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 [ML] models, a neural network NN and an unsupervised clustering algorithm. We evaluated the model performance on how well they classified the trajectories and subsequently predicted which test-subject performed said trajectory and compared the performance of both models statistically. We conclude that there is not a significant difference in the performance between our neural network and the unsupervised clustering model ... *[conclusion & results]* ... In addition we analyzed 16 different experiments and their resulting trajectories and tested whether there was a significant effect of experiment on trajectories. By using a multivariate test-statistics for high dimensional data we are able to conclude that different experiments results in different trajectories this with a signifance level of ¡1%, we obtained a p-value for this very statistics on [result].

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 it. The task for the test-subjects was that they had to move an arbitrary cylinder over another cylinder. The experiments varied between different obstacle and obstacle positions.
Our first aim is to classify the unique trajectories from the resulting 1600 observations and evaluate the performance of our two classifiers and compare their mean squared error using two-sampled t-test. The second aim is to look for a significant effect from the experiments on the trajectories also here we will be using a form of t-test but since we are now looking at a trajectory as a whole we must use a test-statistic which takes the dimensionality into considerations, we end up comparing trajectories by using a generalized form of the Student's t-statistic named Hotelling's t-squared statistics t2 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 TRAJ. A computer doesn't observe data like humans, so we decided to transform the motion data from 3 x 100 observations to 1 x 300, effectively stacking 300 coordinates along one vector.

Person 9 is missing some of the initial 1-4 datapoints for some of the experiments (not experiment 2), we impute the values with the first available datapoints, such that the first 1-5 datapoints are the same for those particular observation. This seems reasonable since it is only a few datapoints and they are located at the beginning of the observation, thus it is equivalent to the person starting from that position and holding it there for the first few measurements.

Include a few plots - (Distribution of curves HOW?) TRAJ, variance between curves (boxplots),

3 Comparing classifiers

We decided to compare two vastly different machine learning models, an ANN and i KNN, by running them multiple times and comparing the mean of our performance measure within each model. We ran each model 30 times and by doing so recording 30 independent and identically distributed random variables for each model. Central limit theorem 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.

We propose the null-hypothesis H_0: The difference in the means is zero.

3.1 Model A

We experimented with several versions of ANNs to find the architecture best suited for the task. We decided to use an ANN because of their already established performance in high dimensional space. The classification network was trained to classify a person from the 1 x 300 long vector of motion data.

3.2 Model B

The second model was a clustering model using the K-Nearest-neighbor KNN algorithm, also this model was trained on the same 1 x 300 motion vector. The KNN was choosen because of its simplicity and because it's very cheap computationally compared to other models such as the NN.
Leave one out Cross Validation LOOCV was performed on both models and their performance evaluated with the zero-one loss function suitable for classification problems. NOTE: Might consider another. SVM hinge loss, TOP-k ... because multiclass.

4 Testing the for experimental influence

We decided on a test statistic of difference between experiments. If the experiment were to have a significant influence on the resulting trajectories, they should all be significantly different from one-another. The multiple test statistics to asses is then a collection of 128 comparisons between each and all experiments from which we will make our conclusion upon.
This raises two points of interest concerning our data: 1. Can the distribution of the repititons withing experiments be considered multivariate **normal**? 2. Having only 100 observations per experiment we are limited to less than 100 explaining variables "Insert a reason here".

A solution for the second concern is to reduce the dimensionallity by performing principal component analysis PCA and choosing the number of principal components PC's by analyzing the resulting decompisition matrix i.e variance explained. But first we must solve the multivariate normal assumption.
We have enough observations to assume normal distribution but we should only do this within single test subjects. Within each test subject we only have 10 observations per experiment. Luckely 10 observations is still enough for the mean of a single test-subjects trajectories to be considered normally distributed.
The solution to the first point then becomes to collect 10 means, one from each test subject withing each experiment but we've just limited ourselves to a maximum of

nine explaining variables down from 300 have we???, see later highlight.
Fortunately performing PCA returns a 63 vs 95% ..see highlighted note.. variance explained by only 9 PC's. What to do, can we use just the 10 means to compute PCA on? Then we can have more variance explained by fewer variables. The reason for all this trickery is that mvn (R) will not consider the 300 dimensional observations to be multivariate normally distributed, it will though significantly contribute most variable to be univarite normal. hence we use central limit theorem and use means. Means = normal

When testing for influence of experiments on the curves we choose a generalized version of the two-sampled Student's t-test namely the t-squared-test ref: "Hotellinger". Thus we propose a null-hypothesis that the mean of the different trajectories are the same. The t-squared test generalizes to multiple dimensions making it a perfect statistics for our 300 dimensional means. With the t-squared test we can handle independent multivariate normal distributions by calculating the mean and standard deviation of the one 100 repetitions in each experiment and finally compare all possible sets of two experiments. This statistics deserves its own explanation which we briefly provide in the appendix.

Evaluating clusters

We use Rand index to evaluate the clustering of our different methods. Specifically we will compare our true labels with the labels given by a model. We define two measures from our clustering:

$$S = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_{ij} \tag{1}$$

$$D = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D_{ij} \tag{2}$$

The notation requires some explanation. We diffine two cluster Z (true) and Q (predicted). $S_{ij} = 1$ if and only if Z and Q agrees that the pair of observation x_i, x_j belong to the same cluster, otherwise $S_{ij} = 0$. Similar $D_{ij} = 1$ only if Z and Q agrees that the pair of observation x_i, x_j doesn't belong to the same cluster. We then calculate the rand index like:

$$R(Q, P) = \frac{S + D}{\frac{1}{2}N(N-1)} \tag{3}$$

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 1.

Table 1: Confidence interval of classifiers

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

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

(4)

6 Appendix

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