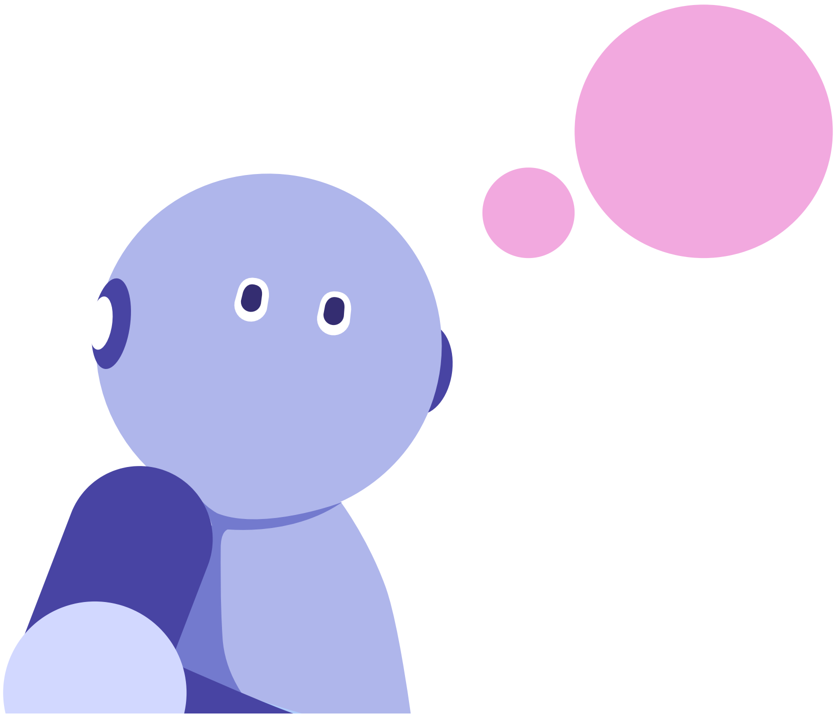
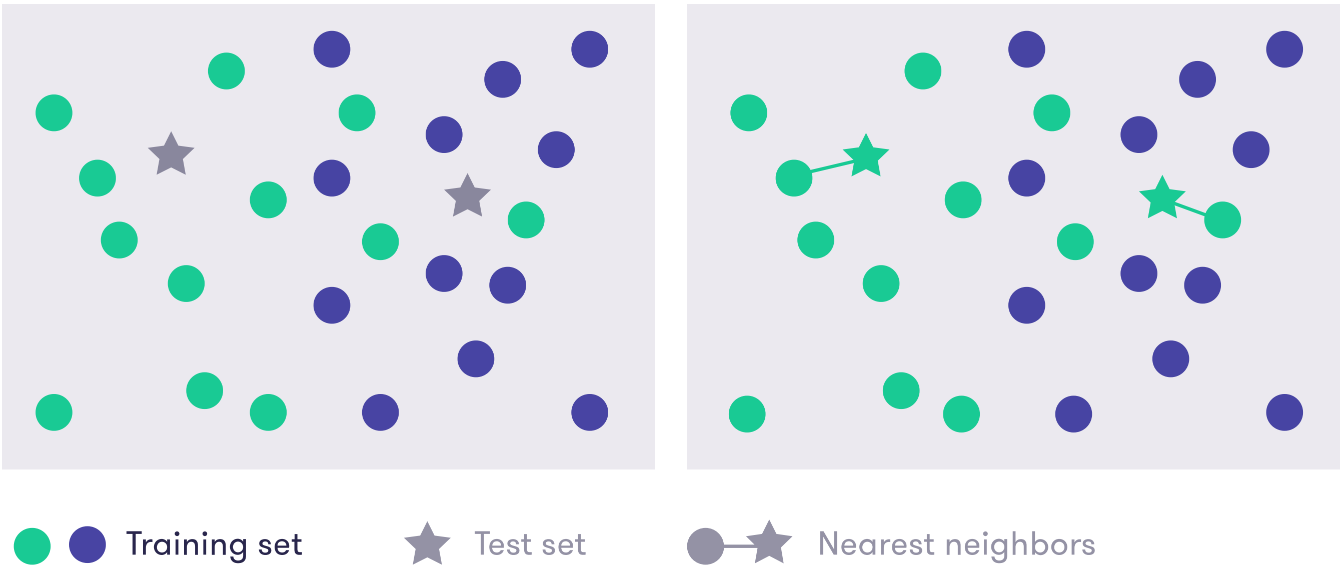


II. The nearest neighbor classifier

The nearest neighbor classifier is among the simplest possible classifiers. When given a item to classify, it finds the training data item that is most similar to the new item, and outputs its label. An example is given in the following diagram.



