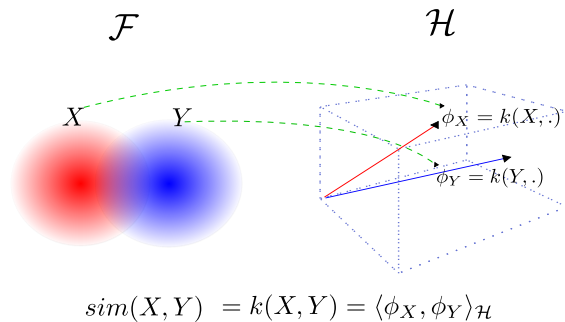


Scikit Fuzzy Kernel - A Toolkit for Kernels on Fuzzy Sets for Data mining and Machine Learning Applications TCC Proposal

Jorge Guevara¹ and Roberto Hirata Jr.²

¹IBM Research

²Department of Computer Science, University of Sao Paulo



1 Introduction

Modeling interval data (each instance is an interval of values), or uncertain data, or imprecise data, is usually done by Fuzzy sets. There are some libraries that help modeling and transforming Fuzzy sets written in Python but there is still no library that implemented Kernel methods for Fuzzy sets. In this work, that will be supervised by Dr. Jorge Guevara and Roberto Hirata Jr, the student will study, implement and document a library for data analysis based on kernel methods on Fuzzy sets. The resulting library will be used in an international tutorial to be presented in July. Therefore, the project will need a concentration of efforts in the first semester and less in the second semester, as usual.

2 Objectives

Construct a tool for data analysis based on kernels on Fuzzy sets. This tool will help researchers and data scientists for investigating new ideas and for dealing with Fuzzy datasets in data mining and Machine Learning tasks. This tool will be comprised by:

- a collection of similarity measures based on kernel on Fuzzy sets;
- a set of procedures for implementing data fuzzification, i.e., generation of Fuzzy datasets from usual datasets;
- integration with existing Kernel machines (from scikit-learn for example): i.e. Support Vector Machines, Kernel PCA, Gaussian process, etc.
- a set of practical examples (Jupyter notebooks) and documentation.

3 Success Criteria

Implement a workflow capable of:

- read a dataset and perform basic statistical analysis
- perform fuzzyfication of a dataset
- construct the Kernel similarity matrix given by Kernels on Fuzzy sets
- perform Machine Learning experiments (predictive modeling) on a Fuzzy dataset
- report those experiments

4 Why is this exciting?

- Create a tool for research on non-conventional data analysis
- Learn and apply the theory behind kernels and Reproducing Kernel Hilbert Spaces in Machine Learning Applications
- Opportunity of presenting this tool in tutorials and conferences of the area.

5 Cronogram

1. Fuzzy sets 101 - study and understand the basic concepts on Fuzzy sets.
2. Kernel methods 1010 - study and understand the basic concepts on Kernel methods.

3. Analysis and understanding of a Scikit tool. (<https://www.scipy.org/scikits.html>)
4. Analysis of Scikit-fuzzy (<https://github.com/scikit-fuzzy/scikit-fuzzy>) for using/reusing of its functionalities inside the aimed library.
5. Implementation of the library.
6. Documentation of the library.
7. Creation of examples of use and practical exercises.
8. Monograph writing
9. Poster writing
10. Poster presentation