Project 1
Course 02445
Project in Statistical evaluation of artificial intelligence

Rasmus J. P. s164564
Nikolaj S. P. s183930

January 2020

Summary

Classifying trajectories is a complex problem with many dimensions. In this report we will attempt to classify trajectories with two different machine learning models, a neural network and the K-Nearest-Neighbors algorithm. We evaluated each model's performances on how well they classify trajectories on new observations and compared the performance of both models. We found a significant difference between model performances in favor of the neural network $\alpha < 0.01$. In addition we analyzed 16 different experiments and their resulting trajectories and tested whether there was a significant effect of experiment on trajectories. Using a multivariate test-statistics for high dimensional data we are able to conclude that 115/120 pairs of two-sample comparisons of means were significantly different than one-another $\alpha = 0.05$ and conclude that for the vast majority of the trajectories there will be an significant effect of the experiment being performed.

1 Introduction

Solving complex problems has been the main drive for development in computer science and the computers has by far overceeded the humans on complex problems such as playing a game of chess or predicting the weather but only because we have been able to present them models simulating the real world for which the computer can react upon. So how do we model the real world? There are many answers, some complicated and some simple. We will be looking at trajectory data from 10 different test-subjects each performing 16 different experiments, repeated 10 times. Each experiment share the same underlying task with slight variation to the task. The task involved having human test-subjects move a cylinder over another cylinder. The experiments varied between different obstacle heights and obstacle positions.
Our first aim is to classify the supposed unique motion between test-subjects within the same experiment. The data from only a single experiment is then the 10 repetitions performed by each of the 10 test subjects, on which we will train and evaluate our models on. Sampling each model 30 times with leave-one-out cross validation LOOCV we will then compare the performance in a two-sampled t-test.
Second aim is to look for a significant effect from the experiments on the trajectories. Since the data is multivariate we will test for multivariate normality and if the results are negative we turn to central limit theorem CLT and reduce our dataset to a dataset of 160 mean trajectories, 10 means for each experiment. Finally we compare mean trajectories by using a generalized form of the Student's t-statistic namely Hotelling's t-squared statistics "t^2 which generalizes to p-dimensionality.

2 Data

The trajectory data was recorded in 3 dimensions using a motion capture camera, resulting in three continuous variables x,y and z, furthermore the data included information about which person performed the motion, in which repetition the motion was captured and which experiments was performed thus giving us three categorical variables.
Each trajectory observation contains 100 recordings of said coordinates - see figure 1. Conveniently computers does not observe data like us humans, so we transform the motion data from 3 x 100 observations to 1 x 300, effectively stacking 300 coordinates along one vector and doing so we have not changed the premise of the problem.

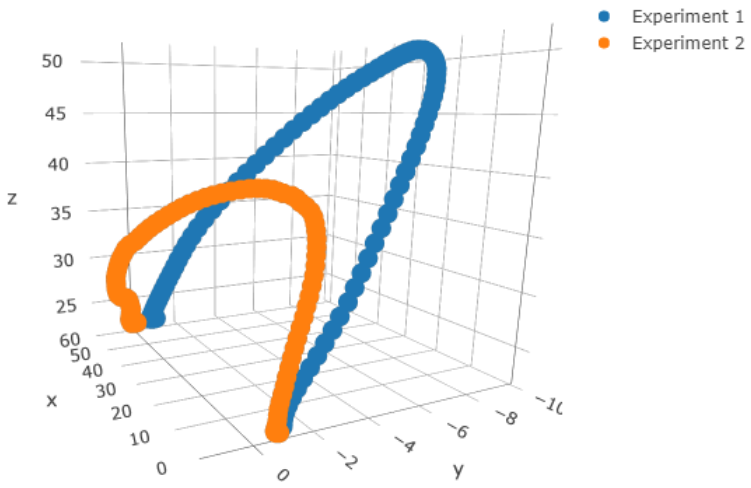

Figure 1: example of data

Subject 9 is missing the first 4 datapoints for some of the experiments, we will impute the values by replicating the first available datapoint, such that the first 5 datapoints become identical for those particular observation.

3 Comparing classifiers

We decided to compare two vastly different machine learning models, an ANN and a KNN, by running them multiple times and comparing the mean of our performance measure within each model. Performance measure was a simple measure of how many correct classification out of all possible. By training our models with the LOOCV algorithm, each fold only contains one test-entry, a measure of "1" for a correct classification and "0" for wrong classification was returned and all values summed after completion. We ran each model 30 times and by doing so recording 30 observations of a mean performance metric. CLT then tells us that these random variables will follow an approximate normal distribution and because of this fact we will be able to compare the two models by comparing the mean of their performance in a two-sampled Student's t-test.

