Statistical Machine Learning
Problem set 5

Alan Joseph Bekker

K-means

- In the clustering problem, we are given a training set x_1,...,x_n , and want to group the data into a few cohesive "clusters." Here, we are given feature vectors for each data point $x_i \in R^d$ as usual, but no labels l_i (making this an unsupervised learning problem). Our goal is to predict k centroids and a label l_i for each datapoint. The k-means clustering algorithm is as follows:

  1. Initialize c_1,...c_k different centroids, were $c_i \in R^d$.
  2. for each data point t set $l_t = \min_j |x_t - c_j|^2$.
  3. For each j set: $c_j = \frac{\sum_{i=1}^{m}\{1|l_i=j\}x_i}{\sum_{i=1}^{m}\{1|l_i=j\}}$
  4. repeat until converge.

- In this assignment you are requested to use the MNIST data set and apply the $K = 10$ algorithm to the data.
- Verify at each iteration the cost function decreases and plot it at the end.
- Remember the cost function is $J = \sum_{i=1}^{k}\sum_{x\in c_i}|x - c_i|^2$
- What is the success rate achieved by this algorithm?