Homework 01: KNN

Yawen Zhang

1. The relationship between the number of training examples and accuracy is shown in Fig. 1. As it shows, generally, the accuracy increases as the number of training examples increases. The maximum accuracy is 0.9727 when the number of training examples equals 50000. However, the increasing speed slows down as the number increases, especially when the number of training examples reaches 30000, the accuracy seems to be stable. More training data tends to provide more samples for each class with a better coverage of different features, which will improve the accuracy for classification.

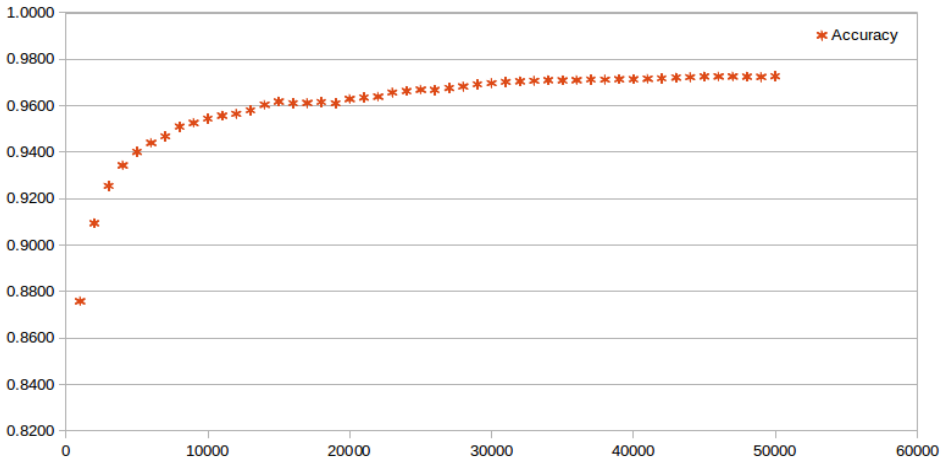

Fig. 1 The relationship between the number of training examples and accuracy (k = 3)

2. The relationship between k and accuracy is shown in Fig. 2. As it shows, generally, the accuracy decreases as k increases. A reason for this is that when the training set size is big enough, for example, 50000, it is easy for a test sample to find its similar training samples nearby. In this condition, if we increase k, we are more likely to include other classes which may results in higher probability of misclassification. The maximum accuracy is 0.9727 when k = 3. There is also a special phenomenon that when k is relatively small, say less than 20, the even k tends to have a smaller accuracy than the nearby odd k, which is especially significant when k = 2. This phenomenon is mainly caused by the Median tie break strategy.

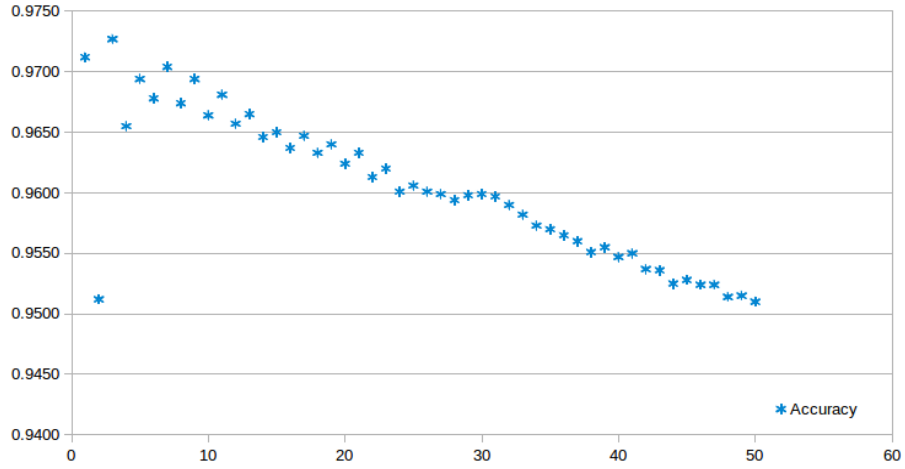

Fig. 2 The relationship between k and accuracy (Training data: 50000)

3. When the number of training data equals 50000 and k equals 3, the numbers get confused with each other most easily are 4 → 9 (19), 2 → 7 (18), 8 → 5 (18), 5 → 6 (16), 5 → 3 (16), 8 → 3 (12), 9 → 4 (11), 9 → 7 (11) (the number on the left of arrow is the true label, the () shows the number of cases). By extracting the cases with Median function, the result shows that the Median function only has some contributions to the misclassification of 5 → 6 (1), 8 → 5 (4), 8 → 3 (6), 9 → 7 (5). Apparently, Median function contributes about 50% to the misclassification of 8 → 3 and 9 → 7.