Machine Learning

|Precision and Recall

Prof. Dr. Mostafa Elhosseini

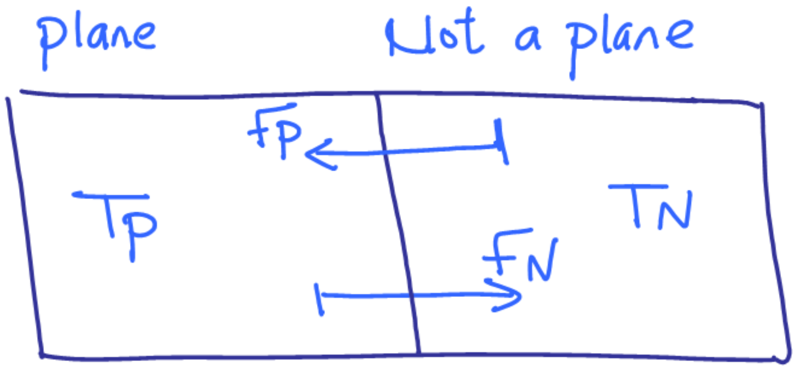
https://youtube.com/@ProfElhosseiniTechVerse

Agenda

- what is more important, Precision or Recall?
- Precision and Recall
- Harmonic mean
- F1 score

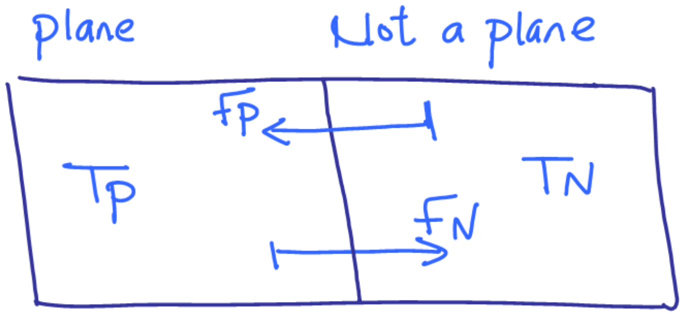
Radar Example

- Classify a point in space to be a plane "enemy" or not
  - All bordered are airplane
  - No false positive
  - $\boldsymbol{Precision} = \frac{TP}{TP+FP}$



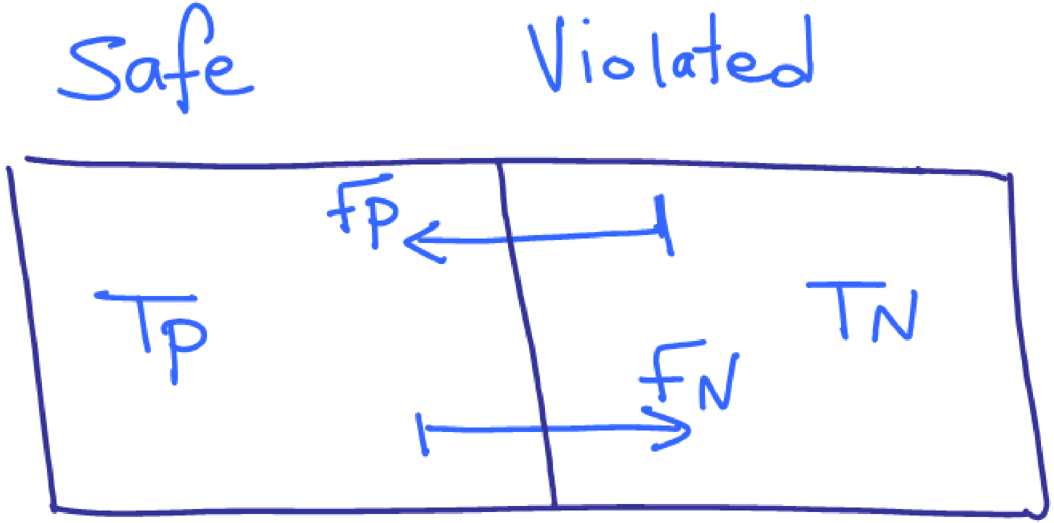
Radar Example

- Classify a point in space to be a plane "enemy" or not
  - Detect all enemy aircraft but, waste a lot of jet fuel
  - Fighter aircraft chase a non-existing enemy
  - $\textbf{Recall} = \frac{TP}{TP+FN}$



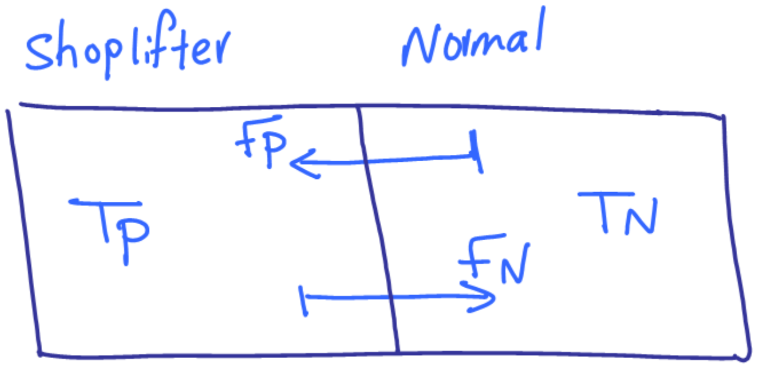
Safe videos for kids

- you prefer a classifier that rejects many good videos (low recall) but keeps only safe ones (high precision)
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$