In the above diagram, we show a collection of training data items, some of which belong to one class (green) and other to another class (blue). In addition, there are two test data items, the stars, which we are going to classify using the nearest neighbor method.

The two test items are both classified in the “green” class because their nearest neighbors are both green (see diagram (b) above).

The position of the points in the plot represents in some way the properties of the items. Since we draw the diagram on a flat two-dimensional surface — you can move in two independent directions: up-down or left-right — the items have two properties that we can use for comparison. Imagine for example representing patients at a clinic in terms of their age and blood-sugar level. But the above diagram should be taken just as a visual tool to illustrate the general idea, which is to relate the class values to similarity or proximity (nearness). The general idea is by no means restricted to two dimensions and the nearest neighbor classifier can easily be applied to items that are characterized by many more properties than two.

What do we mean by nearest?

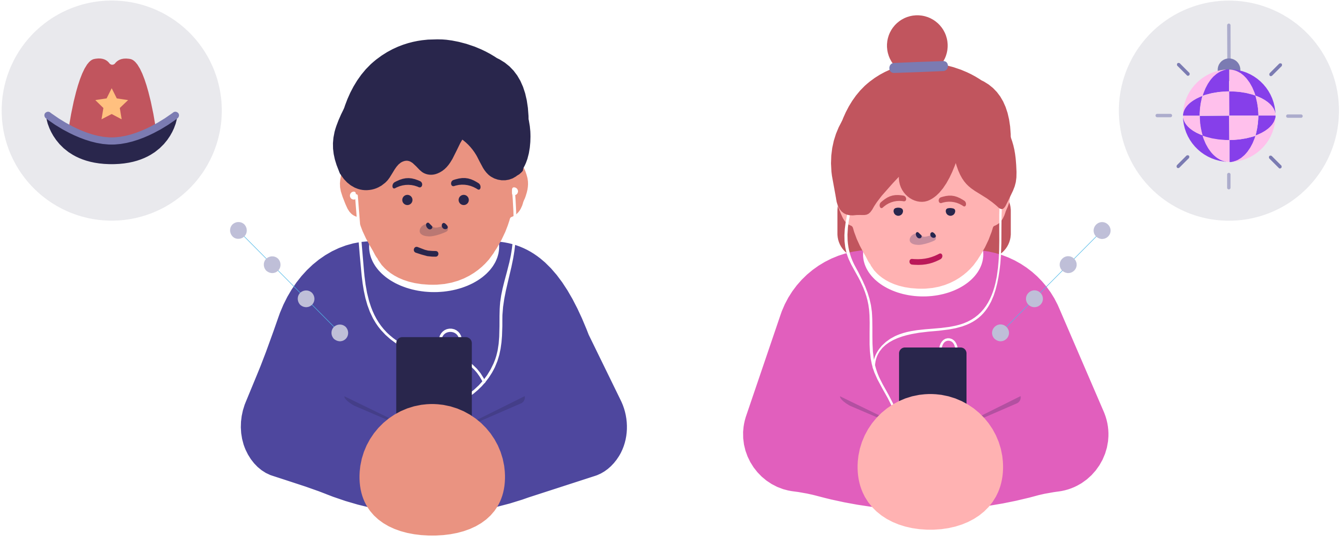
An interesting question related to (among other things) the nearest neighbor classifier is the definition of distance or similarity between instances. In the illustration above, we tacitly assumed that the standard geometric distance, technically called the Euclidean distance, is used. This simply means that if the points are drawn on a piece of paper (or displayed on your screen), you can measure the distance between any two items by pulling a piece of thread straight from one to the other and measuring the length.

Note

Defining ‘nearest’

Using the geometric distance to decide which is the nearest item may not always be reasonable or even possible: the type of the input may, for example, be text, where it is not clear how the items are drawn in a geometric representation and how distances should be measured. You should therefore choose the distance metric on a case-by-case basis.

In the MNIST digit recognition case, one common way to measure image similarity is to count pixel-by-pixel matches. In other words, we compare the pixels in the top-left corner of each image to one another and if the more similar color (shade of gray) they are, the more similar the two images are. We also compare the pixels in the bottom-right corner of each image, and all pixels inbetween. This technique is quite sensitive to shifting or scaling the images: if we take an image of a '1' and shift it ever so slightly either left or right, the outcome is that the two images (before and after the shift) are very different because the black pixels are in different positions in the two images. Fortunately, the MNIST data has been preprocessed by centering the images so that this problem is alleviated.



Music recommendations

Using nearest neighbors to predict user behavior

A typical example of an application of the nearest neighbor method is predicting user behavior in AI applications such as recommendation systems.

