EVALUATING PERFORMANCE

1. Evaluating performance

00:00 - 00:05

You now know more about different model types and their outputs.

2. Evaluate step

00:05 - 00:11

But how do we evaluate a model's output? This is step 4 of the workflow.

3. Overfitting

00:11 - 00:40

The first thing to look for when evaluating is overfitting. That's when our model performs great on training data, but poorly on the testing data. This is bad: it means our model learned the training set by heart and is unable to generalize learnings to new data, which is what we originally want. This is why we need to split our dataset into two sets.

4. Illustrating overfitting

00:40 - 01:01

For instance, the green line here overfits. It's going out of its way to classify all points perfectly, thus performing great on this dataset, but poorly on unseen data. The black line makes more errors on this specific dataset, but generalizes better.

5. Accuracy

01:01 - 01:25

In our college acceptance problem, how would we measure the model's performance? We could use accuracy, which is the number of correctly predicted observations divided by the number of all observations. We had 48 correct classifications out of 50, which is a 96% accuracy.

6. Limits of accuracy: fraud example

01:25 - 02:19

However, accuracy isn't always the best metric. Consider fraud where only a small minority of transactions are fraudulent. Let's say we train a model to predict whether a transaction is fraudulent or legitimate. Here is a graph showing its performance on 30 test data points. It only misclassifies 2 points, giving it an accuracy of about 93%, which sounds good. But the model actually misses the majority of fraudulent transactions, which will be a problem if we deploy this model in the real world. We can even say all points are legitimate, and we get a 90% accuracy, but we miss all the fraudsters. In this case, we need a better metric.

7. Confusion matrix

02:19 - 02:24

Enter the confusion matrix. Let's fill it out.

8. True positives

02:24 - 02:31

True positives are fraudulent points that are correctly classified as fraudulent.

9. True positives

02:31 - 02:36

We only have one and that's the sole red point in the red area.

10. False negatives

02:36 - 02:45

False negatives are fraudulent observations that are incorrectly classified as legitimate.

11. False negatives

02:45 - 02:52

We have two false negatives. These are the red points outside the red area.

12. Remembering False Negatives

02:52 - 02:58

False negatives are like a smoke alarm not going off in the presence of smoke.

13. Fill out the rest...

02:58 - 03:09

Getting a hang of it? Take a pause and try to fill out the remaining squares. How did it go?

14. False positives, true negatives

03:09 - 03:43

There are no false positives, meaning legitimate points that are incorrectly predicted as fraudulent. We see this because there are no blue points in the red area. True negatives are legitimate points correctly predicted as not fraudulent. These are the blue points in the blue area, totaling 27. If you add up all the squares, you should get the total of number of points, which is 30 here!

15. Remembering False Positives

03:43 - 03:48

False positives are like a smoke alarm going off when there's no smoke.

16. Sensitivity

03:48 - 04:35

So why fill out this matrix? Remember, we wanted a better metric than accuracy for our fraud scenario. Sensitivity values accurate prediction of fraudulent transactions specifically by valuing true positives more. Here's the formula, which you don't need to memorize. We get 33% sensitivity, which is a bad score. Optimizing for sensitivity means we'd rather mark legitimate transactions as suspicious than authorize fraudulent transactions. Now, we see our model isn't good at predicting fraud and needs improvement.

17. Specificity

04:35 - 04:54

On the other hand we have specificity, which values true negatives. This is useful metric for spam filters. For an email user, it's better to send spam to the inbox rather than send real emails to the spam folder for deletion.

18. Evaluating regression

04:54 - 05:06

That's classification. But what about regression? Essentially, we want the difference between the actual value and the predicted value.

19. Evaluating regression

05:06 - 05:19

This can be the distance between the points and the predicted line. There are many ways to calculate this error, such as root mean square error, but this is the general idea!

20. Unsupervised learning

05:19 - 05:45

And what about unsupervised learning? Well, remember, unsupervised learning does not have predicted variables, so there's no correct output to compare to. How well your unsupervised learning model performs depends on the problem you're solving. So, you assess the performance based on how well the results advance your initial objective.

1. [1] https://www.flickr.com/photos/micahdowty/8540188997

21. Let's practice!

05:45 - 05:50

It's time for some exercises!