LINMA2472 – Algorithms in Data Science

Homework 2 - PART 2 - Graph Kernel

Groupe 40

21/10/2022

Pierre Remacle
Pierre Merveille
Salma Ismail

Teatchers :
Jean-Charles Delvenne
Rémi Delogne
Estelle Massart
Bastien Massion
Gautier Krings
Brieuc Pinon

2022-2023

Contents

1 Random Fourier Features

1.1 Introduction

Random Fourier Features (RFF) is a widely used, simple, and effective technique for scaling up kernel methods, we'll be exploring more of its efficiency and usefulness. In this part of our assignment, we will be training three different classifiers on the «MNIST» dataset, in order to observe the differences of speed gains when evaluating new instances and when using RFFs.We'll be working with 10.000 training instances and 60.000 testing instances.

1.2 Testing Time: Exploring the Differences between Various SVMs

Aiming for an exploratory understanding of the different SVM methods, we're presenting the testing time consumed of each in table 1 below.

Table 1: Time Taken per Method

Linear SVM	Kernel SVM	Linear SVM with RFF
124.1	254.9	58.8

At first glance, we can observe the variances between different implementation, bringing up a question of distinctness. As observed, the RFF algorithm is successful in reducing the complexity of the process as compared to the time required to predict the outcomes of a linear SVM, and consequently, the time taken. Those results go in line with the concept of theoretical reduction of complexity covered in class.

Moreover, We may also infer that the kernel SVM's prediction time was sufficiently longer than the others. Given that this type of SVM is more complex than a linear one, this makes sense.

1.3 Varying Parameters: An understanding of the influence of parameter D

Furthermore, it is also of great interest to explore the gravity of the impact of adjusting the parameter D. We begin by looking at the summary table that is provided below.

Table 2: Time Taken per Method

D-Parameter Value	Creation Time *in seconds*	Training Time *in seconds*	Prediction Time *in seconds*	Accuracy
10	18.3	5	35.7	0.42
20	10.8	2.9	30.8	0.55
40	42.6	5.6	27.4	0.7
80	50.7	6	57	0.83
160	288.8	7.3	46.9	0.88
320	491.5	19.3	122.7	0.92
640	529.4	17.4	160.7	0.94
1000	725.5	22.3	177.2	0.94

The highlighted rows demonstrate how accuracy does not rise by a margin that would justify the processing expense when we increase the D-parameter value past a certain point. For instance, if we look at the last two D-values, creation time goes up by 0.37, but accuracy virtually stays the same.

This can be seen in a better, more visually attractive way if we look at the scatterplot below (Figure 7), where the dots stand in for various d-values. We can see that while there is a significant variation in the x-coefficient between the two red dots (the final two values for the D parameter), there is hardly any difference in the y-coefficient. There is only a very slight change in accuracy levels, even when noting the space between the D-value: 320 and the last two.

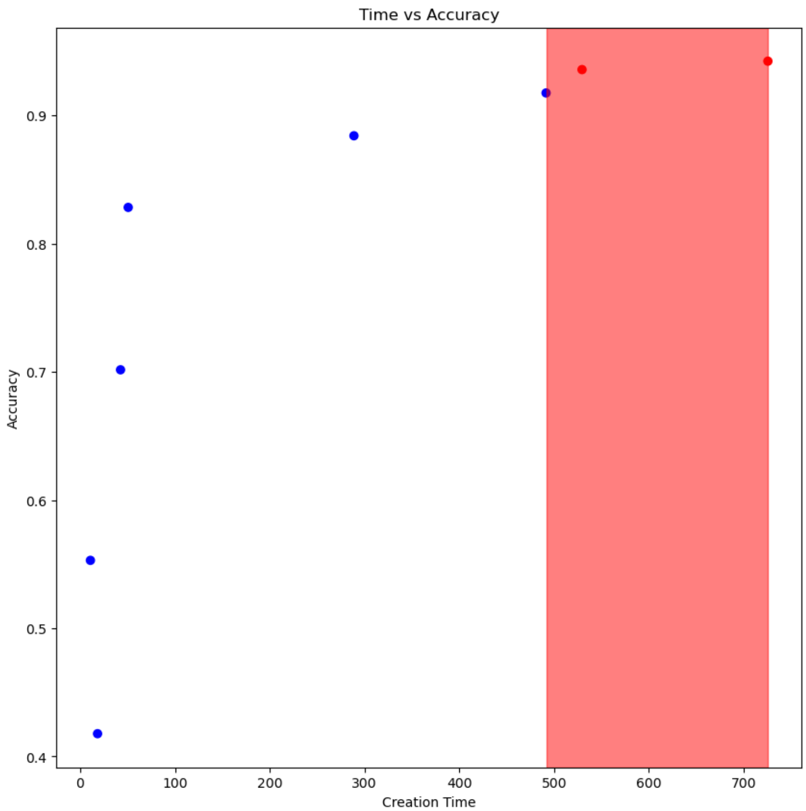

Figure 1: Time vs Accuracy

On the same front, we have also added some extra analysis. charting several perspectives and vantage points to assist us arrive at a logical conclusion.

