Machine Learning Midterm1

Andrea Bacciu

November 2019

1 Dataset

The dataset used for midterm 1 has unbalanced classes for the number of legendary pokémon in the minority compared to the non-legendary. In the dataset, there is a lot of typo or error (e.g. a value of Legendary column was false instead of FALSE), for example we can find the 'Total' column not sum up the other statistics columns (Attack, HP ecc). In the attached .CSV there is the correction of all errors I found in it. In the .ipynb there is print() of the Dataset statistics request in Part 1. For each test, I have set the random state to zero in order to avoid the continuous variation into train e test set during the development.

2 Framework & tools

The whole code is written in Python 3.7 using the Anaconda Environment. The Machine Learning framework used is Sklearn and for parsing the .CSV I used the well know Pandas library.

3 Feature Selection

I have used a different set of features for each classifier.
The features used with DT are [Total, HP, Attack, SpecialAtk, SpecialDef].
For the classification with MLP, I used this set of features [HP, Attack, SpecialAtk, SpecialDef, Speed].
More attributes were removed because they are not a feature like Number (ref. to the pokédex number, a number used as Pokemon ID) and the name is completely arbitrary. The remaining features are the best set I found after several tests. I tested the Type1 and Type2 features turning them from strings to categorical features but they don't improve the predictive power of both classifiers.

4 Classification & Cross Validation

The classifiers used are DecisionTreeClassifier and MLP. In Part 4 I show the performance of both, without tuning the hyperparameters, then in Part 5, I

show the best hyperparameters found by the Grid Search CV algorithm using the Cross-Validation with 5 folds. The use of CV allow to detect the overfitting problem.

5 Final Result

At the end we can conclude that, the predicting power of Decision Tree Classifer is more stronger than the Multi Layer Perceptron. In both tests the Decision Tree reached F1-Score more higher than MLP.
The result shown in tables are produced by the not tuned classifier and train belong the test set (from part 4). I suppose the MLP need more data.

Table 1: Decision Tree

Class	Precision	Recall	f1-score	support
not Legendary	1.00	0.99	0.99	174
Legendary	0.89	1.00	0.94	16

Table 2: Multi Layer Perceptron

Class	Precision	Recall	f1-score	support
not Legendary	0.94	0.98	0.96	174
Legendary	0.62	0.31	0.42	16