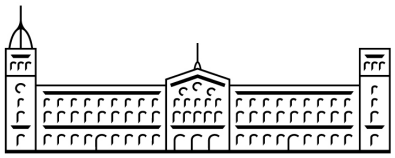

Treball final de grau

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 31, 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 dóna 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

2 Descripció del problema

2.1 Explicar dades

change title

2.2 Ciència de les dades

La ciència de les dades és el conjunt d'etapes per tal d'arribar a un resultat, en forma de coneixement, a partir d'un conjunt de dades. Aquesta aplica un conjunt de tècniques de diferents àreas, ara com matemàtiques, estadística, teoria de la informació o tecnologia de l'extracció d'informació.

Un projecte de ciència de les dades es separa en diverses etapes:

Preguntes Què és el que volem explorar? Té sentit el que ens estem plantejant?

Adquisició de les dades Com és la font d'obtenció de les dades? (Base de dades, *Web Scraping*)

Descripció Aquesta fase abasta tres processos

- **Neteja de dades** Com hem de netejar i separar les dades? (mostres atípiques, filtració, reducció de dimensions, normalització, extracció de característiques)
- **Agregació** Com hem de recolectar i resumir les dades? (promig, desviació estàndard, box plots)
- **Enriquiment** Com podem afegir més informació a les nostres dades? (Cerca a altres fonts de dades adicionals)

Desobriment Podem segmentar les nostres dades per trobar grups naturals i disgregats? (Clusterització, visualització)

2.3 Etapes del projecte

2.4 Plantejar preguntes

change title

3 Planificació

3.1 Tasques

3.2 Diagrama de Gantt

3.3 Evaluació económica

4 Desenvolupament del projecte

4.1 Eines

4.1.1 Eines de 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.

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'emmagatzenament 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.

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'advantatge 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.

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.

4.1.2 Eines de programació

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

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).

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 agrecació d'informació.

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.

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.

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.

Seaborn

Per últim tenim Seaborn que també és una biblioteca informàtica per a visualitazció 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 sbn
palette = sbn.color_palette("hls", 5)
sbn.palplot(palette)
```

Figure 1: Elecció d'una paleta de 5 colors

4.1.3 Eines d'edició

IPython notebook

Texmaker

4.2 Tècniques utilitzades

4.2.1 Clusterització (Agrupacions)

K-means

MeanShift

Mètriques utilitzades

4.2.2 Predicció

Recomanador

Random Forest Regressor

Regressor lineal

Mètriques utilitzades

4.2.3 Reducció de dimensions

PCA

5 Experiments i resultats

Algoritme\Mètriques	**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

Mètriques/Algoritme	**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

6 Conclusions i treball futur

7 Bibliografia