Data Description

We use the presidential election dataset, which contains county demographics and voting outcomes for the last four presidential elections, from 2008 to 2020. It has 123 columns and 3090 counties. We will be using 115 features, and our target column will be whether the Democrats win in each county. The training set will consist of the data from the first three election years of the dataset with the top 50 features. The test set will be the latest year with the same features.

Problem Statement

We plan to predict the presidential election result at the county level. We will try to model the trend in the election years from 2008 to 2016 and see how accurate it is in 2020.

modeling approach

We will build a neural network with binary cross-entropy as the loss function. For the baseline model, we will use a multi-class logistic regression model.

Mathematical Formulation

For baseline/logistic regression $P(y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i X_i)}}$ Where X_i are the features, and y = 1 means democrats win.

For the NN - Model:

$$\text{(input space)} \xrightarrow{\text{linear}} \mathbb{R}^{h_1} \xrightarrow{\text{ReLU}} \mathbb{R}^{k} \xrightarrow{\text{Softmax}} \text{(output space)}$$

A simple linear regression is the linear layer at the beginning.

ReLU activation function means element-wise application of $\max(0, x)$.

The final transformation applies the softmax function S to obtain a probability distribution over the output space:

$$S(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where z_i are the logits from the previous layer.

Key performance indicators

We will make a confusion matrix to see how many counties votes for which party. We will use the F1-score to see the performance of the classification model.

MSE for the logistic model.