Assignment

December 1, 2002

1. Create two random, linearly separable data clouds in R^2 You can do this as follows (or do your own thing): let $f(x, y) = 2x + 5y - 2$ denote a line which passes through the unit square $[0, 1]^2$ (verify this). Then generate some random data (x, y), in $[0, 1]^2$ and assign the points to the two classes depending on the sign of $f(x, y)$.

   Use the perceptron algorithm to locate a separating line.

   Try this with a small set of data, so the line will be easy to locate (i.e., so the perceptron algorithm will have an easy job of it), and a larger set of data so the line will be harder to locate (i.e., the algorithm will take a lot longer).

   Report on the efficacy of the perceptron algorithm.

2. Use the data you have developed for digits' features (the subwindows' black-pixels sums) for two (or more digit categories), and test the separating ability of a hyperplane determined by the pseudoinverse. (By *test* I mean create the confusion matrix.)

3. Add to the 16-element feature set (of the previous problem) the products of pairs of the features, giving $16 + \binom{16}{2} = 136$ features.

   Apply and test the pseudoinverse method (as well as any other method you choose) to this larger, non-linear feature set.

4. Using the artificial data-cloud generators of the first problem set, create a set of data points in R^2 that is the union of two clouds (it should appear so visually). Implement and test the K-means clustering algorithm to attempt to separate the data into two sets (it will be a linear separation).

   Experiment with the union of more than two clouds in various positions.