3.1 Model A

We experimented with several versions of ANNs to find the architecture best suited for the task. ANN was decided on because of their performance on high dimensional

data. The classification network was trained to classify a person from the 1 x 300 long vector of motion data. See a description of the ANN layout in table 1.

Layer no.	Function
Layer 1	linear(300, 150)
	ReLU
	Dropout(0.15)
Layer 2	linear(150, 75)
	ReLU
	Dropout(0.15)
Layer 3	linear(75, 10)
Layer 4	Softmax

Table 1: Network parameters

3.2 Model B

The second model was a clustering model using the KNN algorithm, this model was trained on the same 1 x 300 motion vector. The KNN was chosen because of its simplicity and because it is very cheap computationally, compared to other models such as the ANN.

4 Testing the for experimental influence

If an experiment were to have a significant effect on the resulting trajectories, it should be significantly different from any other trajectory. The multiple test statistics to asses is then a collection of 120 comparisons i.a all possible combination of a single pair of experiments - duplicates and pairs of the same experiments removed.

The Hotelling's T-squared statistics is a generalized version of the Student's t statistic. It generalizes to p dimensions. This proposed test statistics raises two important points concerning our data:

1. Are the 100 repetitions within experiments multivariate **normally** distributed?
2. Having only 100 observations per experiment we are limited to less than 200 explaining variables per Hotelling's T-squared statistcs $p < n_x + n_y - 1$ where p is the number of dimensions in the multivariate observations.

The second point quickly becomes less of an issue, we will reduce the dimensionality of our data using principal component analysis PCA and choosing the number of principal components PCs by analyzing the square roots of the resulting eigenvalues i.e analyzing the variance explained by the eigenvectors. But first we must solve the multivariate normal assumption.

Our dataset must be considered a multivariate normal MVN distribution in order to be used in the t-squared test. The test for MVN resulted in a negative but it also showed that some variables were **uni**variate normally distributed see figure 2, other variables are considered outliers and are the main reason for a negative MVN test. When seeking for a set of variables to use for calculating the test statistic, a variable should be considered univariate normal between all experiments, we find that amongst 300 variables only 12 were lived up to this. Instead of removing the majority of the data we turn to central-limit-theorem CLT.

Univariate- and non univariate normal distributions

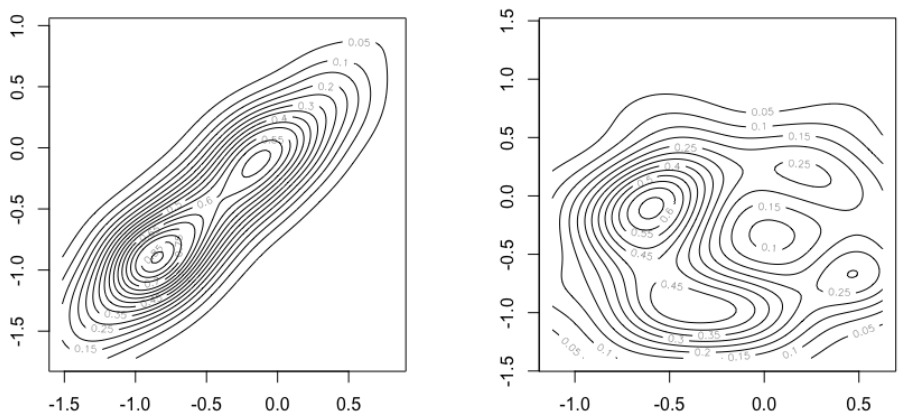

Figure 2: The figure to the right shows two univariate variables (x_8 and x_10), the second two non-univariate variables "outliers variables" (x_1 and y_1). *Two peaks indicate multivariate normallity in two dimensions.*

4.1 Normality in means

We have enough observations to assume a normal distribution within single experiments but we cannot expect the motion of one subject in an experiment to be identically distributed to that of another subject's motion and thus we are forced to reduce our data to one mean observation per subject per experiment. Another problem then arises, we have just reduced our data set from a large set of 1600 individual observation to a mere 160 and when we compare two experiments the number is even lower at only 10 observation, down from 100 and the scenario is that we now must limit ourselves to fewer variables. Hotellinger's t-squared test requires a maximum of $n_x + n_y - 1$ explaining variables. A set of 199 variables is more than we need anyway since our PCA dimension reduction served us with a 82% variance explained by the

first 9 PCs, see figure 3 for an illustration. We project the matrix of means onto the rotation matrix returned from PCA and continue with the t-squared test on only the 9 most explaining PC.

