Leave-One-Out Cross Validation with K in the range 1 to 20. The model with the lowest estimated generalization error is then accepted as the optimal.

3.3 Evaluating the models

Using Leave-One-Out Cross Validation both models are fitted to the data with the same test/train-split and their predictions of the test data are stored in a binary list where 0 repre-

sents the classifier being wrong and 1 is equal to a correct prediction. 95 % Jeffrey's intervals for the true accuracy of the models are computed. These are derived from the beta distribution with parameters $\alpha = m + \frac{1}{2}$ and $\beta = n - m + \frac{1}{2}$, with n being the total number of observations and m being the number of correctly predicted observations.
To compare the accuracies of the two models a 2 by 2 contingency table as the one below is constructed and the McNemars test is used.

	Classifier2 Correct	Classifier2 Wrong
Classifier1 Correct	n_{11}	n_{10}
Classifier1 Wrong	n_{01}	n_{00}

Table 1: Contingency Table for the Mcnemars Test

The McNemars test only takes into account when the two models differ in their predictions, since n_{11} observations might be trivial and n_{00} observation might be impossible to predict [1]. The null hypothesis of the McNemars test is that there is no difference between the number of predictions where one classifier is correct and the other is wrong and vice versa:

$$H_0 : \quad \frac{n_{10}}{n_{10} + n_{01}} = \frac{1}{2}$$

A p-value from the McNemars test can thus be obtained by

$$p = 2 \cdot cdf_{binom}(\min(n_{10}, n_{01})|\theta = \frac{1}{2}, N = n_{10} + n_{01})$$

with $\min(n_{10}, n_{01})$ being the more extreme of the two observations.

3.4 Effect of experiment on curves

To test whether or not the experiments have a significant effect on the curves Analysis of Variance (ANOVA) is a suitable choice of method. In order to compare the curves they are reduced to three parameters: length, height and width - these parameters are calculated respectively by $\max(x) - \min(x)$, $\max(z) - \min(z)$ and $\max(y) - \min(y)$. As a result each experiment is reduced to 100 observations of the above 3 parameters, which are compared using ANOVA testing whether the curve parameters mean of each experiment differ from one another. Thus 3 ANOVA tests will be conducted (one for each parameter) resulting in three p-values, stating whether or not there is a significant difference in the means of the parameters between the experiments.

To avoid the multiple comparison problem, adjustments of the p-values will be made. This is done as we have 3 tests and the probability of rejecting a hypothesis that is true and the other way around increases. These types of errors are known as type 1 and type 2 errors. Specifically we want to address the probability of a type 1 error. Normally this set to a 5 % significance level, but as we perform 3 tests this increases to be a $1 - (1 - 0.5)^3 = 0.14\%$ significance level. To address this problem we choose to use the adjustment methods Bonferroni and Benjamini-Hochberg (BH). The Bonferroni method is a simple method where the significance level is divided by the number of comparisons n, thus the adjusted p-values will be given by

$$\widehat{p_{val}} = \frac{p_{val}}{n}$$

When using the BH method the p-values are sorted from smallest to largest and assigned a rank i from 1 to m, where m is the number comparisons, i.e. 3. The adjusted values are then calculated by

$$\widehat{p_{val}} = p_{val} \cdot \frac{m}{i}$$

4 Results

4.1 Classification on experiment 5

In the pre-experiment parameter tuning of the KNN, $K = 1$ is found as the optimal choice of parameter as $K = 1$ and $K = 3$ results in the smallest loss and $K = 1$ is less computationally expensive (see figure of CV in appendix A.2).

Below is the contingency table containing correct and wrong predictions of the classifiers and a table containing the accuracies of the model as well the accuracies' 95 % Jeffrey's interval and a p-value from the McNemar test.

	MLR Correct	MLR Wrong
KNN Correct	49	18
KNN Wrong	20	13

Table 2: Contingency Table for the McNemars Test

	MLR	KNN
Accuracy (%)	69	67
95% Confidence (%)	[59-77]	[57-76]
p-value	0.87	

Table 3: Accuracy of models and 95 % confidence interval as well as p-values from the McNemar test

4.2 Influence of experiments on curves

In figure (3) below, boxplots of max-min of x, y and z for each experiment are shown. They illustrate that the parameters appear to be normally distributed and the between group variance is fairly equal - both good reasons for the use of ANOVA.

