Evaluation Metrics

David Robinson

Metrics

Accuracy measures the number of correct predictions relative to the total predictions.

$$\textbf{Accuracy} = \frac{TP + TN}{P + N}$$

Precision measures how many of the predicted positives are actually positive.

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Recall measures the ability to identify all relevant instances.

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Generalized Fβ-Measure

F_β allows you to adjust the balance between precision and recall, where $\beta > 1$ puts more emphasis on recall and $\beta < 1$ puts more emphasis on precision.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

F1-Score is the harmonic mean of precision and recall, which gives more weight to low values.

Paired t-Test

The **paired t-test** compares two related groups, like scores or results from two different methods applied to the same dataset.

1. Calculate the difference between each pair of scores
2. Calculate the mean of the differences
3. Find the standard deviation of the differences
4. Calculate the t-statistic
$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$
5. Compare the t-statistic to a critical value from a t-distribution table and the difference is significant if the t-statistic is larger than the critical value