The idea is to use the very simple principle that users with similar past behavior tend to have similar future behavior. Imagine a music recommendation system that collects data about users' listening behavior. Let's say you have listened to 1980s disco music (just for the sake of argument). One day, the service provider gets their hands on a hard-to-find 1980 disco classic, and add it into the music library. The system now needs to predict whether you will like it or not. One way of doing this is to use information about the genre, the artist, and other metadata, entered by the good people of the service provider. However, this information is relatively scarce and coarse and it will only be able to give rough predictions.

What current recommendation systems use instead of the manually entered metadata, is something called collaborative filtering. The collaborative aspect of it is that it uses other users' data to predict your preferences. The word "filter" refers to the fact that you will be only recommended content that passes through a filter: content that you are likely to enjoy will pass, other content will not. (This kind of filters may lead to the so called filter bubbles, which we mentioned in Chapter 1. We will return to them later.)

Now let's say that other users who have listened to 80s disco music enjoy the new release and keep listening to it again and again. The system will identify the similar past behavior that you and other 80s disco fanatics share, and since other users like you enjoy the new release, the system will predict that you will too. Hence it will show up at the top of your recommendation list. In an alternative reality, maybe the added song is not so great and other users with similar past behavior as yours don't really like it. In that case, the system wouldn't bother recommending it to you, or at least it wouldn't be at the top of the list of recommendations to you.

The following exercise will illustrate this idea.

Unanswered

Exercise 14: Customers who bought similar products

In this exercise, we will build a simple recommendation system for an online shopping application where the users' purchase history is recorded and used to predict which products the user is likely to buy next.

We have data from six users. For each user, we have recorded their recent shopping history of four items and the item they bought after buying these four items:

User	Shopping History				Purchase
Sanni	boxing gloves	Moby Dick (novel)	headphones	sunglasses	coffee beans
Jouni	t-shirt	coffee beans	coffee maker	coffee beans	coffee beans
Janina	sunglasses	sneakers	t-shirt	sneakers	ragg wool socks
Henrik	2001: A Space Odyssey (dvd)	headphones	t-shirt	boxing gloves	flip flops
Ville	t-shirt	flip flops	sunglasses	Moby Dick (novel)	sunscreen
Teemu	Moby Dick	coffee	2001: A	headphones	coffee

	(novel)	beans	Space Odyssey (dvd)		beans

The most recent purchase is the one in the rightmost column, so for example, after buying a t-shirt, flip flops, sunglasses, and Moby Dick (novel), Ville bought sunscreen. Our hypothesis is that after buying similar items, other users are also likely to buy sunscreen.

To apply the nearest neighbor method, we need to define what we mean by nearest. This can be done in many different ways, some of which work better than others. Let’s use the shopping history to define the similarity (“nearness”) by counting how many of the items have been purchased by both users.

For example, users Ville and Henrik have both bought a t-shirt, so their similarity is 1. Note that flip flops doesn't count because we don't include the most recent purchase when calculating the similarity — it is reserved for another purpose.

Our task is to predict the next purchase of customer Travis who has bought the following products:

User	**Shopping History**				**Purchase**
Travis	green tea	t-shirt	sunglasses	flip flops	?

You can think of Travis being our test data, and the above six users make our training data.

Proceed as follows:

1. Calculate the similarity of Travis relative to the six users in the training data (done by adding together the number of similar purchases by the users).
2. Having calculated the similarities, identify the user who is most similar to Travis by selecting the largest of the calculated similarities.
3. Predict what Travis is likely purchase next by looking at the most recent purchase (the rightmost column in the table) of the most similar user from the previous step.

Who is the user most similar to Travis?

Your answer...

What is the predicted purchase for Travis?

Your answer...

Submit

In the above example, we only had six users' data and our prediction was probably very unreliable. However, online shopping sites often have millions users, and the amount of data they produce is massive. In many cases, there are a hoard of users whose past behavior is very similar to yours, and whose purchase history gives a pretty good indication of your interests.

These predictions can also be self-fulfilling prophecies in the sense that you are more likely to buy a product if it is recommended to you by the system, which makes it tricky to evaluate how well they actually work. The same kind of recommendation systems are also used to recommend music, movies, news, and social media content to users. In the context of news and social media, filters created by such systems can lead to filter bubbles.

Unanswered

Exercise 15: Filter bubbles

As discussed above, recommending news of social media content that a user is likely to click or like, may lead to filter bubbles where the users only see content that is in line with their own values and views.

1. Do you think that filter bubbles are harmful? After all, they are created by recommending content that the user likes. What negative consequences, if any, may be associated with filter bubbles? Feel free to look for more information from other sources.
2. Think of ways to avoid filter bubbles while still being able to recommend content to suit personal preferences. Come up with at least one suggestion. You can look for ideas from other sources, but we'd like to hear your own ideas too!

Note: your answer should be at least a few sentences for each part.

Your answer...

Words: 0

Submit

Next section

III. Regression →

Course overview

About

FAQ

Privacy Policy

My profile Sign out