Increase the Efficiency of Capturing a Pokemon: Predict Pokemon's Showing up Based on Historical Data

Wei Li

Project type: application oriented
Dataset: Predict'em All

1 DATASET DESCRIPTION

Predict'em All is a dataset containing roughly 293,000 historical Pokemon sighting (more detail below). There are 206 raw features, each describe a condition where the sighting happens, like time, weather, location, whether the sighting co-occur with other Pokemons, whether the place is close to water, etc. There are 151 types of Pokemons in the game, so the target to be predict is Pokemon's identifier, PokemonId, a numeric feature ranges between 1 and 151.

2 PROJECT DESCRIPTION

Pokemon Go is an augmented reality mobile game where users can catch Pokemons they encountered in real-world and train them to fight with other people's. Users may be interested in the type of Pokemon they will encounter in the future as they want to capture specific types. Assume that given time and environment, all Pokemons have a possibility to show up, if we know the possibility, we know where to find them. A real-life environment like weather, location, time may be factors that affect the possibility. The dataset contains 293,000 historical sightings, including the environment and the Pokemons' id. The objective of this project is to train a classifier that generalizes well to output the possibility of each class's showing up. With the classifier, subsequent works can be done to build a system that:

- predicts what kind of Pokemon has the highest chance to appear nearby (says 500 m).
- predicts when and where a type of Pokemon will appear in the future (says 30 mins).
- visualizing the result.

A wide range of models will be considered in the project. The performance will be compared using different metrics. Besides, some extent of data modification and cleaning need to be done given the quality of the dataset.

Finally, since the algorithm to generate Pokemons may not be consistent, the hidden distribution may be messy, which may limit the accuracy of our best model.

3 STATEMENT

I hereby certify I will not use these dataset(s) and algorithm(s) in any other courses, past, present or future. I understand that submitting overlapping material for more than one course is not allowed. This work is also not part of my thesis or capstone project research. I understand that any such action on my part will lead to a grade of INC being awarded for CSI5155 and may lead to academic sanctions imposed by the faculty of Engineering.