San Francisco Crime Classification
Kaggle competition

Łukasz Rados, Wojciech Kusa

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza w Krakowie

24 stycznia 2016

1 Wprowadzenie

Celem projektu było stworzenie oprogramowania pozwalającego dokonać klasyfikacji przestępstw na podstawie danych czasoprzestrzennych z raportów policyjnych dla miasta San Francisco w Stanach Zjednoczonych.

2 Dane

Zbiór danych zawiera incydenty zgłoszone policji w San Francisco pomiędzy 01.01.2003r. a 13.05.2015r.. Podzielony jest na dwie podgrupy (prawie równoliczne, w każdej po około 850 tysięcy elementów) :

- zbiór treningowy - zawierający zgłoszenia z tygodni parzystych,
- zbiór testowy - zawierający zgłoszenia z tygodni nieparzystych.

Przykładowe wiersze danych treningowych znajdują się na Rysunku **??**. Dane składają się z następujących pól:

- Dates - znacznik czasu przestępstwa
- DayOfWeek - dzień tygodnia
- PdDistrict - nazwa departamentu policji odbierającego zgłoszenie
- Address - przybliżony adres przestępstwa
- X - długość geograficzna
- Y - szerokość geograficzna

- Category - kategoria przestępstwa (tylko dla zbioru treingowego). Zmienna, którą należało przewidzieć
- Descript - szczegółowy opis przestępstwa (tylko dla zbioru treingowego)
- Resolution - jaki był wynik działania policji (tylko dla zbioru treingowego)

2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

Rysunek 1: Przykładowe dane treningowe. Źródło: `https://www.kaggle.com/c/sf-crime/data`

2.1 Wstępna analiza zbioru treningowego

tutaj rysunki

2.2 Zastosowane deskryptory

3 Zastosowane algorytmy

3.1 Lasy losowe - Random Forest

3.2 Generalized Linear Model

4 Implementacja

kilka słów o użytych bibliotekach