Taylor Huntington
Machine Learning Project Paper
11/22/2016

**1) Introduction and Motivation**

Pokemon Go is this year's highest grossing mobile game  released this Summer by Niantic Labs (a Google subsidiary). Termed "immersive reality", the game relies on users moving throughout areas and catching virtual "Pokemon" that appear on the screen.

I frequently play this game walking to and from class. As a computer science major, I naturally looked into the online developer community for the game. Pokemon Go turned out to be much less random than it appeared. Using an application from a github programming team (Christopher 2016), I was able to create and use bot accounts to scan areas I would otherwise be unaware of. As I used the application, I began to notice trends and patterns in spawns.

**2) Problem definition**

Not all Pokemon spawns are created equal. Pokemon, by definition, vary in name, shape, combat power, and other attributes. Therefore, certain species of Pokemon are more desirable to acquire than others. Knowing this disparity, it becomes valuable to know spawn points and patterns. Scanning revealed two main quantitative structures, spawn points and nests.

Spawn points are the methodology on which Pokemon spawn. Throughout the scan area we are able to identify various location points that spawn Pokemon. 'Spawning' Pokemon occurs at an exact minute of every hour the spawn point is active and produces a catchable specimen for 15 minutes. Spawn points are "active" for 15  of every 60 hours, or about 25% of the time.

Nests consist of a collection of spawn points in specific areas. These spawn points are extremely likely to spawn a specific species of Pokemon. The scanning program revealed that Audubon Park's spawns were unlike other areas on the map.

**3) Solution?**

Seeing Audubon Park's trends and patterns of spawn points I wanted to find a quantifiable way of defining a nest. Whenever I would visit Audubon Park I would notice a strange density of Squirtles. In an immersive reality game, it would make sense for a turtle to spawn by the water, but there was such a unique density to it. In my attempt to define a nest, I knew there were squirtles, but specifically I wanted to know how many squirtles and when over a period of time.

This disparity in spawns are only notable inside the Audubon Park area. Fifteen steps off of the green section of your map pointed to spawns returning to normal. To prove the existence of a nest at Audubon Park, I would compare data from spawn locations inside Audubon Park versus data from spawn locations outside of Audubon Park.

**4) Data?**

To collect data for this experiment, I scanned 1600m in every direction from McAlister stadium for a week. For simplicities sake, I decided to focus upon 10 spawn points. Every spawn point has two characteristics; a longitude latitude location, and a 15 hour spawn timer based on an hour and minute combination specific to each spawn point. This combination operates on a 60 hour timer that resets back to 00:00 at 12:00AM Sunday every week.

The first five of these spawn points are within Audubon Park. The sixth spawn point sits on the LBC quad. The seventh is located near our very own Stanley Thomas. Our eighth point is contained at Devlin Fieldhouse. Our ninth point greets local patrons at Bruno's. Lastly our tenth point is located over by Felipe's.

Data reveals that three of the five spawn points will be active on any given day. Spawns frequently passing over multiple days due to the 15 hour active duration. All spawn points are in consistent with a 15 hour active duration followed by a 45 hour sleep duration. These amounts never changed throughout the experiment.

**5) Machine learning design**

Due to the nature of the 60 hour timer, the weekly midnight Sunday reset to 00:00, and making exact Pokemon spawn/despawn times, we are able to accurately predict spawn points. I created a learning program to predict Pokemon spawns for our ten spawn points.

To help read the data, I created a PokeCsvReader converts data from a .csv file(PokeInput) into ArrayLists of a custom class called Pokemon. Each Pokemon stored in these arraylists have a ID that corresponds to their species (1=Squirtle, 2=Bulbasaur, 3=Eevee, 4=Pidgey, 5=Other), an integer value that stores when in the spawn cycle they spawned, and a classifier of whether that

Pokemon is in Audubon Park(sourced from spawn points 0-5) or not from Audubon Park(sourced from spawn points 6-10).