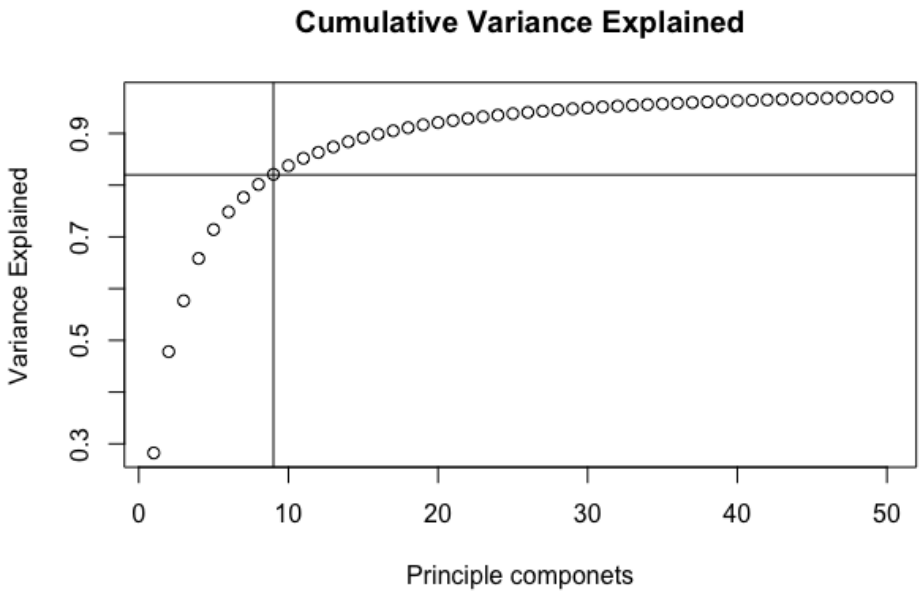

Figure 3: "Cumulative variance explained by principal components"

5 Results

LOOCV was performed 30 times to get a good estimate of the variance of the generalization accuracy of the two classification models, from this a confidence interval was calculated 2. Table 2 show the confidence intervals for both the ANN and the KNN model, the first value in second column is the generalized accuracy measure i.e estimated percentage of successful classification. Adding and subtracting the second value from the first in the second column return the 95% confidence intervals. We see that the two intervals does not overlap, provided a 0.95 confidence level.

Table 2: Confidence interval of classifiers

Model	**CI of Generalization Accuracy**
ANN	0.708 ± 0.0078
KNN	0.644 ± 0.0066

The same story follows from the paired test seen in table 3. We notice a very high test score of 10.934 and subsequently the resulting p-value which is approximately zero. The very reason for such high test score is that all 30 observations per model each come from the LOOCV algorithm were we effectively train and test on all the data available and the resulting values will be very close to equal which gives us very little variance also evident from the small confidence interval in table 2 as a result we can with very high confidence say that difference in performance between the two model is not zero and that the artificial neural network is the better classifier.

Table 3: Two-sampled paired comparison of means

Test	**Test Statistic**	**p-value**
Paired t-test	10.934	$8e-12$

Testing for effect on trajectories from experiments we compared 120 unique pairs of two experiments. Out of the 120 pairs 5 pairs where found to have no significant difference see 4. Interestingly most non-significant pairs, 3/5, comes from the setup of "S" small height of obstacle, and when the obstacle is moved one position of 7.5cm between the two experiments. And the two other non-significant pairs are also between two experiments with one positions difference of the obstacle on the same height of obstacle. There is a significant difference between the curves of the experiments, which means the experiments has an influence on the resulting curves.

Table 4: Non-significant pairs of Experiments

Experiment pair	**p-value**
1, 4	0.137
4, 7	0.177
7, 10	0.278
5, 8	0.0855
6, 9	0.162

6 Conclusion

We found that it is possible to classify a person from the motion of their hand within an experiment. The two classifiers trained ANN and KNN had an generalization accuracy of 70.8% and 64.4% respectively, thus way outperforming the 10% generalization accuracy one would get by random guess. Of the two classification models the ANN performed the best. The experiment has an influence on the resulting curves when heavily considering what data we use. iI is hard to assume multivariate in motion data from multiple test-subjects and thus we feel intrigued to further pursue methods for this subject.

7 Appendix

Public Github link for code and files: https://github.com/realnikolaj/02445