# Question

The training data has a large class imbalance as described in this table.

|  |  |  |  |
| --- | --- | --- | --- |
| Dir | Interictal Files=0 | Preictal Files=1 | E(Class==1) |
| Dog\_1 | 480 | 24 | 0.0476 |
| Dog\_2 | 500 | 42 | 0.0775 |
| Dog\_3 | 1440 | 72 | 0.0476 |
| Dog\_4 | 804 | 97 | 0.1077 |
| Dog\_5 | 450 | 30 | 0.0625 |
| Patient\_1 | 50 | 18 | 0.2647 |
| Patient\_2 | 42 | 18 | 0.3000 |

When training and validating the classifier, is it better to present training examples drawn from these classes in this ratio or to present an equal number of examples from the two classes?

# Accomplishments

Saved shuffled file name lists in each dataset subdirectory to serve as my standard for reproducible training / validation splits with any training/validation ratio and any subset size.

Used SetShuffle.ShuffleAll('C:\\Users\\Mark\\Documents\\GitHub\\IS-2014\\Datasets\\Kaggle Seizure Prediction Challenge\\Raw') to write ‘Shuffle.csv’ to each subdirectory Raw directory. Note that the filenames in the .csv have .mat extensions.

# Pseudocode

For each individual…

For each epoch, load a random subset of the training files

Generate GPU batch files for 16, 256, 4096, and 65536 samples.

Modify the RBM training program to pretrain an hierarchical autoencoder with constant depth per layer.

Compare its RMSE to a single layer network with the same hidden layer dimension.

