Product Classification 2

Submissions are evaluated using multi-class logarithmic loss. Each row in the dataset has been labeled with one true Class. For each row, you must submit the predicted probabilities that the product belongs to each class label. The formula is:

$$\text{log loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of rows in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if observation *i* is in class *j* and 0 otherwise, and p_{ij} is the predicted probability that observation *i* belongs to class *j*.

The submitted probabilities for a given product are not required to sum to one; they are rescaled prior to being scored, each row is divided by the row sum. In order to avoid the extremes of the log function, predicted probabilities are replaced with $\max(\min(p,1-10^{-15}),10^{-15})$.

Submission File

You must submit a csv file with the product id and the predicted probability that the product belongs to each of the classes seen in the dataset. The order of the rows does not matter. The file must have a header and should look like the following:

id,Class_1,Class_2,Class_3,Class_4,Class_5,Class_6,Class_7,Class_8,Class_9

200000,0.05,0.14,0.21,0.05,0.20,0.04,0.00,0.20,0.11

200001,0.21,0.06,0.10,0.20,0.13,0.01,0.04,0.10,0.15

200002,0.15,0.12,0.18,0.10,0.16,0.16,0.03,0.01,0.09

Etc.

Files

- train.csv - the training data, one product (id) per row, with the associated features (feature_*) and class label (target)
- test.csv - the test data; you must predict the probability the id belongs to each class
- sample_submission.csv - a sample submission file in the correct format