Blog > Engineering

Guide to accuracy, precision, and recall

First published on October 13, 2022
Last updated at June 1, 2023

8 minute read

Tommy Dang
Engineering

TLDR

> Accuracy tells you how many times the ML model was correct overall. Precision is how good the model is at predicting a specific category. Recall tells you how many times the model was able to detect a specific category.

Outline

Introduction

Most machine learning (ML) problems fit into 2 groups: **classification and regression**. The main metrics used to assess performance of classification models are accuracy, precision, and recall.

To demonstrate each of these metrics, we'll use the following example: We're a mage on a quest to save the villagers from a horde of monsters. There are 100 monsters attacking the village. We're going to use machine learning to predict what element each monster is weak against. There are 3 possible elements that monsters are weak against: fire, water, and wind.

Accuracy

After summoning our machine learning model, we approach the village and begin the battle. After hours of intense combat, we are victorious and the villagers are all saved. We sit down by the fire, have a few refreshments (e.g. boba from the local village tea shop), and review how the battle unfolded.

For every monster we fought, our model predicted an element for us to use that would make the monster the weakest possible. Out of 100 predictions, our model predicted the correct element 73 times. Therefore, the accuracy of our model is 73%; 73 correct predictions out of 100 predictions made.

Precision

Measuring our model's accuracy told us overall how many times we were correct when predicting a monster's weakness. However, if we want to know how many times we were correct when predicting a monster was weak against fire, we need to use the precision metric.

Let's say we predicted 30 monsters were weak against fire. At the end of the battle, we saw that of those 30 monsters, 6 of them were actually weak against water and 3 of them were actually weak against wind. That means our precision for predicting fire is 70% (21/30). In other words, we predicted 21 fire weaknesses correctly out of 30 fire predictions.

After further review of the battle, we predicted 40 monsters were weak against water and 30 monsters were weak against wind. Out of the 40 monsters who we predicted were weak against water, we were correct 28 times resulting in 70% precision (28/40). Out of the 30 monsters who we predicted were weak against wind, we were correct 24 times resulting in 80% precision (24/30).

The following chart is called a confusion matrix:

The left y-axis is what was predicted. The top x-axis is what the monster was truly weak against. Look at row 1 and column 2. It's saying that out of the 30 monsters (21 + 6 + 3) that the model predicted a fire weakness for, 6 of those monsters were actually weak against water.

For more information on how to read a confusion matrix, check out this **article**. Here are 2 helpful images:

Source: Anuganti Suresh

Source: NillsF blog

Recall

Now that we know how many times our model predicted correctly overall and how good our model is at predicting a specific weakness (e.g. fire, water, wind), it’s time to find out how good our model is at finding monsters with a particular weakness.

For all the monsters who were actually weak against wind (there are 36 of them), how many did our model find out? Our model found 24 monsters who were weak against wind. That means our model has 67% recall when predicting monsters who are actually weak against wind.

How to determine if your model is doing well

Before evaluating how good a model is doing, we must establish a baseline score. A baseline score is a value that our model should perform better than. For example, if we have a baseline score of 60% accuracy and our model achieved an accuracy greater than that (e.g. 61%), then the model is good enough to use in real life scenarios (e.g. production environment).

There are many ways to establish a baseline score for your model. In an upcoming tutorial, we'll show you multiple other methods that can be used.

For our monster weakness prediction model, we'll use a baseline score method called mode category. This method takes the most abundant weakness category and divides that by the number of predictions. Specifically, more monsters are weak against wind than any other element (only 29 monsters are weak against fire and only 35 monsters are weak against water).

Our baseline accuracy score will be 36% (e.g. 36 monsters weak against wind divided by 100 monsters in total).

Our model had a 73% accuracy score. That is well above the baseline score. We can conclude that our model is performing well.

Note: you may want to have different baseline scores and use different metrics other than accuracy (e.g. precision, recall, etc) to assess whether your model performed well or not. Depending on your use case, precision or recall may be more important.

When to use precision or recall

On our quest to save the village from monsters, our goal was to defeat all the monsters. If we can predict their weakness, we have a higher chance of victory. Therefore, we want a model with the highest accuracy; in other words, a model with the highest chance of predicting the weakness correctly.

However, there are scenarios where you may want a model with higher precision. For example, you may have a fire spell that drains a lot of your energy and you want to use it sparingly. In that case, you'll want a model that has high precision when predicting a monster is weak against fire. That way, when you use your fire spell, you know with high certainty that the monster is weak against it.

Source: Lifehack

The scenario in which you want to choose a model with high recall is when the predicted outcome leads to a critical decision. For example, let's say we had a model that predicts whether a villager has been poisoned by a monster and needs immediate medical treatment. We want a model that can find every villager that is poisoned even if it incorrectly labels a healthy villager as needing medical treatment. The reason why we want this model to have high recall is because it's better to give a healthy person unnecessary treatment than to accidentally pass over someone who was poisoned and needs treatment. In other words, better safe than sorry.

Conclusion

Accuracy, precision, and recall are used to measure the performance of a classification machine learning model (there are other metrics for regression models, read more here). The metrics alone aren't enough to determine if your model is usable in real life scenarios. You must establish a baseline score and compare your model's performance against that baseline score. In a future tutorial, we'll show you how to assess a regression machine learning model; a model predicting a continuous numerical value (e.g. housing prices, forecasting sales, customer lifetime value, etc).

You can also check out:

- Building your first machine learning (ML) model
- How to interpret machine learning (ML) models with SHAP values
- A product developer's guide to machine learning (ML) regression model metrics

Check out our open source tool for building and running data pipelines that transform you data!

Join our Slack channel! Come, chat and collaborate with our Mage slack community.

AI/ML Developers

Every AI needs quality data

Kickstart your next project by creating pipelines that deliver better, more purposeful data.

View GitHub

Magical powers for data teams

GET STARTED

Documentation

Get instant help

COMMUNITY

GitHub

Slack

LinkedIn

Twitter