We want to find and label a fruit automatically. Fruits are characterized by: weight (in grams),

colour as an integer

1. red

2. orange

3. yellow

4. green

5. blue

6. purple

Let us also represent the labels as (apple (A) and banana(B)).

We are given some samples such as < 303; 3; apple > meaning the fruit with 303 gram

weight, and yellow colour is an apple. A set of such \training examples" is given in \01-

train.csv". This has a small set of \labeled" 17 examples.

We are given a set of test data where only weight and colour is given (eg. < 373; 1 >). We

should design a simple Nearest Neighbour classifier that will nd the fruit label. i.e., apple

or banana. We have 102 such test cases.

As an objective of finding, how good is our prediction, we are also given labelled csv les,

which have labels for all the test cases. If your predicted label (fruit name) is correct, you

have done well !!. But there is also a chance that your predictions can go wrong. Let us be

realistic. We can use 17 examples the same or similar as in [1]. We can create test samples

random but around the labeled examples.

Exercises

1. Plot the training data on a 2D plane. How similar fruits come close in the feature space?

2. Now plot the test samples as “black" on the same plot.

3. Find the accuracy of the prediction (percentage of the samples that are correctly

predicted) with a simple KNN classifier with K=3 and distance we Euclidean distance.

You may observe that with a small number of labeled examples (17), we are predicting

the labels “accurately" on a larger novel set of examples.

4. Find the accuracy of another test set.

5. Find the accuracy of samples in the test set with K = 1, K = 5 and K = 17, . What

do you observe? Why is K = 17 a bad choice? What happens to all the labels now?

6. If we have used the weight in Kg (i..e, dividing the weight in grams by 1000), what

will be the accuracy for K = 3?

7. Assume we have used only the weight in grams (by discarding the colour feature) for

calculating distance. Calculate the accuracy on the test set. (modify only the distance

function).

8. If we have used sum of absolute differences, can the accuracies change?