Javier Pollak, Vanessa Amar, Elliot F. Poirier   McGill University
February 9th 2022   Applied Machine Learning

MiniProject 1: Getting Started with Machine Learning

Task 1: Acquire, preprocess, and analyze the data

Dataset 1. We chose to use numPy for the first dataset since the values were numerical and numPy seemed like a better tool. There were no duplicate rows in this dataset

Dataset 2. We made the choice of using pandas for the second dataset since this one used categorical classifications and Pandas seemed like a better tool for that. There were no rows with missing values or values that were not in the specified format (P,A or N; B or NB), so none were removed because of that. However, there were 147 duplicate rows that we chose to remove from the Dataset.

$$cost = (-1/m) * \sum_{i=1}^{m}[y_i * log(h_i) + (1 - y_i) * log(1 - h_i)]$$