Machine Learning Challenge

Prepared for 7Next by Joseph Rivas

Introduction

The purpose of this assignment is to develop a machine learning classifier for predicting a given Pokemon's type. For this assignment, a support vector machine (SVM) model was used to provide a performance baseline for the classifier.

Dataset & Preparation

The dataset used for this assignment was obtained from Kaggle [1]. This dataset was inspected in order to select the relevant features and match them to the information that can be retrieved from the public PokeAPI. *NOTE:* This dataset only includes Pokemon up to the 7th generation. Pokemon from the newest 8th generation are not available in this dataset; as such, predictions for an input for a Pokemon from the 8th generation will result in an error.

Exploratory Data Analysis

The dataset was first explored in order to determine if there were any class imbalances in the dataset. The results for the type class labels are shown below in Figure 1. It can be observed that the class labels are not equally distributed. To prevent bias in the model training, training data will be split and shuffled.

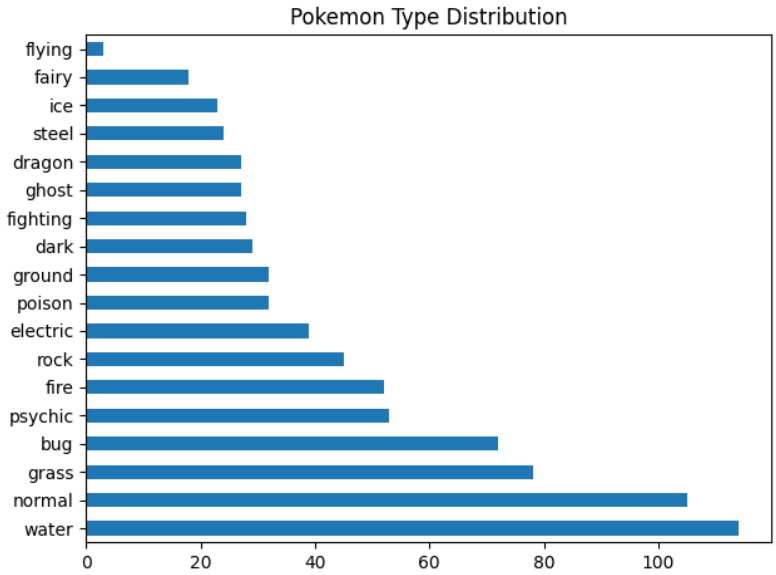

Figure 1. Pokemon type distribution

Next, a feature correlation heatmap was created in order to determine which features may be correlated with each other. The correlation heatmap is shown below in Figure 2. It can be observed that height and weight are correlated, as can be expected. Though typically one feature may be dropped if correlated with another, for the purpose of this assignment they will both be kept as features for the classifier model.

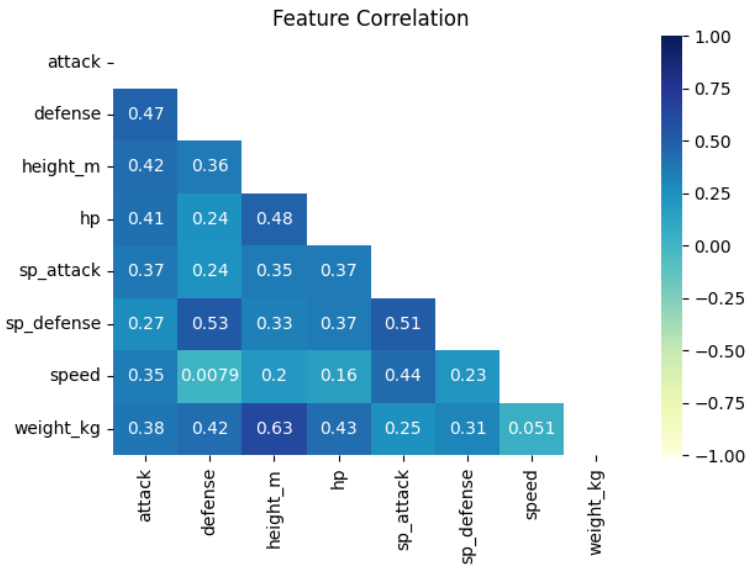

Figure 2. Feature correlation heatmap

Model Output & Results

The SVM model was split into train/test groups with 70% of the data being sent to the training group. The default radial basis function (RBF) was used as the kernel. Given the limited data and additional parsing, the model's performance clearly does not produce production-level performance. The values for a cross-validation score with 5 folds were determined to be: [0.15454545 0.13761468 0.12844037 0.14678899 0.13761468]. Given the imbalanced nature of the dataset, scikit-learn's "f1_macro" scoring was also used, with resulting values of: [0.02445394 0.02560554 0.02655943 0.02565032 0.01461988]. It can be seen that the F1 scores are significantly lower than the accuracy scores. Finally, a classification report of the data is shown below in Table 1.

	precision	recall	f1-score	support
grass	0	0	0	22
fire	0	0	0	9
water	0	0	0	8
bug	0	0	0	12
normal	0	0	0	5
poison	0	0	0	9
electric	0	0	0	15
fairy	0	0	0	1
fighting	0	0	0	8
psychic	0	0	0	23
ghost	0	0	0	8
rock	0	0	0	7
ground	0	0	0	30
ice	0	0	0	9
dragon	0	0	0	16
dark	0.125	0.166667	0.142857	12
steel	0	0	0	7

flying	0.152074	0.970588	0.262948	34
accuracy	0.148936	0.148936	0.148936	0.148936
macro avg	0.015393	0.063181	0.022545	235
weighted avg	0.028385	0.148936	0.045338	235

Table 1. Classification report

Future Work & Discussion

Given the time constraints of this challenge, it was not possible to implement several more classifier models. Much of the time was dedicated to the initial inspection and analysis of the dataset and to creating functions for cleaning the data input. In the future, perhaps a custom neural network classifier could be implemented in Tensorflow/Pytorch in order to compare its performance to the SVM.

Since this dataset was limited in the amount of input values for the classifier, additional data could have been used to improve the classifier. In this dataset, 20 rows were removed due to missing NaN values for heights and weights of certain Pokemon, decreasing the size of an already limited dataset. As mentioned previously, this dataset also did not include the newest generation of Pokemon. Future work could involve tying in the PokeAPI to this dataset and augmenting them missing values. Additional values could also be added to the dataset. For instance, the type of moves that the Pokemon learns would be a useful indicator for predicting its type – it is unlikely that a water Pokemon would learn fire-type moves and this would give the classifier more information for its training. Another set of information would be the Pokemon's abilities. From the dataset, it appears that some types may have specific ability attributes that other types are unlikely to have. For example, an electric-type Pokemon would not likely have the "chlorophyll" ability, which is more closely associated with grass-type Pokemon. To incorporate this information, these text values would likely have to be encoded to integer values using a method such as one-hot encoding. This would lead to an increase in the space of the dataset and create a sparse matrix of values as the Pokemon would have types associated with certain abilities and moves. A neural network classifier could be trained with this sparse dataset, though its performance would need to be evaluated against this SVM baseline model.

References

1. https://www.kaggle.com/rounakbanik/pokemon