1

00:00:00,560 --> 00:00:06,350

[Auto-generated transcript. Edits may have been applied for clarity.]

We're talking about model evaluation and metrics. For classification models, we use several important metrics.

2

00:00:06,680 --> 00:00:12,800

For regression models, we use others such as r squared, r squared and mean squared error and MSE.

3

00:00:13,520 --> 00:00:15,740

First, let's discuss regression metrics.

4

00:00:16,130 --> 00:00:23,780

MSE can be used both as a loss function during training and as an evaluation metric after training on the test set in regression,

5

00:00:23,900 --> 00:00:26,960

you have a numerical target and one or more attributes.

6

00:00:27,260 --> 00:00:32,780

You're fitting a line to the training data and minimizing the squared distances from the line to the data points.

7

00:00:33,200 --> 00:00:39,800

When a new data point arrives in the test set, you compare the predicted value from the line to the true value.

8

00:00:40,130 --> 00:00:47,630

The squared difference becomes part of the evaluation. The lower the average MSE across test points, the better the model.

9

00:00:48,170 --> 00:00:55,700

R squared is another metric for regression. It represents how much of the variation in the target variable is explained by the predictor.

10

00:00:56,000 --> 00:01:06,290

For instance, if you're predicting mouse weight from mouse height and get an R-squared of 81%, it means 81% of the variation is explained by height.

11

00:01:06,740 --> 00:01:13,460

In contrast, if you predict mouse weight based on time spent sniffing a rock and get an R-squared of 6%,

12

00:01:13,820 --> 00:01:19,700

that means sniffing time explains only 6% of the variation, making it a poor predictor.

13

00:01:20,680 --> 00:01:24,430

Now let's explore classification metrics for classification.

14

00:01:24,550 --> 00:01:31,000

Common metrics include accuracy, F1 score, and area under the ROC curve AUC.

15

00:01:31,330 --> 00:01:34,660

These help evaluate the performance of classification models.

16

00:01:35,290 --> 00:01:38,320

Let's assume you're identifying whether someone has a disease.

17

00:01:38,590 --> 00:01:41,770

The positive class one represents sick.

18

00:01:41,980 --> 00:01:45,310

The negative class zero represents healthy.

19

00:01:45,760 --> 00:01:52,300

Here's how predictions are categorized. True positive TP you predict sick and the person is sick.

20

00:01:52,630 --> 00:01:57,580

False positive FP you predict sick, but the person is healthy.

21

00:01:58,090 --> 00:02:02,590

True negative TN you predict healthy and the person is healthy.

22

00:02:03,130 --> 00:02:07,600

False negative f n you predict healthy but the person is sick.

23

00:02:08,050 --> 00:02:11,740

This structure forms the confusion matrix in statistics.

24

00:02:12,550 --> 00:02:18,640

False positives are called type one errors. False negatives are called type two errors.

25

00:02:19,180 --> 00:02:22,510

Accuracy is defined as the number of correct predictions.

26

00:02:22,630 --> 00:02:26,860

TP plus ten divided by the total number of predictions.

27

00:02:27,340 --> 00:02:32,020

A high accuracy is good, but it can be misleading in imbalanced data sets.

28

00:02:32,590 --> 00:02:38,260

For example, if only 1 in 100,000 people has a disease,

29

00:02:38,650 --> 00:02:45,880

a model that classifies everyone is healthy will have very high accuracy, but it will miss all disease cases.

30

00:02:46,330 --> 00:02:54,100

To address this, we use the F1 score, which combines precision, which is how many predicted positives are correct,

31

00:02:54,580 --> 00:02:59,200

and recall how many actual positives are correctly identified.

32

00:02:59,680 --> 00:03:03,910

F1 is more balanced and sensitive to imbalanced classes.

33

00:03:04,540 --> 00:03:11,770

It's calculated from the confusion matrix and penalizes false negatives and false positives based on context.

34

00:03:12,280 --> 00:03:15,400

Finally, let's discuss thresholds and ROC curve.

35

00:03:16,030 --> 00:03:19,210

Classification models often output probabilities.

36

00:03:19,630 --> 00:03:27,070

By default, values above 0.5 are considered plus one and below 0.5 are class zero.

37

00:03:27,400 --> 00:03:35,380

Adjusting the threshold affects classification by. Lowering the threshold may reduce false negatives, but increase false positives.

38

00:03:35,800 --> 00:03:44,680

Raising the threshold may reduce false positives, but increase false negatives, and visualizing the impact of threshold changes is helpful.

39

00:03:44,710 --> 00:03:49,390

Some resources offer animations to show how the threshold affects outcomes.

40

00:03:50,020 --> 00:03:56,350

A key metric for comparing thresholds is the ROC curve receiver operating characteristic curve.

41

00:03:56,950 --> 00:03:59,890

It plots true positive rate TPR,

42

00:03:59,980 --> 00:04:10,120

which are correctly identified positives out of all actual positives and false positive rate FPR which are incorrectly identified positives.

43

00:04:10,120 --> 00:04:17,680

Out of all actual negatives, we want a classifier that achieves a TPR close to one and an FPR close to zero.

44

00:04:18,250 --> 00:04:22,480

The area under the ROC curve AUC quantifies this performance.

45

00:04:23,080 --> 00:04:26,110

A perfect classifier has an AUC of one.

46

00:04:26,740 --> 00:04:32,230

If two classifiers have different ROC curves, the one with the higher AUC is better.

47

00:04:32,860 --> 00:04:38,590

During training, engineers can also penalize false negatives or false positives in the loss function.

48

00:04:39,130 --> 00:04:46,300

For example, in rare disease detection, false negatives are costly, so models are trained to reduce them.

49

00:04:47,380 --> 00:04:50,410

All right. This concludes the key topics for week two.

50

00:04:50,530 --> 00:04:51,700

See you in week three.