Brianna Shade

CS545: Machine Learning

1/23/14

Project 1: Perceptron Training Distinguishing Letters

OVERVIEW

This project explored training perceptrons to recognize handwritten alpha characters, specifically distinguishing **A**s from the other letters of the alphabet. A text file consisting of 16 feature values for several hundred instances of each letter was used to train and test the perceptrons, calculating final accuracy statistics.

IMPLEMENTATION

This project was coded in Java 1.7.0\_09, using Eclipse on a MacBook Pro running Mavericks. All data points for both the target letter and test letter are first extracted from the input file, alternatingly populating a designated training set and a test set (approximately half of the input data used for training and half for testing). The training data features are then scaled to between 0 and 1, and the sgn value is calculated for each instance, beginning with randomly generated weights and adjusting them stochastically after each training example. The entire training set is processed repeatedly until the change in weights converge to below a specified threshold. Finally, the test data is processed with the learned weights, calculating and outputting the perceptron’s accuracy, precision, recall, and confusion matrix. Data points for an ROC graph are also calculated, slicing the range between everything classified as positive and everything classified as negative according to a provided value.

I first ran some tests to determine what would be the optimal weight delta convergence threshold. I tested only **A** vs **B** to get a rough idea for thresholds, finding 0.0000001 to produce the best accuracy. However, using this threshold revealed many of the other letters’ weight changes never seemed to converge to below this value, so I also implemented a cap of 500 epochs. Adjusting this threshold, I found a balance between overall accuracy of all letters at about 0.000001.

OBSERVATIONS

With this threshold, most letters fall in the 98-99% accuracy range, with **H** and **O** closer to 95-97%. **A** tends to get confused the most with **H** and **O**. In particular, these letters were not commonly classified as **A**s, but rather **A**s were incorrectly classified as **H**s or **O**s. **H** would be understandable, as this is likely the most similar letter to **A** when written. However, **O** surprises me, as I wouldn’t suspect these to be easily confused. This could depend entirely on the style of a written **A**. If it is cursive or lower-case, it would, indeed, look very much like an **O**.

Letters such as **D**, **H**, **O**, **Q**, and **U** had the lowest recall values (highest false negative rates). Not surprising that these are similar in shape to **O** and **H**, with which the perceptrons most commonly confuse **A**. Conversely, **J**, **K**, and **S** have the lowest precision rates (highest number of false positives). These have similar qualities to **A**, depending on how they are written.

RESULTS

**Learning rate of 0.2**

**Learning rate of 0.8**

This higher learning rate resulted in poorer accuracy. Weights took longer to converge - in fact, many more letters maxed out the instituted epoch cap. Also accuracy values fell to 97-98%.

**Learning rate of 0.05**

The lower learning rate understandably initiated more epochs being executed. However, I felt this led to some perceptrons becoming over-trained, resulting in some accuracy values falling into the 84% range.

CONCLUSION