# UNIVERSITAT POLITÈCNICA DE CATALUNYA

## FACULTAT D'INFORMÀTICA DE BARCELONA

GEP — Deliverable 1

*GEP tutor:* Andujar Larios

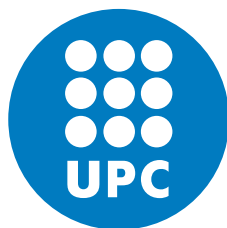# Analysis of the SVM-RFE algorithm for feature selection

*Author:*
Robert PLANAS

*Director:*
Luis A. BELANCHE

Bachelor Degree in Informatics Engineering
Specialization: Computing

March 2, 2021

# Contents

# Chapter 1

# Context and scope

## 1.1 Context

In statistics, machine-learning, data-mining, and other related disciplines, it is often the case that there is redundant or irrelevant data in a dataset[1]. Indeed, before we can start working with the data, some form of data analysis and cleaning is required. Data cleaning may include removing duplicated rows or rows with missing values, removing observations that are clearly out-layers, removing irrelevant variables (e.g. name, surname, email address), etc.

With the new era of Big Data, datasets have increased in size, both in number of observations and in dimensions. Applying classical data-mining and machine-learning algorithms to this high-dimensional data rises multiple issues collectively known as "the curse of dimensionality". One such issue is the elevated, usually intractable, cost and memory requirements derived from the non-linear (on number of dimensions) complexities of the algorithms. Another issue has to do with data in a high-dimensional space becoming sparse and negatively affecting the performance of algorithms designed to work in a low-dimensional space. And finally, a third issue is that with a high number of dimensions the algorithms tend to overfit, that is, they don't generalize enough and end up producing models that perform worse with real data that their predicted performance with the training data.

Simple manual data cleaning is not enough to achieve satisfactory amounts of dimensionality reduction. In this case we can use automatic techniques. We can classify such techniques in two categories, feature extraction, and feature selection.

### 1.1.1 Feature Extraction

Feature extraction techniques transform the original high-dimensional space into a new low-dimensional space by extracting or deriving information from the original features. The premise is to compress the data in order to pack the same information at the expense of model explainability[2]. Continuing with our data compression analogy, virtually all feature extraction techniques perform lossy compression, that is, some data is lost which makes the process irreversible. Notice that, if the process was reversible then feature extraction would not decrement explainability.

Some well known feature extraction algorithms include Principal Component Analysis (PCA) and auto-encoders, the first being a linear transformation over the feature-space and the second a neuronal network. PCA may be extended with a kernel method in order to make non-linear transformations. Similarly, we will also use kernels methods to extend SVM-RFE.

---

[1]A table with rows / records / observations and columns / variables / features / dimensions / predictors / attributes.

[2]The ability to explain why certain predictions are made. Also, interpretability.

### 1.1.2   Feature Selection

In contrast, feature selection only selects a subset of the existing features, ideally the most relevant or useful. This often implies a greater loss of data (which may result in lower predictability) but it doesn't reduce explainability. Knowing which features are relevant is often desired, but for some problems it is, not only desired, it's the whole point. In domains such as genetic analysis and text mining feature selection is not necessarily used to build predictors. For example in microarray analysis feature ranking is used to identify genes (i.e. features) that discriminate between healthy and disease patients (Guyon and Elisseeff, 2003).

*How are feature selection algorithms classified?*
*Where does SVM-RFE stand in?*

## 1.2   State of the art

Some some some.

## 1.3   Objective

Explore methods to improve on the existing SVM-RFE algorithm. Such methods can be divided in either methods that improve on the SVM utilization or methods that improve on the recursive feature elimination strategy.

For the first category we will explore ways to use non-linear kernels in order to improve the candidates, in particular the use of RFK kernel.

On the second category well explore options such as using results from previous iterations, and using subsets of data for optimization purposes.

## 1.4   Methodology

Some some some

## 1.5   Scope

# Bibliography

Guyon, Isabelle and André Elisseeff (2003). "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3.Mar, pp. 1157–1182. ISSN: 1533-7928.