# Assignment 54-1

Riley Paddock

October 16,2020

## Setup

The dataset below displays the ratio of ingredients for various cookie recipes.

['ID', 'Cookie Type', 'Portion Eggs', 'Portion Butter', 'Portion Sugar', 'Portion Flour']

[[1, 'Shortbread', 0.14, 0.14, 0.28, 0.44],
[2, 'Shortbread', 0.10, 0.18, 0.28, 0.44],
[3, 'Shortbread', 0.12, 0.10, 0.33, 0.45],
[4, 'Shortbread', 0.10, 0.25, 0.25, 0.40],
[5, 'Sugar', 0.00, 0.10, 0.40, 0.50],
[6, 'Sugar' , 0.00, 0.20, 0.40, 0.40],
[7, 'Sugar' , 0.10, 0.08, 0.35, 0.47],
[8, 'Sugar' , 0.00, 0.05, 0.30, 0.65],
[9, 'Fortune' , 0.20, 0.00, 0.40, 0.40],
[10, 'Fortune' , 0.25, 0.10, 0.30, 0.35],
[11, 'Fortune' , 0.22, 0.15, 0.50, 0.13],
[12, 'Fortune' , 0.15, 0.20, 0.35, 0.30],
[13, 'Fortune' , 0.22, 0.00, 0.40, 0.38]]

Suppose you're given a cookie recipe and you want to determine whether it is a shortbread cookie, a sugar cookie, or a fortune cookie. The cookie recipe consists of 0.10 portion eggs, 0.15 portion butter, 0.30 portion sugar, and 0.45 portion flour. We will infer the classification of this cookie using the " k nearest neighbors" approach.

## Part 1: How to do k nearest neighbors

### 1.a

This cookie can be represented as the point P(0.10,0.15,0.30,0.45). Compute the Euclidean distance between P and each of the points corresponding to cookies in the dataset.

```
['ID','EuclideanDistance']
[[1, 0.04690415759823429],
[2, 0.037416573867739396],
[3, 0.061644140029689765],
[4, 0.1224744871391589],
[5, 0.15811388300841897],
[6, 0.158113883008419],
[7, 0.08831760866327845],
[8, 0.24494897427831783],
[9, 0.21213203435596428],
[10, 0.18708286933869708],
[11, 0.3959797974644666],
[12, 0.17320508075688776],
[13, 0.22759613353482086]]
```

## 1.b

Consider the 5 points that are closest to 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 points are cookes 2,1,3,7,4

```
['ID','EuclideanDistance','Type']
[2, 0.037416573867739396, 'Shortbread'],
[1, 0.04690415759823429, 'Shortbread'],
[3, 0.061644140029689765, 'Shortbread'],
[7, 0.08831760866327845, 'Shortbread'],
[4, 0.1224744871391589, 'Sugar'],
```

## 1.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 P ?
The 'Shortbread' Cookie showed up the most in the 5 nearest neighbors. From this we can infer that the recipe is for a Shortbread cookie.

## Part 2: The danger of using too large a k

### 2.a

What happens if we try to perform the k nearest neighbors approach with k=13 (i.e. the full dataset) to infer the cookie classification of point P? What issue occurs, and why does it occur?
If we include all of the dataset in the nearest neighbors we will run into the problem where the classification that shows up the most in the nearest neighbors isnt the classification closest to your data point but the classification that you have the most data on. So it becomes less about the closest points and more about the most points of a certain classification.

### 2.b

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

$['Type',$ 'Average Euclidean Distance']
$['Shortbread', 0.06710983965]$
$['Sugar', 0.16237358724]$
$['Fortune', 0.23919918309]$

This resolves the issue of having too many k's because by taking the average we are acknowledging every point in the set but we are deciding the nearest classification by finding the average distance from each thing of that classification. So this way we are giving equal weight to every classification, where before we were giving more weight to classifications with larger data sets, and acknowledging every data point.