



Treball final de grau

GRAU D'ENGINYERIA
INFORMÀTICA

Facultat de Matemàtiques
Universitat de Barcelona

AQUI EL TÍTOL DEL TREBALL

Autor: Xavier Moreno Liceras

Director: Laura Igual

Realitzat a: Departament

Matemàtica Aplicada y Anàlisis

Barcelona, May 29, 2015

Abstract

Goldbach's weak conjecture asserts that every odd integer greater than 5 is the sum of three primes. We study that problem and the proof of it presented by H. A. Helfgott and D. Platt. We focus on the circle method. Finally, we describe a computation that confirms Goldbach's weak conjecture up to 10^{28} .

Resum

Normalment un tutor tutoritza a un conjunt d'alumnes i no dona temps de mirar detingudament alumne per alumne, això fa que possiblement no es realitzin les accions corresponents per a un alumne.

Agraïments

Vull agrair a ...

Contents

1	Introducció	1
2	Planificació	2
2.1	Problemes de planificació	2
I	Coneixement de l'àrea	3
3	Eines	3
3.1	Suport	3
3.1.1	GitHub	3
3.1.2	Bitbucket	3
3.1.3	Trello	3
3.2	Eines de programació	4
3.2.1	Python	4
3.2.2	Pandas	4
3.2.3	NumPy	4
3.2.4	Scikit-learn	4
3.2.5	Bokeh	4
3.2.6	Seaborn	5
3.3	Editors	5
3.3.1	IPython notebook	5
3.3.2	Texmaker	5
4	Tècniques utilitzades	6
4.1	Clusterització (Agrupacions)	6
4.1.1	K-means	6
4.1.2	MeanShift	6
4.1.3	DBScan	6
4.1.4	Mètriques utilitzades	6
4.2	Predicció	7
4.2.1	Regressors	7
4.2.2	Recomanador	7
4.2.3	Mètriques utilitzades	7
4.3	Reducció de dimensions	7
4.3.1	PCA	7
II	Procés pràctic	8

5	Etapas del projecte	8
5.1	Hipòtesis: preguntes plantejades	8
5.2	Data cleaning de les primeres dades	8
5.3	Data cleaning de les noves dades	8
5.4	Procés de resolució de preguntes	8
6	Preguntes	9
6.1	Hi ha perfils d'estudiants basats en les notes?	9
6.2	Un dels clusters correspon als alumnes que abandonen? . . .	9
6.3	Quin perfil d'entrada té cada cluster?	9
6.4	Ranking de dificultat de les assignatures que ha de fer (que ha matriculat) un alumne	9
7	Treball futur	10
8	Conclusions	11
9	Bibliografia	12

1 Introducció

2 Planificació

2.1 Problemes de planificació

Part I

Coneixement de l'àrea

3 Eines

3.1 Suport

Aquestes són les eines de suport que m'han ajudat al llarg del treball per tal de fer més còmode la seva organització tant personal com per equip.

3.1.1 GitHub

GitHub és una plataforma online per desenvolupar projectes software de forma col·laborativa. Aquesta plataforma utilitza un control de versions anomenat Git. La finalitat de GitHub és l'emmagatzement massiu de projectes amb codi font obert. Per això hem optat per la utilització de GitHub, ja que que volem que el nostre codi el pogui veure tothom i que qualsevol que el necessiti per fer la seva investigació, el pogui utilitzar.

3.1.2 Bitbucket

Bitbucket és una plataforma semblant a GitHub, però amb el servei d'un altre control de versions com Mercurial a més de Git. Bitbucket té l'advantage de permetre crear repositoris privats de forma gratuïta. Aquesta plataforma va bé per a l'inici d'un projecte on es fan molts canvis en el codi, ja que pots tenir el codi en privat, i un cop el codi ja agafa forma es pot migrar a GitHub. Això és el que hem fet nosaltres en el projecte, començar amb Bitbucket i després passar-nos a GitHub amb el codi font obert.

3.1.3 Trello

Per últim com eina de suport, hem fet servir Trello, una plataforma online que permet una comunicació més clara entre els membres d'un projecte. Amb Trello pots crear projectes i cada projecte conté un conjunt de llistes que s'omplen de tasques. Nosaltres hem fet servir Trello, per comunicar-nos amb la tutora i tenir present una planificació per tal d'organitzar-nos millor.

