Machine Learning - Exercise 0: Dataset description

A, B, C

Chosen data sets

The first data set we chose is titled *"Crimes in Los Angeles from 2010"*. It is a transcription from original crime reports and contains relatively few missing attributes while at the same time the total number of samples is quite large. Also, being government-issued, we expect the data to be quite precise; allowing for interesting insights into distribution of reported crime by year as well as area.

The other set contains data collected analysing a single player's experience with the competitive ladder of the video game *"Overwatch"*. It contains fewer entries and often times - because of the chosen data categories - a lot of missing values. What made this data set especially interesting for us is the fact that the creator included their psychological condition during each match they played.

Additionally, we chose these two sets together to also contrast a "professionally" created table with one created by a layperson.

Characteristics of data set

Crimes in Los Angeles from 2010

The data are comprised of 1692056 samples with 26 attributes. They contain mostly nominal values, e.g. the area where a crime occurred, the description of the crime or the victim's sex. The important interval quantities are date and time of a crime being reported, the victim's age marking the single significant ratio.

Looking at the distribution of total reported crime per year, we notice a considerable incline starting at the year 2015.

Also, interestingly, a majority of crimes occurs at noon, as can be seen in the following histogram:

In terms of victim characteristics, slightly more males than females are reported as victims. The data also show that most victims' descent is Hispanic, followed by White and Black.

Finally, the following scatter plot shows the rough location of the reports as coordinates:

Overwatch competitive ladder

This data set contains 3299 samples with 47 features. Many of the features are ratio quantities, e.g. the amount of healing done during a match, the number of deaths or weapon accuracy. A lot of nominal quantities are also of significance, most notably if a match resulted in a win, a loss or a draw, as well as which map was played or character used.

In terms of numerical values, a lot of preprocessing was done to convert them from their original string format. Over the five seasons recorded (seasons 3 - 7) the player did an average amount of healing of 5852.7 per match, with a maximum of 19226. The player deaths yield an average of 9825.5 which is unrealistically high. It is most likely that this data is either mislabelled or incorrect. Lastly, the player's accuracy sits at 9.45% on average with a maximum of 59%.

Looking at match results, we can see that there are slightly more lost games than victories, draws are significantly rarer.

Further, there is a number of maps that are notably less voted for than others. These include *"Junkertown"*, *"Oasis"* and *"Nepal"*.

Looking at heroes played we see that the player mostly chooses support roles. This corresponds nicely with the relatively high average healing output noted earlier.

Lastly and most interestingly, a look at the player's psychological condition reveals that most of the time they felt indifferent/neutral and in that state of mind their win/loss ratio is close to 1. "Good" and "great" conditions tend to have much more wins than losses. "Bad" and "tilted" conditions appear to have more losses than wins, suggesting that a player's negative mood might adversely affect their gameplay.