CS 145 Project: K-Nearest Neighbors

Surui Sun 104700648

November 26, 2016

Summary:
1. Preprocess the data, including assigning labels for classes and standardizing the observed value for margins, shapes and textures.
2. Using stratified shuffle split, using 80% data for every class as training data, and the remaining 20% as testing data. Iterating this stratified sampling for 10 times.
3. Calculated the average accuracy and log loss for given k (the number of neighbors, k ranges from 1 to 9) and found that the accuracy is highest when k=1 and p=1(which means we are using manhattan distance instead of traditional euclidian distance) and visualize the result.
4. Do the same method but only consider margin, shape or texture as input data matrix X. The result showed that the accuracy has reduced significantly and log loss is increased.
5. Combining all those observations and results, we finally chose k=1,p=1 with all attributes (margins, shapes, textures) to build the K Nearest Neighbors model. (Accuracy = 0.9803, log loss = 0.68)