3.2 Eines de programació

En aquesta secció trobarem amb el llenguatge de programació i conjunt de llibreries que hem treballat.

3.2.1 Python

Python és un llenguatge d'alt nivell interpretat. Remarquen molt la fàcil lectura del seus codis, per això té una sintaxis molt semblant a un pseudocodi. Python és un llenguatge de codi obert i desenvolupat per *Python Software Foundation*, una organització sense ànim de lucre. Vam escollir Python en el seu moment per dues simples raons; per ser un llenguatge de scripting i per la seves llibreries relacionades amb el tractament de dades (com [Pandas](#), [NumPy](#) o [Scikit-learn](#)).

3.2.2 Pandas

Pandas és una biblioteca informàtica escrita en python per a la manipulació i anàlisi de dades. Especialment va bé per al tractament de taules alhora de fer consultes, o per a l'agrupació i agregació d'informació.

3.2.3 NumPy

Numpy és una biblioteca informàtica de Python per operar amb vectors i matrius d'una forma més extensa a la que et permet el mateix llenguatge Python, la qual conté tot un conjunt de funcions matemàtiques d'alt nivell per treballar amb aquests vectors i matrius.

3.2.4 Scikit-learn

Scikit-learn és una biblioteca informàtica orientada a l'aprenentatge automàtic per a Python. Té suport per classificadors, regressors i clustering. Per aquest projecte hem fet servir clustering i regressors. En la secció de [Tècniques utilitzades](#) es detalla cada tècnica utilitzada d'aquesta biblioteca informàtica.

3.2.5 Bokeh

Bokeh és una biblioteca informàtica per a la visualització interactiva de dades dirigida als navegadors per a la seva presentació a través d'HTML i JavaScript. Bokeh té el suport per a gràfiques específiques com diagrames de barra, box plots o time series, però a banda d'aquests gràfics pots dibuixar sobre un gràfic amb elements bàsics com cercles, línies, rectangles, entre altres.

3.2.6 Seaborn

Per últim tenim Seaborn que també és una biblioteca informàtica per a visualització de dades com Bokeh, amb gràfiques molt més específiques. A més té una part de la biblioteca informàtica dedicada a les paletes de colors i la qual permet escollir un conjunt de colors afavorits per mostrar les dades.

```
%matplotlib inline
import seaborn as sns
palette = sns.color_palette("hls", 5)
sns.palplot(palette)
```



Figure 1: Elecció d'una paleta de 5 colors

3.3 Editors

3.3.1 IPython notebook

3.3.2 Texmaker

4 Tècniques utilitzades

4.1 Clusterització (Agrupacions)

4.1.1 K-means

4.1.2 MeanShift

4.1.3 DBScan

4.1.4 Mètriques utilitzades

4.2 Predicció

4.2.1 Regressors

4.2.2 Recomanador

4.2.3 Mètriques utilitzades

4.3 Reducció de dimensions

4.3.1 PCA

Part II

Procés pràctic

5 Etapes del projecte

5.1 Hipòtesis: preguntes plantejades

5.2 Data cleaning de les primeres dades

5.3 Data cleaning de les noves dades

5.4 Procés de resolució de preguntes

6 Preguntes

- 6.1 Hi ha perfils d'estudiants basats en les notes?
- 6.2 Un dels clusters correspon als alumnes que abandonen?
- 6.3 Quin perfil d'entrada té cada cluster?
- 6.4 Ranking de dificultat de les assignatures que ha de fer (que ha matriculat) un alumne

7 Treball futur

8 Conclusions

9 Bibliografia

<i>Algoritme\Mètriques</i>	MAE	MSE	PCC
Recomanador col·laboratiu	1.231	2.997	0.335
Recomanador basat en contingut	1.197	2.905	0.403
Random Forest Regressor	1.134	2.584	0.490
Linear Regressor	1.175	2.720	0.462

Table 1: Dades no normalitzades

<i>Mètriques/Algoritme</i>	MAE	MSE	PCC
Recomanador col·laboratiu	0.558	0.669	0.069
Recomanador basat en contingut	0.531	0.660	0.358
Random Forest Regressor	0.509	0.565	0.393
Linear Regressor	0.538	0.648	0.462

Table 2: Dades normalitzades