The nearest neighbor classifiers

The nearest neighbor rule

A set of n pairs $(\mathbf{x}_1, t_1)$,..., $(\mathbf{x}_n, t_n)$ is given, where $\mathbf{x}_i$ takes real values and t_i takes values in the set $\{1, ..., M\}$. Each $\mathbf{x}_i$ is the outcome of the set of measurements made upon the ith individual. Each t_i is the index of the category to which the ith sample belongs. For brevity we say:

$\mathbf{x}_i$ belongs to category t_i

A set of measurements is made upon a new individual as $\mathbf{x}$, and we wish to assign $\mathbf{x}$ a label in $\{1, ..., M\}$. Let $\mathbf{x}_k$ be the sample nearest to $\mathbf{x}$, then the nearest neighbor rule is to assign $\mathbf{x}$ the label associated to $\mathbf{x}_k$.

$$\min\{d(\mathbf{x}, \mathbf{x}_i)\} = d(\mathbf{x}, \mathbf{x}_k), \quad i = 1, ..., n$$

A commonly used distance measure is the sum of squares. Suppose $\mathbf{x} = [x_1, x_2]^T$ and $\mathbf{x}_k = [x_{k1}, x_{k2}]^T$

$$d(\mathbf{x}, \mathbf{x}_k) = (x_1 - x_{k1})^2 + (x_2 - x_{k2})^2$$

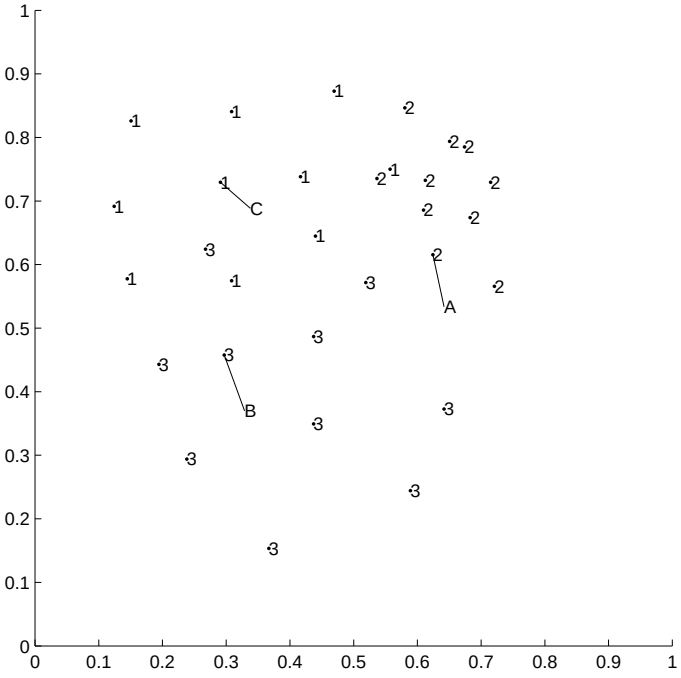

Table: There are three classes each having 10 known samples. Three new samples A,B and C are presented unlabelled. The algorithm can output the class label, for each new sample, as the label of its nearest neighbor. The results are generated by $nn.m$.

Example 1: In order to select the best candidates, an over-subscribed secondary school sets an entrance exam on two subjects of English and Mathematics. Suppose that we know the marks and the classification results of 5 applicants as in the Table below. If an applicant has been accepted, this is denoted as class 1, otherwise class 2. Use the nearest neighbor rule to determine if Andy should be accepted if his marks of English and Mathematics are 70 and 70 respectively.

Candidate No.	English	Math	Class
1	80	85	1
2	70	60	2
3	50	70	2
4	90	70	1
5	85	75	1

Solution:

1. Calculate the distance between Andy's marks and those of 5 applicants.

$$d_1 = (70 - 80)^2 + (70 - 85)^2 = 225$$

$$d_2 = (70 - 70)^2 + (70 - 60)^2 = 100$$

$$d_3 = (70 - 50)^2 + (70 - 70)^2 = 400$$

$$d_4 = (70 - 90)^2 + (70 - 70)^2 = 400$$

$$d_5 = (70 - 85)^2 + (70 - 75)^2 = 150$$

2. Find out the minimum value amongst $\{d_1, d_2, d_3, d_4, d_5\}$, which is $d_2 = 100$.

3. Look for the value of the Class for the No.2 applicant, which is 2. Hence the applicant is determined as not acceptable by the algorithm.

