Machine Learning Terminology

This page holds a list of key machine learning terms and their definitions.

The name and location of the markdown file is "\my-repo\MachineLearningTerms.Rmd".

Types of Machine Learning

Supervised machine learning, which works with labeled data, and **unsupervised machine learning**, which works with unlabeled data.

Supervised machine learning falls into two categories - *classification* and *regression*. Classification algorithms predict the discrete classes (categories) to which samples belong. Regression models predict a continuous output.

Accuracy, Precision and Recall

Accuracy

Accuracy is a statistical measure which is defined as the quotient of correct predictions (both True positives (TP) and True negatives (TN)) made by a classifier divided by the sum of all predictions made by the classifier, including False positves (FP) and False negatives (FN).

Precision

Precision answers the question: "**What proportion of positive predictions were actually positive?**"

In other words, **Precision** is the ratio of the correctly identified positive cases to all the predicted positive cases, i.e. the correctly and the incorrectly cases predicted as positive.

When you are in a Precision mindset:

- You don't care about getting every positive sample correctly classified.
- But it's important that every positive prediction you get should actually be positive.

For instance, consider a book recommendation system. Say a positive prediction means you'd like the recommended book.

In a Precision Mindset:

- You are okay if the model does not recommend all good books in the world.
- But what it recommends should be good.

So even if this system recommended only one book and you liked it, this gives a Precision of 100%.

Recall

Recall answers the question: "**What proportion of actual positives was identified correctly by the model?**"

In other words, **Recall**, also known as *sensitivity*, is the ratio of the correctly identified positive cases to all the actual positive cases, which is the sum of the "False Negatives" and "True Positives".

When you are in a Recall Mindset:

- You care about getting every positive sample correctly classified.
- It's okay if some positive predictins were not actually positive.
- But all positive samples should get classified as positive.

For instance, consider an interview shortlisting system. A positive prediction means that the candidate should be invited for an interview.

In a Recall Mindset:

- You are okay if the model selects some incompetent candidates.
- But it should not miss out on inviting any skilled candidate.

So even if this system says that all candidates (good or bad) are fit for an interview, it gives you a Recall of 100%.

Conclusion

Which metric to choose entirely depends on what's important to the problem at hand:

Optimize **Precision** if:

- You care about getting **ONLY** quality positive predictions.
- You are okay if some quality (or positive) samples are left out.

Optimize **Recall** if:

- You care about getting **ALL** quality (or positive) samples correct.
- You are okay if some non-quality (or negative) samples also come along.