Shoplifter on Surveillance Images

- Suppose you train a classifier to detect shoplifters on surveillance images: it is probably fine if your classifier has only 30% precision as long as it has 99% recall (sure, the security guards will get a few false alerts, but almost all shoplifters will get caught).
- $\textbf{Precision} = \frac{TP}{TP+FP}$
- $\textbf{Recall} = \frac{TP}{TP+FN}$



- You can not have it both ways: increasing precision reduces recall and vice versa

F_1 score

- It is often convenient to combine precision and recall into a **single metric** called the **F_1 score**, in particular if you need a **simple way to compare two classifiers**.
- The F_1 score is the **harmonic mean** of precision and recall
- $F_1 = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$
- $F_1 = \frac{2 \times Precision \times Recall}{Precision+Recall}$

Harmonic Mean

- The harmonic mean is a **type of average** that is calculated by dividing the number of elements (n) by the sum of the reciprocals of each element
- Harmonic Mean = $n/\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)$
- The most common application is in the calculation of the **F1 score**, which is the **harmonic mean of precision and recall**.

Harmonic Mean

- The harmonic mean is **more sensitive to smaller values** compared to the arithmetic mean. In the case of **imbalanced datasets**, where the model might have high precision but low recall (or vice versa), the harmonic mean (F1 score) provides a more balanced evaluation of the model's performance.
- Using the harmonic mean in the F1 score ensures that both precision and recall are taken into account and that the **model is penalized for having a low value in either of these metrics**. This encourages the development of models that perform **well on both** precision and recall, rather than favoring one over the other.

Harmonic Mean

- The F1 score **ranges from 0 to 1**, with **1 being the best** possible score, indicating perfect precision and recall.
- The reason for using the harmonic mean instead of the arithmetic mean is that it **gives more weight to lower values**. In the context of the F1 score, this means that if either precision or recall is low, the F1 score will be impacted more significantly.
- This is desirable because in many cases, having a high score in one metric while the other is low is not considered a good outcome.

Harmonic Mean

- The F1 score strikes a balance between precision and recall, providing a single metric that reflects the model's overall performance. It is especially useful when false positives and false negatives have similar costs or when the class distribution is imbalanced.

Example

- Certainly! Let's consider a binary classification problem where we have a dataset of 100 instances, and our model makes predictions for each instance. We'll compare the F1 score with the arithmetic mean of precision and recall to demonstrate the benefit of using the F1 score.

Scenario 1:

- True Positives (TP) = 40
- False Positives (FP) = 10
- False Negatives (FN) = 20
- True Negatives (TN) = 30

Example

Scenario 1:

- True Positives (TP) = 40
- False Positives (FP) = 10
- False Negatives (FN) = 20
- True Negatives (TN) = 30

In this scenario, the arithmetic mean is slightly higher than the F1 score.

- Precision = TP / (TP + FP) = 40 / (40 + 10) = 0.80
- Recall = TP / (TP + FN) = 40 / (40 + 20) = 0.67
- Arithmetic Mean = (Precision + Recall) / 2 = (0.80 + 0.67) / 2 = **0.735**
- F1 Score = $\frac{2 \times Precision \times Recall}{Precision + Recall}$ = 2 * (0.80 * 0.67) / (0.80 + 0.67) = **0.73**

Example

Scenario 2:

- True Positives (TP) = 20
- False Positives (FP) = 5
- False Negatives (FN) = 40
- True Negatives (TN) = 35

In this scenario, the F1 score is significantly lower than the arithmetic mean

This is because the model has a **high number of false negatives**, resulting in **low recall**. The F1 score penalizes this more heavily compared to the arithmetic mean.

- Precision = TP / (TP + FP) = 20 / (20 + 5) = 0.80
- Recall = TP / (TP + FN) = 20 / (20 + 40) = 0.33
- Arithmetic Mean = (Precision + Recall) / 2 = (0.80 + 0.33) / 2 = 0.565
- F1 Score = $\frac{2 \times Precision \times Recall}{Precision + Recall}$ = 2 * (0.80 * 0.33) / (0.80 + 0.33) = **0.47**

Example

- The benefit of using the F1 score becomes evident in Scenario 2. While the arithmetic mean is relatively high (0.565), it doesn't properly reflect the model's poor performance in terms of recall. The F1 score, on the other hand, provides a more balanced assessment by giving more weight to the lower recall value.
- In situations where both precision and recall are important, and we want to avoid models that have a high score in one metric while performing poorly in the other, the F1 score is a more appropriate choice. It encourages models to have a balance between precision and recall, penalizing cases where one metric is significantly lower than the other.

Example

- By using the harmonic mean, the F1 score ensures that models with imbalanced performance (high precision but low recall, or vice versa) receive a lower score compared to models that have a more balanced performance in both metrics.
- The arithmetic mean is sensitive to extreme values, meaning that if one of the values (precision or recall) is significantly higher or lower than the other, it can skew the average.
- The harmonic mean, on the other hand, is more sensitive to smaller values. If either precision or recall is low, the harmonic mean will be pulled down more significantly compared to the arithmetic mean.

Next Lecture

- Precision/ Recall Tradeoff