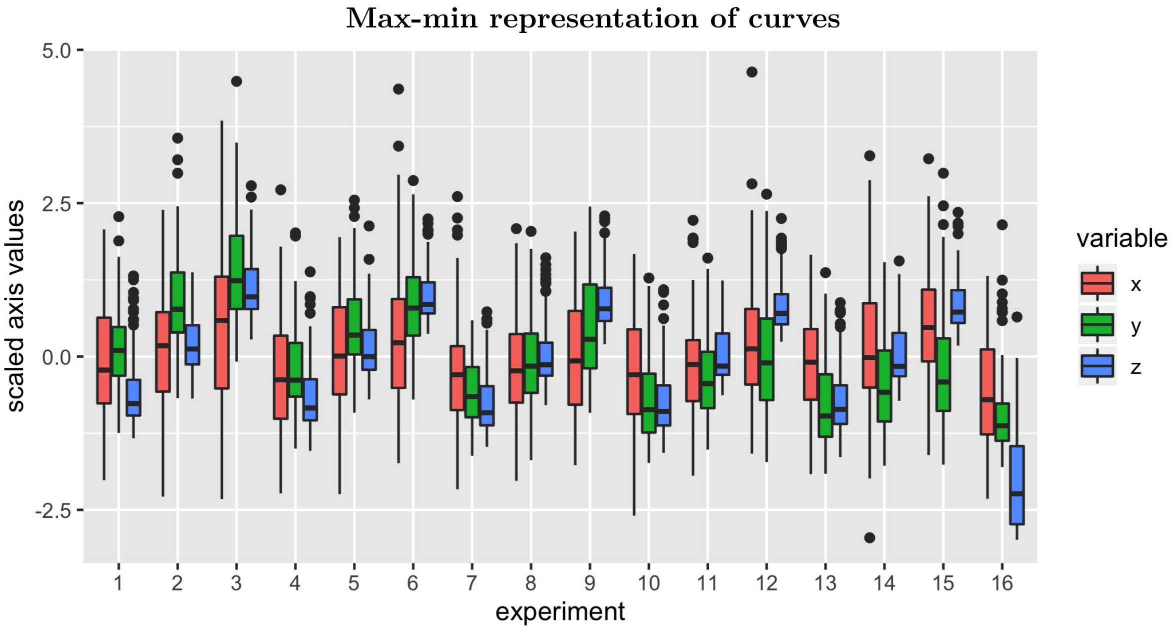


Figure 3: Boxplots of each experiment with max-min values from x, y and z axis, where each axis is scaled individually

The results of the ANOVA with 15 and 1584 degrees of freedom and adjusted p-values can be seen below in table 4. The p-values are all significant on a 95% significance level.

p-adjustment/ Axis	F-value	None	Bonferroni	BH
x	9.9261	6.2e-23	1.9e-22	6.2-23
y	83.805	2.2e-188	6.5e-188	3.2e-188
z	240.14	0	0	0

Table 4: Various correction methods applied to p-values from 3 different one-way ANOVAs

5 Discussion and conclusion

5.1 Classification

Both machine learning methods MLR and KNN predicted participants fairly well in experiment 5, with accuracies of 69% and 67% respectively and 95% confidence intervals of [57;77] and [57;76]. There was no significant difference in accuracy as the p-value for the null hypothesis was 0.87. As the performance does not differ, we could look at the computational power necessary to determine which model to use. KNN needs an extra layer of cross validation to determine the optimal parameter K, however MLR is far more computationally expensive and KNN therefore seems to be the optimal classifier of the two. As each curve was flattened into a 300 long vector with 100 values of each axis, the models did not exploit the sequential nature of the data, which could have been beneficial when classifying the person.

5.2 Experiment and curves

Our representation of the curves is plotted for each axis on figure 3. These plots indicate a strong effect of experiment on curves, and is consistent with the experimental setup, where the height of the obstacle was increased within each distance, and the distance was increased every third experiment. The last experiment had no obstacle.