As can be seen in the 'Accuracy vs D-Value' plot below, accuracy rises as D rises, but on a log scale.

In other words, increasing D significantly won't be helpful because accuracy will only increase somewhat before complexity begins to skyrocket. The figure depicting 'RFF Creation Time' serves as strong verification for this.

Another finding from the analysis is that the time needed to build the RFF increases more quickly than it does for the others metrics, and for high D's, this time virtually makes up the entire procedure. The time takes the form of an exponential function based on D. The accuracy takes the form of a logarithmic function based on the same D. To select D, We have to make a compromise between time and accuracy.

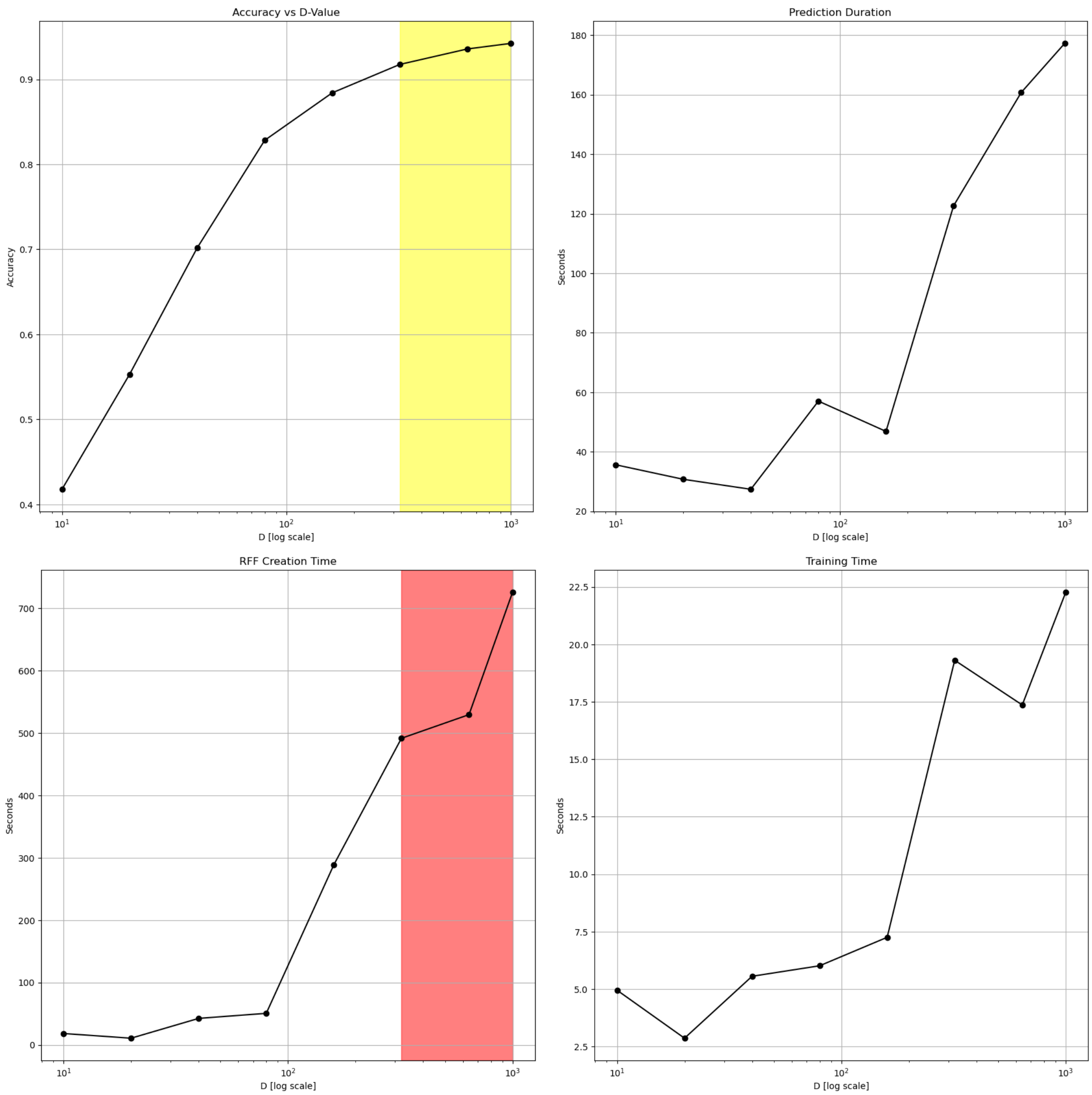

Figure 2: D-Parameter Analysis (Logarithmic Scale)

1.4 Conclusion

In the field of data science, high dimensional data will always constitute a significant problem. Large datasets would not be feasible even for powerful computers without the ongoing development of algorithms and techniques to increase effectiveness and reduce complexity. We were able to improve our classificational strategy thanks to the investigation of Random Fourier Features, which also demonstrated the advantages of such an explicit mapping over the true kernel. However, parameter value must be taken into account in order to ensure effectiveness and the best outcomes.