Problem 54-1

David Gieselman

October 2020

1 How to do k nearest neighbors.

a: This cookie can be represented as the point $\boldsymbol{P}$(0.10,0.15,0.30,0.45). Compute the Euclidean distance between $\boldsymbol{P}$ and each of the points corresponding to cookies in the data set:

Sorted list:
ID, Cookie Type, Euclidean distance
2, 'Shortbread', 0.037416573867739396
1, 'Shortbread', 0.04690415759823429
3, 'Shortbread', 0.061644140029689765
7, 'Sugar', 0.08831760866327845
4, 'Shortbread', 0.1224744871391589]
5, 'Sugar', 0.15811388300841897
6, 'Sugar', 0.158113883008419
12, 'Fortune', 0.17320508075688776
10, 'Fortune', 0.18708286933869708
9, 'Fortune', 0.21213203435596428
13, 'Fortune', 0.22759613353482086
8, 'Sugar', 0.24494897427831783
11, 'Fortune', 0.3959797974644666

b: Consider the 5 points that are closest to $\boldsymbol{P}$. (These are the 5 "nearest neighbors".) What cookie IDs are they, and what types of cookies are represented by these points?

The 5 closest have the IDs 2,1,3,7,4 and 4/5 of them are Short-breads with the exception of 7 which is a Sugar cookie.

c: What cookie classification showed up most often in the 5 nearest neighbors? What inference can you make about the recipe corresponding to the point $\boldsymbol{P}$?

Most of the 5 are Short-breads which leads me to believe that $\boldsymbol{P}$ is also a Short-bread.

2 The danger of using too large a k.

a: What happens if we try to perform the k nearest neighbors approach with k=13 (i.e. the full data set) to infer the cookie classification of point $\boldsymbol{P}$? What issue occurs, and why does it occur?

We would immediately come to the conclusion that $\boldsymbol{P}$ is whatever is the most common cookie in our data set regardless of what $\boldsymbol{P}$ is. This occurs because the method we use is designed to narrow down the search compared to others, and using the whole data set defeats the purpose.

b: For each classification of cookie, find the average distance between $\boldsymbol{P}$ and the points corresponding to the cookies in that classification. Explain how this resolves the issue you identified in part (a).

Shortbread: 0.05368787172696447
Sugar: 0.12989886979168686
Fortune: 0.19933265257513944

We can use this to see that the cookie category with the smallest average is the most accurate to what we're looking for and this allows us to have a k of the length of the data set.