The experiments had a significant influence on the curves according to the ANOVA as well (see table 4), when the curves are represented by length, height and width (max-min on each axis). When using ANOVA the variance within each group must be the same, which may not have been entirely the case and therefore might have had an influence on the results. This method of using max-min neglects most of the data in each curve, but we regard it a meaningful representation, as the experiments themselves are differentiated by the dimensions of, and distance to, the obstacles. This method has another drawback as it is sensitive to outliers. Each parameter is based on only two coordinates, so a single outlier value can influence the entire representation of the curve. As mentioned earlier (Figure 1), there is in fact an outlier in experiment 5, but it seems to be the entire curve at fault, not a single coordinate, so using e.g. the average value on each axis as parameters, would not have helped. As there was only 1 outlier curve (3 x 100 points), we did not remove or substitute the curve.

A Appendix

A.1 R-markdown used for project 1

https://github.com/s183920/02445_Statistical_evaluation_of_AI/blob/master/Proj_1/proj1.Rmd

A.2 Selecting K

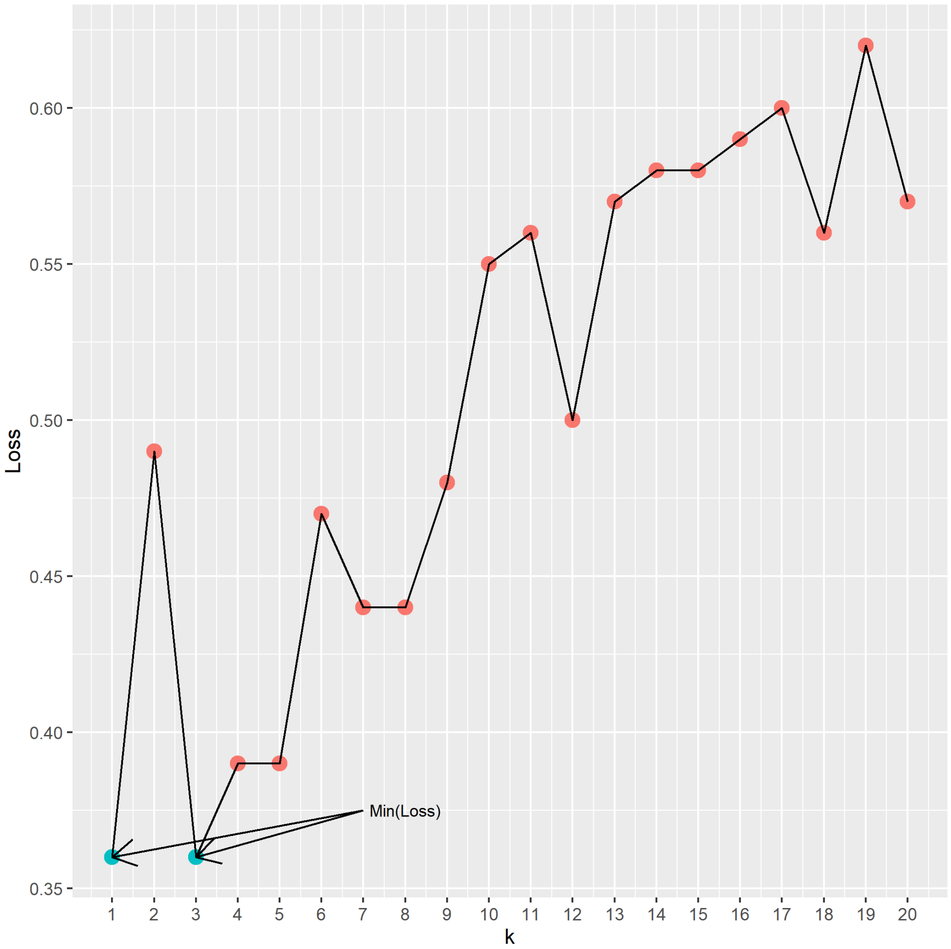


Figure 4: The loss for different values of k when performing KNN

References

[1] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. *Introduction to Machine Learning and Data Mining*. Technical University of Denmark, 2019.