

SUPERVISED LEARNING

1. Supervised learning

00:00 - 00:14

Hi! My name is Hadrien, I'm a Content Developer here at DataCamp. Great job on Chapter 1! Now that you know more about machine learning, let's dive in deeper.

2. Modeling

00:14 - 00:21

Machine Learning is all about modeling. Not that kind of modeling.

3. Types

00:21 - 00:28

As you saw in Chapter 1, there are several types to machine learning, including supervised learning.

4. What is supervised learning?

00:28 - 00:37

Supervised learning is basically a labeling machine. It takes an observation, and assigns a label to it.

5. Classification and regression

00:37 - 00:44

There are two flavors of supervised learning: classification and regression.

6. Classification

00:44 - 00:48

Let's focus on classification first.

7. Classification

00:48 - 01:16

Classification consists in assigning a category to an observation. We're predicting a discrete variable, a variable that can only take a few different values. Is this customer going to stop its subscription or not? Is this mole cancerous or not? Is this wine red, white or rosé? Is this flower a rose, a tulip, a carnation, a lily?

8. Observations

01:16 - 01:25

Remember, we feed the model observations. Let's take college admissions where we want to predict acceptance.

9. Features

01:25 - 01:40

For simplification, we show two features here: GPA and admission test results. We could have more features like: involvement in student organizations and sports or prizes applicants have won.

10. Target

01:40 - 01:58

And, the target is what we want to predict. There are two possible labels: those who are accepted and those who are not. The target can only be one of these two labels, making it a classification problem.

11. Graphing our data

01:58 - 02:15

Here are our observations plotted. GPA on the x axis and test results on the y axis. Blue points represent accepted applicants, red points represent rejected applicants.

12. Splitting data

02:15 - 02:20

We keep 80% of our data to train our model.

13. Manual classifier

02:20 - 03:01

Because we use just two features, we're able to plot and interpret the results. For us humans, it's pretty clear. If you score above 4 on both the GPA and entrance test, you're accepted. If we added more features like extracurriculars or prizes, we would need more axes. It would be very hard for us to interpret the data with our human eye. However, an model wouldn't struggle at all. We can use a Support Vector Machine. It sounds scary, but it's just a line separating our points.

14. Support vector machine - linear classifier

03:01 - 03:25

We train our algorithm and classify the 20% of observations we left aside. It only misclassifies two blue points as red - meaning two applicants were wrongly predicted as rejected. The problem is, it tries to separate with a straight line, so it's unlikely to do better than that.

15. Support vector machine - polynomial classifier

03:25 - 03:40

What if we allow a curved line instead? Then it classifies everything correctly! There are ways we can tweak our model's behavior, like allowing curves. We will talk about that later.

16. Regression

03:40 - 03:44

Now, what about regression?

17. Regression

03:44 - 04:06

While classification assigns a category, regression assigns a continuous variable, that is, a variable that can take any value. For example, how much will this stock be worth? What is this exoplanet's mass? How tall will this child be as an adult?

18. Predicting temperature

04:06 - 04:15

Let's look at some new data: weather readings. Can we predict temperature based on humidity?

19. Training data

04:15 - 04:25

We use 80% of the data to train our model. Seems like when humidity rises, temperature decreases.

20. Linear regression

04:25 - 04:30

Indeed, our linear regression model catches that.

21. Model

04:30 - 04:33

Making this is our model.

22. Given humidity...

04:33 - 04:39

According to our model, if the humidity rate is 0.5,

23. ...find temperature

04:39 - 04:44

Then the temperature is 18.5 °C.

24. Testing data

04:44 - 05:01

And this is how the model performs on real data. It identified the trend, but is still bad at predicting. More features, like wind, cloudiness, location, season could make it more accurate.

25. Classification vs regression

05:01 - 05:33

It's up to you to choose whether you want to frame your problem as a regression or classification problem. For example, we could predict an exact temperature, or a categorical range of temperatures like "Cold", "Mild", and "Hot". This is true for something like age, where we can have categories like baby, child, teenager, adult, and so on! Do you see the difference?

26. Let's practice!

05:33 - 05:40

Let's make sure everything is clear with some exercises.