Using Random Fourier Features with Random Forests

Deliverable 5: Justification of the competences and specialization adequacy

Albert Ribes Marzá

April 9, 2018

1. **Description of the objectives of the project**

   Random Forest is a popular machine learning algorithm which can be used both for classification and regression problems. Among others, it has de advantages that it is very fast to train and that it does not over-fit on the data. Together with Neural Networks and Support Vector Machines, the Random Forest is one of the best algorithms used in the state-of-the-art problems.

   However, this method is based on the vote of many trees, and as any committee, it is good on the members to disagree and arrive to a consensus. If everybody answered the same, the committee wouldn't be needed, just one person. In Random Forest, this disagreement usually is achieved by randomizing the part of the dataset that each tree is able to "see" in each step of the learning process. This way, they have different "points of view" and together can arrive to a more general answer.

   In this project I will try to formalize, implement and test a different method to achieve this discrepancy among the trees of the forest. It is based on the idea of performing a pseudo-random mapping on the data using the Fourier Features. This mapping has a linear complexity, so it will not affect to much the speed of the hole algorithm, and by the other hand it could help to increase the accuracy of the predictions taken.

   There are many approaches which could be taken using this idea, and the results are not guaranteed to be successful. This project will tackle this issues.

2. **Summary of the scope of the project**

   The whole project will be focused on the idea of mixing Random Fourier Features with the algorithm of Random Forest, in particular, with classification problems.

   I will study three possible approaches to this mix. The first one, which is the simplest, consists on simply generating one mapping from the original dataset to a different feature space, and then use this new data with the orignal Random Forest algorithm without any modification.

   The second approach is to generate a different mapping for each of the trees of the forest and then separately train the tree, each one using the orignal tree building algorithm. It is also possible for the tree building algorithm to need a modification to fit with the new data.

   The final approach considered in this project consists on performing the pseudo-random mapping in each of the nodes of the tress. It seems clear that a new formulation of the algoritm will be needed for this approach.

   Among the three options, the first one is the fastest and is not expected to effectively increase the accuracy of the algorithm. The second one seems to be the most realistic option, and the last one looks like a very hardcore method.

   In this project, I will study thes three approaches, implement them and finally perform test for each of them in order to see how they behave.

3. **What knowledge of the subjects from the specialization do you thing will be useful to develop the project. Justify it**

This project is fully focused in the field of Machine Learning, and as such all the knowledge which has anything to do with this subject will be useful in this project.

From the subject of *Algorithmics* (A), all the concepts about computing complexity provide a good introduction for understanding why a method performs better than another one. It is also useful the ability to distinguish between the different types of problems (P, NP, etc.) and the typical methods used to solve (or at least try) them.

Although the themes of *Artificial Intelligence* (IA) don't have to much to do with this project, they supply a notion of the complexity of the kind of problems we have to solve nowadays, and show that many alternatives exists to address a specific subject.

But a huge pat of the project will need the knowledge acquired in the subject of *Machine Learning* (APA). From the vey basics about Probability and Statistics to the formulation of the Random Forest algorithm, crossing themes like differentiation about Linear and Non-linear methods, optimizations and a big list of algorithms used in the field; all of this is very useful knowledge to develop the project.

4. **Justify why the suggested project suits the characteristics of the computing specialization**

   As Machine Learning is the core point of this project, it requires knowledge about complex computing problems, considering themes like efficiency and stability.

   A big background of algorithmics is needed, since you need to be able to formalize a complex problem, analyze it, know the ways to deal with it and be able to evaluate the solution proposed.

   In this process, it is important to have concern for the fiability of the solution proposed, and for the efficiency reached. Knowing the best programming language to implement the methods proposed is also needed to get the results as fast as possible.

   All this skills, which are required for this project, are the ones which are developed in the specialization of Computing, and this is why this project suits in it.

5. **List the technical competences of the project and the level of achievement chosen. Justification on why have each one been chosen and how is it planned to reach that level of achievement**

   - **CCO1.1**: Avaluar la complexitat computacional d'un problema, conèixer estratègies algorísmiques que puguin dur a la seva resolució, i recomanar, desenvolupar i implementar la que garanteixi el millor rendiment d'acord amb els requisits establerts. [A little bit]

     In the field of machine learning the computational complexity is an important factor to take into account. The best predictive method is useless if we can't get the answer in the required time.

The main objective of this project is to be able to tackle problems that cannot be solved precisely because a faster method is needed.

During the workflow of the project the complexity of the original and the implemented algorithm will be taken into account when taking a decision.

- **CCO2.1**: Demostrar coneixement dels fonaments, dels paradigmes i de les tècniques pròpies dels sistemes intel·ligents, i analitzar, dissenyar i construir sistemes, serveis i aplicacions informàtiques que utilitzin aquestes tècniques en qualsevol àmbit d'aplicació. [Enough]

  The whole project is based on being able to build and intelligent system, from a theoretical point of view. We will implement an algorithm which has to be the main building block for many services and computing applications.

- **CCO2.2**: Capacitat per a adquirir, obtenir, formalitzar i representar el coneixement humà d'una forma computable per a la resolució de problemes mitjançant un sistema informàtic en qualsevol àmbit d'aplicació, particularment en els que estan relacionats amb aspectes de computació, percepció i actuació en ambients o entorns intel·ligents. [A little bit]

  It's real problems what we are trying to solve, and as such they are presented in a human way. In order to predict a good answer it is required a good mapping from the knowledge owned by humans to something that machines can understand.

  In this project our algorithm needs to be fed with a dataset which represents human knowledge, it has to process it and finally produce and answer understandable by humans.

- **CCO2.4**: Demostrar coneixement i desenvolupar tècniques d'aprenentatge computacional; dissenyar i implementar aplicacions i sistemes que les utilitzin, incloent les que es dediquen a l'extracció automàtica d'informació i coneixement a partir de grans volums de dades. [In depth]

  I'm trying to improve the Random Forest algorithm, which is a machine learning technique. This improvement includes the design and implementation of the new features of the algorithm.

- **CCO3.1**: Implementar codi crític seguint criteris de temps d'execució, eficiència i seguretat. [A little bit]

  Part of the project consists on the implementation of part of the algorithm, and efficiency is a main point in the whole process.