Experiment 3

Distribution-based loss functions

Instruction: Create a new .Rmd file and write code for each section in separate R code block

All machine learning (ML) models are based on minimizing a loss function. The parameters are learned such that the desired loss is minimized. One type of loss function, frequently used in ML, is a Distribution-based loss function. These loss functions are based on **likelihood of assumed data distribution**. Instead of minimizing arbitrary errors (like MSE), the model minimizes the negative log-likelihood i.e., it learns parameters that make the observed data most probable under a chosen distribution.

These losses embed distributional assumptions directly into training. Such models are more robust and interpretable. This technique is key bridge between statistics and machine learning.

1. Download "epil" Dataset. Evaluate mean and variance for the label ('y').

2. Define a Random Variable (X) as number of epileptic seizure ('y'). Draw the histogram and find the probability mass function (PMF) for X.

3. Fit a Poisson mass function to the PMF obtained in 2. Evaluate parameters.

4. Fit a Negative binomial mass function to the PMF obtained in 2. Evaluate Parameters. Based on your observations in 1,2,3, identify which model (Poisson/Negative Binomial) is more appropriate and why?

5. Develop a linear regression model to predict y as a function of features (f) (lbase+lage+trt). Evaluate MAE, RMSE and log likelihood values.

6. Develop a Generalized Linear Model (GLM) regression using poisson deviance as the loss to model y vs f in 5. Evaluate MAE, RMSE and log likelihood values.

7. Develop a Generalized Linear Model (GLM) regression using negative binomial likelihood as the loss to model y vs f in 5. Evaluate MAE, RMSE and log likelihood values. Which model (5,6,7) is best (best in what metric?)?

8. Exploratory: Perform a log transformation on y(z=ln(1+y)). Implement regression (z~lbase+lage+trt) using the methods in steps 5,6, and 7. Explain your observations.

9. Answer the questions in the shared report.