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 199 features and 3103 counties. We will not use all the features. I plan to use the top 50 features with the most combined votes over the four elections (none would be a linear combination of others). The logic behind it is that the features with the highest volume will dictate who wins the elections.

The target variable is whether the Democrats win. The training set will consist of the data of 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 overall outcome of the presidential election based on demographic data. 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

I will build a multilayer perceptron with a hidden layer and MSE as the loss function. For the baseline model, I 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 is one of the parties winning.

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 in 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

Accuracy—See if the model accurately guesses the correct winners. Compare per county for both models.

MSE for the perception model.