Using Random Fourier Features with Random Forest

Albert Ribes

Director: Lluís A. Belanche Muñoz

Computer Science

Grau en Enginyeria Informàtica

Computació

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) – BarcelonaTech

Fecha de defensa

Contents

1 Ideas generales

El guión que me propuso LLuís es:

1. Problema que ataco
2. Por qué es importante
3. Qué propongo en mi TFG
4. Estado del arte en el problema que ataco
5. Nociones generales del tema
    - Machine Learning
    - Árboles
    - Logit
    - RFF
    - Nystroem
    - Bootstrap
    - Boosting

6. El trabajo propiamente dicho (explicar lo que voy a hacer)
7. Experimentos
8. Conclusiones y Trabajo futuro
9. Referencias
10. Apéndices

2 Introduction

2.1 Problem to solve

Todavía no se consigue suficiente precisión con el Machine Learning

2.2 Why it is important to solve this problem

Con precisión más alta se podría aplicar el machine learning en otros campos

2.3 Project proposal

Incrementar el accuracy que se puede conseguir con algunos problemas mezclando la técnica del bagging (y quizá del boosting) con el tema los RFF

Actualmente el bagging solo se usa con Decision Tree porque es muy inestable. Con lo que propongo aquí, podría ser factible usarlo con otros algoritmos más estables

3 Background

Más o menos, cada uno debería ocupar entre media y una página

3.1 Machine Learning

3.2 Classification and Regression

3.3 Review de los principales modelos que existen

3.3.1 Decision Tree

3.3.2 Logistic Regression

3.3.3 SVM

3.4 Las técnicas ensembling

3.4.1 Bagging

- Inventado por Leo Breiman
- Pretende reducir el sesgo

3.4.2 Boosting

- Adaboost (adaptive boosting)
- El siguiente estimador es más probable que contenga los elementos no no se han predicho bien en el anterior
- Se trata de ir modificando los pesos que tiene cada una de las instancias
- El entrenamiento de los modelos es secuencial, a diferencia del bagging
- Enterarme de quien lo inventó, y para qué ámbitos es útil

3.5 El bootstrap

- En bagging es bueno que los estimadores estén poco relacionados entre ellos
- Idealmente, usaríamos un dataset distinto para cada uno de los estimadores, pero eso no siempre es posible
- Una alternativa es usar un resampling con repetición sobre cada uno de los estimadores para tener datasets un poco distintos entre ellos.
- Enterarme de la cantidad de elementos distintos que se espera que queden en el subconjunto, y quizá hablar de la cantidad de aleatoriedad

3.6 Las funciones kernel

3.7 Las Random Fourier Features

3.8 Nystroem

3.9 PCA

3.10 Cross-validation

4 Workflow of the project

4.1 La idea general un poco desarroyada

4.2 State of the art con las RFF

5 Experimental results

6 Conclusion

De momento, parece que algunos problemas sí que se benefician de esto, mientas que otros no lo hacen

	Project development	**Exploitation**	**Risks**
Environmental	Consumption design	Ecological Footprint	Environmental risks
Economic	Project bill	Viability plan	Economic risks
Social	Personal impact	Social impact	Social risks

7 Future work

- El trabajo se ha centrado en problemas de clasificación, pero no hay ningún motivo para que no se pueda aplicar el regresión. Se ha omitodo por simplificar
- Aquella teoría de que quizá se puede regular la cantidad de aleatoriedad que añade el bootstrap, y quizá inventar un bootstrap con un parámetro para regular la cantidad de aleatoriedad
- Pensar en aquella teoría de que quizá se puede inventar un procedimiento para, dato un problema determinado con sus datos, sacar un número que sea representativo de la cantidad promedio de ruido que tiene. Puesto que quizá es útil para este proyecto conocer la cantidad de aleatoriedad que tienen los datos, para que se pueda regular

8 Sustainability Report

Consumo del equipo CO2 Materiales Peligrosos Materiales que vienen de zonas de conflicto Qué sabemos de nuestros proveedores, y si la fabricación de la maquinaria se ha hecho en una instalación sin riesgo, ni para las personas ni para la naturaleza

El impacto que tiene el proyecto directa e indirectamente en la gente Ha sido diseñado pensando en cómo se ha de reciclar y reparar, siguiendo conceptos de economía circular?

8.1 Environmental

El impacto medio-ambiental de mi TFG ¿Cuál es el coste medio-ambiental de los productos TIC? ¿Cuantos recursos son necesarios para fabricar un dispositivo? ¿Cuánto consumen estos dispositivos durante su vida útil? ¿Cuantos resuduos se generan para fabricarlos? ¿Qué hacemos de los dispositivos una vez ha terminado su vida útil? ¿Los tiramos y contaminamos el medio de nuestro entorno? ¿Los enviamos al tercer mundo, y contaminamos el suyo?

La huella ecológica se puede medir, por ejemplo, en Kilovatios hora y en emisiones de CO2

La huella ecológica que tendrá el proyecto durante toda su vida útil ¿Mi TFG contribuirá a reducir el consumo energético y la generación de resudios? ¿O los incrementará?

8.2 Economic

Costos de portarlo a terme i assegurar la seva pervivència

Estimar los recursos materiales y humanos necesarios para la realización del proyecto. Sería como preparar la factura que se le pasará al cliente, teniendo en cuenta que se terminará en un término bien definido

Una planificación temporal y Una explicación de si he pensado algún proceso para reducir costes

Estimar las desviaciones que he tenido respecto a las planificaciones iniciales.

Para evaluarlo durante su vida útil, he de hacer un pequeño estudio de mercado, y analizar en qué se diferencia de los ya existentes, si se mejora en algún aspecto, o no. Reflexionar sobre si mi producto tendrá mercado o no, y sobretodo, explicarlo

Además de estudiar los costes, plantearse si es posible reducirlos de alguna manera

Plantear posibles escenarios que por razones de limitaciones en el tiempo y de recursos no tendré en cuenta, pero que podrían perjudicar la viabilidad económica del proyecto

Suele salir a unos 20000 € el TFG

8.3 Social

Implicaciones sociales, tanto sobre el colectivo al que se dirige el proyecto como sobre otros colectivos

8.3.1 Impacto personal

En qué me ha afectado a mí, y a mi entorno cercano, la realización de este proyecto. En qué me ha cambiado la vida, o si ha cambiado mi visión sobre el tema. ¿Ha hecho que me de cuenta de situaciones que antes ignoraba?

8.3.2 Impacto social

Implicaciones que la realización de este proyecto puede tener sobre la sociedad. Hay que identificar al colectivo de los afectados. Los colectivos pueden ser: los propietarios, los gestores del proyecto, los trabajadores, los proveedores, los consumidores (usuarios directos), o terceros (usuarios indirectos o pasivos) Puedo consultar los estándares del GRI

8.3.3 Riesgos sociales

Explicar posibles escenarios probables, pero no significativos, que no puedo abordar por falta de tiempo o de recursos o de capacidad, y que podrían perjudicar a personas relacionadas con mi proyecto