DSBA 2020: M1 Assignment

Machine Learning with Pokémon

Description

Back to the teens. Your first assignment will be on Pokemon data. No data munging needed. Just old-school ML.

Data

The data is available through the URL:
https://sds-aau.github.io/SDS-master/00_data/pokemon.csv

It contains data on 800 Pokemon from the 1st to the 6th generation.

Tasks

You will have to perform a series of standard tasks in unsupervised as well as supervised machine learning.

EDA & Unsupervised ML

1. Give a brief overview of data, what variables are there, how are the variables scaled and variation of the data columns.
2. Execute a PCA analysis on all numerical variables in the dataset. Hint: Don't forget to scale them first. Use 4 components. What is the individuel and cumulative explained variance?
3. Use a different dimensionality reduction method (eg. UMAP/NMF) – do the findings differ?
4. Perform a cluster analysis (KMeans) on all numerical variables (scaled & before PCA). Pick a realistic number of clusters (up to you where the large clusters remain mostly stable).
5. Visualize the first 2 principal components and color the datapoints by cluster.
6. Inspect the distribution of the variable “Type1” across clusters. Does the algorithm separate the different types of pokemon?
7. Perform a cluster analysis on all numerical variables scaled and AFTER dimensionality reduction and visualize the first 2 principal components.
8. Again, inspect the distribution of the variable “Type 1” across clusters, does it differ from the distribution before dimensionality reduction?

Supervised ML

Your task will be to predict the variable “legendary”, indicating if the pokemon is a legendary one or not.

1. Perform necessary ML preprocessing of your data if deemed necessary.
2. Split the data in a training (75%) and test (25%) dataset.
3. Define a n-fold cross-validation workflow for your model testing.
4. Fit three separate models on your training data, where you predict the “legendary” variable. Use a 1. Logistic regression, 2. Decision tree, and 3. (minimum) on adittional SML algorithm of choice to do so.
5. Use the fitted models to predict the “legendary” variable in your test data.
6. Evaluate the performance of these 3 models by comparing the predicted and the true values of “legendary” in the test data. To do so, also create a confusion matrix, provide and discuss further useful metrics of model performance.

Hand-in

- Hand in on eksamen.cbs.dk
- Deadline: 5th Oct. at 12:00
- Format
    - Python: PDF of notebook + original ipynb (zipped)
    - R: Python: Html (self-contained) of notebook + original rmd (zipped)