Likelihood Machine Learning Seminar II

Lecture 2: K Nearest Neighbors

Likelihood Lab

XingYu Fu

1. How K Nearest Neighbors Work?

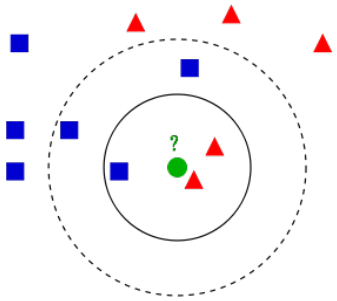
K Nearest Neighbors (KNN) is one of the most basic machine learning algorithms and is widely used in industry due to its simplicity. The core idea of KNN can be intuitively illustrated by a famous saying: 'Birds of a feather flock together.'

We use an example of fish to explain the decision-making process of KNN: say we have a group of fishes with two categories (e.g. A and B) and a single fish whose category is unknown to us. The task is to classify the unknown fish wisely. Here is how KNN works:

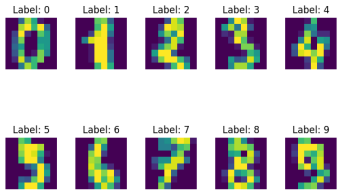
- Step1: find K fishes within the group which are most similar to the unknown fish.

- Step2: classify the unknown fish using majority voting.

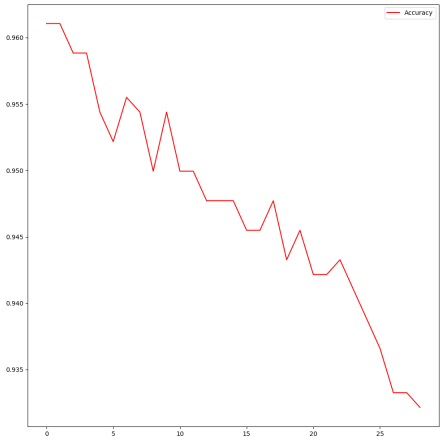
Another example of KNN from wikipedia is: The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If K = *3* (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If K = *5* (dashed line circle) it is assigned to the first class (3 squares versus 2 triangles inside the outer circle).



2. KNN on Hand Written Single Digit Classification



We apply the KNN algorithm in the hand written single digit classification problem, where each input sample is an eight by eight digit image. Half of them are used as training data and others are used for testing. Different K are selected to compare their prediction performances:



As we can see from the above figure, the choice of K has a great impact on the prediction accuracy, which is worth paying special attention to in practice.

3. High Performance KNN Algorithm

For those who may concern, there are some fast adaptations of the original KNN algorithm, to name a few:

- KD Tree

- Ball Tree

Both of them are implemented in sklearn python package.