Aplicação de Clusterização para identificar dígitos

O dataset presente no github contém 60000 pares, o primeiro elemento de cada um deles sendo um vetor de 784 casas e o segundo um dígito de 0 a 9. O vetor como um todo representa uma matriz de 28X28 pixeis, cada uma de suas casas armazenando um valor entre 0 (branco) e 255 (preto). Cada matriz é um dígito que foi escrito a mão.

Dividimos o dataset em dois pedaços, uma parte para treino e outra para validação. A parte de treino será usada para construir as clusters, que serão usadas para separar os vetores do conjunto de validação em cada um dos 10 dígitos. Depois disso, seremos capazes de conferir quantos dígitos foram classificados corretamente.

Usaremos a clusterização k-means. Implementando o algoritmo de Lloyd, encontramos as clusters e suas respectivas centroides. Classificamos cada uma das clusters com o dígito mais frequentemente encontrado dentro delas, o que é possível já que o dataset nos informa o dígito correspondente a cada vetor.

Para cada vetor presente no dataset de validação, encontramos a centroide mais próxima dele, assim identificando a qual cluster ele pertence, classificando-o como o dígito daquela cluster.

Podemos, no final, fazer uma análise de eficiência dessa abordagem, vendo quantos dígitos não foram corretamente identificados.

Fonte: <http://johnloeber.com/docs/kmeans.html>