The k nearest neighbor rule (k-nn)

An obvious extension of the nearest neighbor rule is the k nearest neighbor rule. This rule classifies the new sample $\mathbf{x}$ by assigning it the label most frequently represented among the k nearest samples.

We will restrict our discussion on the case of two classes.

A decision is made by examining the labels on the k nearest neighbors and taking a vote (k is odd to avoid ties).

Using the same example, we can determine if Andy should be accepted with k nearest neighbor rule, with $k = 3$.

1. Calculate the distance between Andy's marks and those of 5 applicants. $d_1 = 125$, $d_2 = 100$, $d_3 = 400$, $d_4 = 400$ and $d_5 = 150$.

2. Find out the 3 smallest values amongst $\{d_1$, d_2, d_3, d_4, $d_5\}$, which is d_1, d_2, d_5.

3. Look for the values of the Class labels for No.1, No. 2 and No.3 applicants, which are $\{1,2,1\}$.

4. There are more ones in the set of $\{1,2,1\}$, so the applicant is determined as acceptable by the $3 - nn$ algorithm.

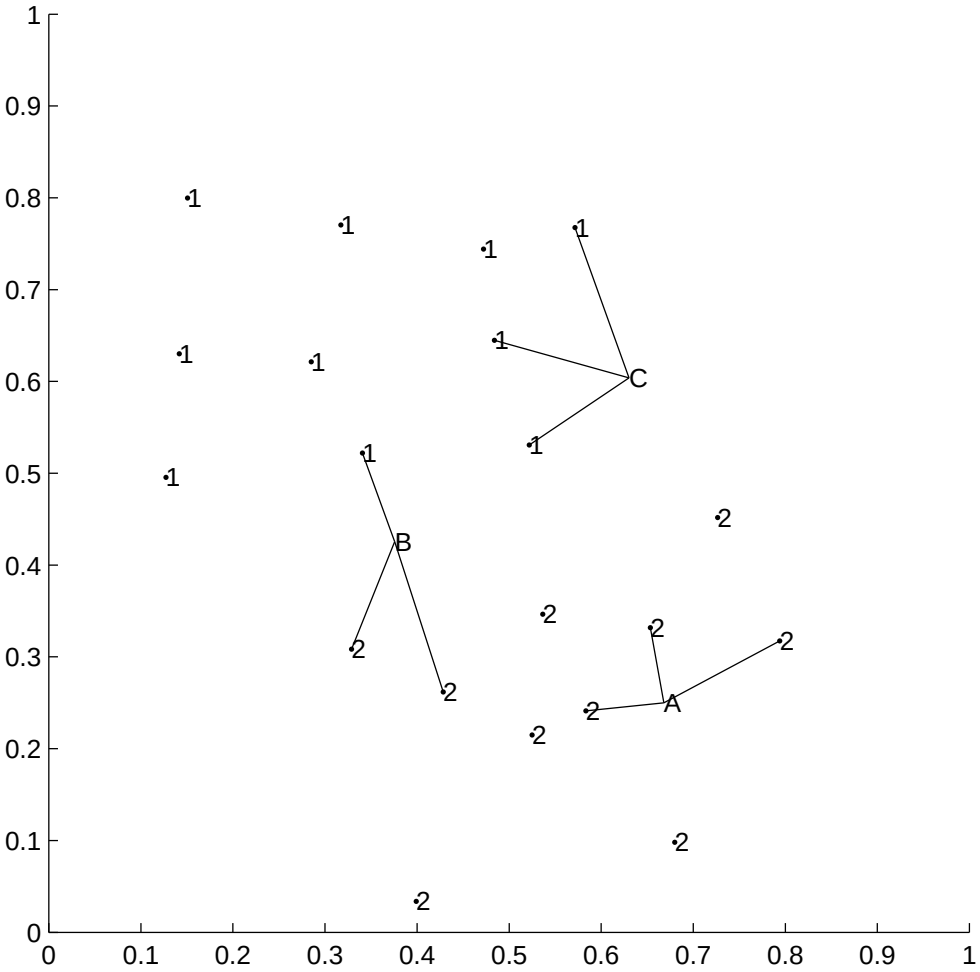

Table: There are two classes each having 10 known samples. Three new samples A,B and C are presented unlabelled. The algorithm can output the class label, for each new sample, as the label of the most represented 3 nearest neighbors. The results are generated by $knn.m$.