STA2201 Final Project Proposal

Justin Zhang, Isaac Baguisa, Alex Faassen

2025-03-09

As kids growing up in the late 2000's, early 2010's, we witnessed the rise in popularity (in the Western world at least) of anime, trading card games, and the Nintendo DS. Though nowadays it's considered ancient technology, back then every kid was immersed in their Gameboys and DS's, playing games such as Super Mario Bros, Zelda, and various versions of Pokémon. Now a decade down the road, Pokémon has maintained its popularity through various TV shows, Pokémon GO, and most recently a trading card app. We thought it would be interesting to explore the world of Pokémon through a statistical lens.

The Kaggle datasets we have chosen to work with are a set of Pokemon images and a Pokemon stats dataset. Our analysis will centre on the ability to predict a Pokemon's type based on images and statistics. More formally, we aim to answer the question: can clustering algorithms predict a Pokemon's type based on its visual, physical, and fighting attributes?

The Pokemon images are a folder of 809 PNG files, each file representing a Pokemon. Each file consists of a 3-dimensional array of dimensions 120 by 120 by 4. The first 2 slices represent a 120 by 120 grid of pixels, and the 3rd slice represents the RGBA colour assignment to each pixel. Each array will be turned into a (120x120x4 = 57600) element vector, each pixel and RGBA value representing a feature. The Pokemon stats dataset consists of 801 rows representing Pokemon and 41 columns representing a variety of features including physical attributes (ex. height, weight), fighting attributes (ex. hit points, attack, defense) and other descriptive attributes (ex. legendary status and ability), each encoded as a real (decimal) number. We only consider the 801 Pokemon in both data sources. Note that there are no missing values in these datasets.

We will conduct the following data analysis:

1. Run PCA on the dataset to reduce dimensionality.
2. Use t-SNE and UMAP and compare with PCA to see if Pokemon types have non-linear clustering patterns
3. Run K-means on each dataset with k = number of types to determine if type can be recovered. Consider a weighted k-means when working with combined dataset to not overweigh the image data.
4. Run K-means on each dataset and tune for the optimal number of clusters.
5. Since we have the true labels, compute the accuracy of these methods and compare.
6. Run supervised learning algorithms - such as multinomial logistic regression, KNN, and LDA

Our final output will consist of:

1. An introduction to dataset – printing images data distributions of interesting features
2. PCA results – autoplots, scree plots, and visualization of compressed images
3. Clustering results – visualization (interesting pairs of features), CH plots
4. Classification results - determine which features contribute to predicting Pokemon type, compare classification accuracy and clustering results