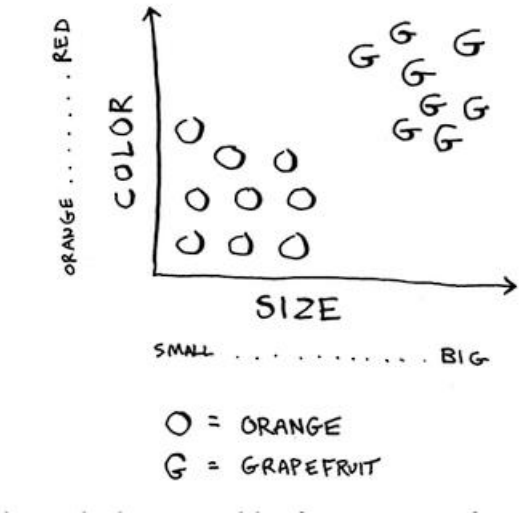
Chapter 10

In this chapter we learn to build a classification system and some concepts of machine learning

Orange vs grapefruit

If I put a fruit in front of you and asked whether it's an orange or grapefruit so you will begin to think about the main features of each one to figure out which features fits the fruit in front of you at the end you probably get the idea that grapefruit is bigger and more red than orange you may have this graph in your mind by now



This graph here makes it a bit easy but there is a slight problem you notice the blank area in the graph this area represents an uncertainty what if our fruit lies in this area here, we cannot be sure what is this fruit in this case we check the nearest neighbors of the fruit if the more of the neighbors are orange then the fruit is probably orange and vice versa you just use the KNN or K-nearest neighbors algorithm for classification let's see another example

Recommendation system

Suppose you are Netflix, and you want to build a movie recommendation system just like the previous problem plot all you users on a graph and users with similar interests are going to be close to each other and this way if one of the users likes a movie it will be recommended to the users nearest to him

Feature extraction

We just discussed two different examples and we could figure out the similarities ourselves quite easily but the process isn't that easy for the computer in the grapefruit example we could till easily if the fruit is red or not and how red it is but the computer can't do this so we can give a scale from 1 to 5 where 1 is least red and the redness increases as we go up and the same

thing for size in this way we could plot them easily on a graph the computer would understand also you have to bear in mind that you need to choose only the features that will help in the classification process for example you choose the roundness of the fruit as a feature it will be redundant and may lead to wrong classification then how you choose features for your recommendation system when you sign up for Netflix they ask you to rate categories comedy, action, drama, horror and romance based on that your graph is plotted but the user here has five dimensions and we don't know how to plot a five dimensional graph the truth is you don't need to remember the equation for the distance between two points $\sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$ this equation can work even if you had million dimension to you point and according to this equation we get to know who the nearest neighbors to a point are let's say these are two Netflix users (3,2,1,5,4) and (5,2,1,3,0) the distance is going to be $\sqrt{((5- 3)^2+ (2 - 2)^2+(1 - 1)^2+(3- 5)^2+(0 -4)^2}$ this evaluates to $2\sqrt{6}$ this number here represents the distance between the two points on the graph (how near are they to each other)

regression

sometimes your aim isn't only to classify but you want to predict also after recommending a movie to user you want to know how he will rate it so let's say you recommended a movies based on the nearest five users to him those 5 users rated the moves on scale from 1 to 5 as follows 5,4,4,5,3 you can predict the rating the user will give by averaging the ratings of the users near to him in this case you will get 4.2 this is what we call regression and it's very useful

Note: the number of nearest neighbors you choose can vary that is why it is called K but be careful because too many neighbors or very few neighbors will lead you to wrong classifications

Introduction to machine learning

Machine learning is making your computer more intelligent. You already saw examples for that using KNN let's look at another example.

OCR: stands for Optical character recognition that means you can take a picture of a page of text and your computer will automatically read it. Google for example uses it to digitize books so how does it work?
if you are given a picture of number 7 to make the computer understand what is in the picture you will do the following
1-go through pictures of numbers and extract features (here you are saving the features of each number you teaching your computer how number 7 looks like and how number 3 looks like and so on)
2-when you get a new image extract it's features and see what it's nearest neighbors are

After that it is like the grapefruit example

Note: features for numbers and characters can be lines, points, and curves

Building a spam filter

Spam filters use a simple algorithm called Naïve Bayes classifier first you need to train your classifier.

training set.

subject	Spam?
“Reset your password”	Not spam
“You have won 1 million dollars”	Spam
“Send me your password”	Spam
“Nigerian prince sends you 10 million dollars”	spam
“Happy birthday”	Not spam

Let us say that you get a message that says “collect your million dollars the classifier will break the message into words and check the probability of existence of each word in a spam message in this simple model all the spam messages had the word million in it so the new message will be considered spam.

Predicting stock prices

That something that is hard to do using machine learning because there is no clear features to extract I can’t say since the stock went up yesterday it will go up tomorrow or that stocks always go up in may there is no guaranteed way to use the past to predict the future so it’s almost impossible.

Ending

That was all for this chapter and for the book as a whole this summary can be useful for any one whether you have read the book and using to go back to some stuff that you have forgot or you are using it to get a general idea of the book I hope it was of any help and finally there is a remaining chapter but it’s better to read it from the book it’s giving you an idea of what you could start to study after you finish this book so I thought it’s better to read it from the book.