Machine Learning Challenge

Accenture

Introduction

This is a small series of exercises to evaluate your knowledge in machine learning. Please respond the questions detailing the steps you took to solve each task. All questions are simple, but this is your chance to show us your technical knowledge, so do not hold out on your math skills.

We also expect to receive your code with your solution and analysis, including any pre-processing steps. Feel free to use the programming language with which you are the most comfortable, but please provide us with instructions to run your code. This goes without saying, but you do not need to code the algorithms from scratch. You can use any library available to you, such as *scikit-learn*. Remember that the organization and readability of your code will also be evaluated.

We want this challenge (and hopefully your job with us) to be fun, so we selected a super-heroes dataset for you to play with! You can download the data from Kaggle at https://www.kaggle.com/claudiodavi/superhero-set/data. It comprises two csv files, super_hero_powers.csv and heroes_information.csv, and you will use both of them.

Finally, we designed this challenge to take at most three hours. Obviously, we are not timing you, but you should not take much longer than that to finish this exercise. Mind that we are not looking for the best performing models, as it is not a competition and we will mainly evaluate your approach to solving these tasks. Hence, invest your time in detailing and explaining your solutions rather than trying to fit the best model. Also, note that we are not defining any data splits or evaluation metrics. These are entirely up to you and will also be evaluated.

If you have any problem with the exercise, do not hesitate to contact us, but we are not going to comment on your solutions or elaborate on the technical details of the questions.

Clustering

Question 1

First, we want to cluster our superheroes according to both their powers and information. Run an unsupervised clustering method using the number of clusters that you judge the most appropriate.

1. Which algorithm did you pick and why?
2. Which features did you use and why? Please explain any pre-processing or feature engineering (selection) you have performed.

Question 2

One of the challenges in clustering is defining the right number of clusters. How did you choose that number? How do you evaluate the quality of the final clusters?

Spotting the Bad Guys

In this section, we will deal with the supervised learning problem. More concretely, we will formulate a classification task, and our target is the super-heroes alignment (good or bad).

Question 3

First, we will use the Naive Bayes algorithm. Run the algorithm on the superheroes data to predict the alignment variable and evaluate the results. Again, please detail any pre-processing and feature engineering you applied in the process.

1. Which hypotheses do we assume when using the Naive Bayes algorithm?
2. How do the specific characteristics of this dataset influence your modeling choices and results?
3. How do you evaluate the results?

Question 4

Now feel free to run the classification algorithm that you judge the most appropriate for this task.

1. What motivated your choice of algorithm?
2. How does this algorithm compare with the Naive Bayes regarding modeling assumptions and results?

Beyond Good and Evil

Let's turn our problem into a regression task and try to predict the super-heroes' weight given the other features.

1. Which algorithm did you pick and why?
2. How do you evaluate the performance of your algorithm in this case?

Analysis

Which aspects of this dataset pose problems for clustering, classification, and regression? How did you solve these issues?

Bonus

If you enjoyed playing with the super-heroes dataset, this section is for you to showcase any further aspects of the data we have not explored in the questions. As a bonus section, this is totally optional, but we would love to see the insights you can get from this data.