Kaggle-Challenge: San Francisco Crime Classification

Raphael Emberger

January 5, 2019

Date: January 5, 2019
Instructor: Professor Yukawa

Contents

1 Preface **2**

2 Abstract **3**

3 Introduction **4**
3.1 Initial situation 4
3.2 Objective 4

4 Theoretical Principles **5**
4.1 Loss Function 5

5 Methods **6**

6 Results **7**

7 Conclusion **8**

8 Listings **9**

A Appendix **10**

1 Preface

Firstly, I want to express my gratitude to Professor Yukawa for guiding me in this project and to the Kokusaika staff members to arrange my stay here at the Nagaoka University of Technology(subsequently referred to as "NUT"). I was given the generous opportunity to study at the NUT for one semester, for which I am very grateful. During that time I could choose from the following six Kaggle challenges to work on as project work:

- Toxic Comment Classification Challenge (Kaggle 2018*c*)
- TalkingData AdTracking Fraud Detection Challenge (Kaggle 2018*b*)
- Quora Question Pairs (Kaggle 2016*a*)
- Expedia Hotel Recommendations (Kaggle 2017)
- San Francisco Crime Classification (Kaggle 2016*b*)
- Inclusive Images Challenge (Kaggle 2018*a*)

Of those, I was most interested in the classification of reported crimes (Kaggle 2016*b*), as in my opinion this was an interesting challenge, given the dataset to be only consisting of spatial and time data. As such, this report is dedicated to take on this challenge.

2 Abstract

3 Introduction

3.1 Initial situation

The challenge has been out since roughly 3 years and since then, many teams have participated and submitted their results. This lead the leader-board to fill up with 2335 submissions which were ranked and their results displayed online(see "Leaderboard" at Kaggle (2016*b*)). The results vary from 34.53877 up to 1.95936(see 4.1 for the ranking principle).

3.2 Objective

> From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.
>
> Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.
>
> From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.
>
> We're also encouraging you to explore the dataset visually. What can we learn about the city through visualizations like this Top Crimes Map? The top most up-voted scripts from this competition will receive official Kaggle swag as prizes.

(Kaggle 2016*b*)

4 Theoretical Principles

4.1 Loss Function

The ranking of the results on the Kaggle leader board are based on the multi-class logarithmic loss function:

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}\right) \tag{I}$$

N : Number of cases in dataset.

M : Number of classes.

y_{ij} : Label for class. 1 if i is in j. Otherwise 0.

p_{ij} : Predicted probability that i belongs to j.

This basically boils down to a format as follows:

Class 1	Class 2	Class 3
0.24	0.48	0.38

With the labels being:

Class 1	Class 2	Class 3
0.00	1.00	0.00

When those values are applied to I, we get a value of 0.49548. Of course, the closer the prediction is to the actual labels, the smaller the loss value will be.

To calculate examples quickly on the python console, the following code can be used:

```
import numpy as np
from sklearn.metrics import log_loss
labels = np.array([0.0, 1.0, 0.0])
prediction = np.array([0.04, 0.78, 0.18])
print(log_loss(labels, prediction))
```

5 Methods

6 Results

7 Conclusion

8 Listings

References

Kaggle (2016*a*), 'Quora Question Pairs'.
URL: *https://www.kaggle.com/c/quora-question-pairs* 2

Kaggle (2016*b*), 'San Francisco Crime Classification'.
URL: *https://www.kaggle.com/c/sf-crime* 2, 4

Kaggle (2017), 'Expedia Hotel Recommendations'.
URL: *https://www.kaggle.com/c/expedia-hotel-recommendations* 2

Kaggle (2018*a*), 'Inclusive Images Challenge'.
URL: *https://www.kaggle.com/c/inclusive-images-challenge* 2

Kaggle (2018*b*), 'TalkingData AdTracking Fraud Detection Challenge'.
URL: *https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection* 2

Kaggle (2018*c*), 'Toxic Comment Classification Challenge'.
URL: *https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge* 2

A Appendix