For our seven days of data, I was able to create a learning program that can predict future spawn locations and times across different days. For our training and testing, I compared our days of spawn data against each other. The LogisticRegressionModel class represents a Java representation of logistic regression. The constructor contains a double, rate, to represent the rate of change in training test data. The weights array in the constructor represents our eventual output. The weights array in our experiment contains two doubles, one to represent the number of squirtles as a ratio of all spawns in Audubon Park, and outside of Audubon Park. The intent of this model is to show that Audubon Park operates as a squirtle nest and spawn points inside will have significantly higher ratios than outside the park.

**6) Machine learning algorithm**

The user designates the specific day of the training data and has our LogisticRegressionModel instance learn from that day's spawn data through the train command. The model then produces predictions about that data through the predict function. The user then has the data we will test our trained model against designated in the testList variable. By using the arrayClassify command, the model produces two separate ratios output meant to represent the predicted ratio of squirtles inside and outside of Audubon park for the set of data. These two ratios are then compared against the actual ratios of squirtles inside and outside of audubon park and presented to the user. [0] represents the ratio of Audubon squirtle spawns compared to all other areas in Audubon and [1] the ratio of Non-Audubon squirtles against every other spawn. Next, the difference of these two ratios is then calculated and presented as output to the user as a single variable that shows the correlation of Audubon Park to squirtle spawns as compared to all other areas.

**7) Your Results**

By comparing spawn diversity and times across different days, patterns began to emerge. Because Sunday resets the 60 hour spawn clock, there is a schedule every week that spawn points hold to. Due to the patterning of 24 hour days into a 60 hour cycle, 2.8 iterations of the 60 hour cycle pass per week. This means that the last 12 spawns of a 60 hour cycle are not spawned three times in a week, but rather only two, regardless of location. This makes spawn points that do not occur in the 48-60 hour segment of the cycle inherently more valuable.

The data showed high correlation between Audubon Park spawn similarities across days, but less so for spawn points outside the park. In addition, data shows that Tuesday, Wednesday, and Thursday all have unique spawning patterns and times. Data from these three days produce similar spawn diversity when compared to test overall data from different days, but do not share spawn timings or patterns with any other day of the week.

This means that there exists two pairs of spawn cycles, Friday paired with Sunday and Saturday paired with Monday.. Friday and Sunday have matching spawn times and locations and therefore produce identical data. This means that using Friday to train our PokeLearner will always produce predictions that perfectly match Sunday's testing data. The same can be said if we train with Sunday's data to test against Friday's data. This property also exists for the pairing of Saturday and Monday.

Our model predicted a Squirtle correlation to Audubon Park of about .75 or 75%. In other words, our model predicted a ratio of about 3 Squirtles spawns inside of Audubon Park for every Squirtle spawn outside of it across an equal number of spawn points representing inside and outside of Audubon park.

**8) Related work, are there other results?**

The next logical step would be to find another nest and run a similar test. City Park is widely known as a nest, and a similar train and test to predict spawn times, locations, and diversity would produce similar, repeatable results. If these results are confirmed, we can reproduce these ratios for different species dependent on the nest. If City Park is a Bulbasaur nest, we could perform a similar test to mine checking for the ratio of Bulbasaur spawns inside and outside of City Park.

**9) Conclusion**

By using patterns in our data, we were able to perfectly predict a spawn's time, place, and species. Since Friday/Sunday and Saturday/Monday share spawn schedules, we are able to train data from one and test it against another with perfect accuracy. If we use data from other, unconnected days, we able to get a good guess at the spawn diversity and their locations, but unable to get a good grasp on predicting when those spawns will occur.

Our Logistic Regression Model produces a prediction output that three squirtles would spawn inside Audubon Park for every Squirtle outside of it, assuming an equal number of spawn points. This clustering of a specific species proves our intention of quantifying a nest.

Sources:

Christopher, M. (2016, August). Mchristopher/PokemonGo-DesktopMap. Retrieved December 05, 2016, from https://github.com/mchristopher/PokemonGo-DesktopMap/releases