Comprehensive Revision Notes on Classification Metrics

Overview

This set of notes covers key concepts and metrics critical to understanding and evaluating classification models. It specifically focuses on issues like imbalanced data, the inadequacy of accuracy as a sole measure, and the application of confusion matrix-related metrics such as precision, recall, and the F1 score.

Imbalanced Data

Imbalanced data is a common issue in classification tasks, where the distribution of the classes is unequal. For instance, if there are significantly more instances of one class than another, it can lead to biased model performance, where the model might predict the dominant class more often, which appears accurate but isn't effective for real-world applications .

Identifying Imbalance

- **Example**: If there are 100 instances of spam and 900 instances of non-spam in a dataset, this is considered imbalanced .

Handling Imbalanced Data

- Techniques such as oversampling the minority class or utilizing different models that can handle such imbalances are discussed .

Accuracy

- **Definition**: The ratio of the number of correct predictions to the total predictions made .
- **Limitation**: While a model might have high accuracy, it can be misleading in the case of imbalanced datasets as it might classify all samples as the

Confusion Matrix

A confusion matrix is a tool to visualize the performance of a classification model with counts of true positive, false positive, true negative, and false negative .

- **True Positive (TP)**: Correctly predicted positive observations.
- **False Positive (FP)**: Incorrectly predicted as positive.
- **False Negative (FN)**: Incorrectly predicted as negative.
- **True Negative (TN)**: Correctly predicted negative observations .

Precision

- **Definition**: Precision is the ratio of correctly predicted positive observations to the total predicted positives【4:4†transcript.txt】 .
- **Importance**: High precision means that there are very few false positives .
- **Formula**: $Precision = \frac{TP}{TP+FP}$

Recall

- **Definition**: Recall (also known as Sensitivity) measures the ratio of correctly predicted positive observations to all actual observations in that class .
- **Importance**: High recall indicates that the model is capturing most of the positive class cases .
- **Formula**: $Recall = \frac{TP}{TP+FN}$

F1 Score

- **Purpose**: Combines precision and recall into a single metric by calculating their harmonic mean .
- **Importance**: Useful when both precision and recall are needed to be balanced .
- **Formula**: $F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Practical Examples

- **Cancer Detection**: Recall is crucial because missed positive cases (false negatives) need to be minimized .

These notes provide a foundational understanding of how to evaluate classification models beyond simple accuracy, using metrics that offer a deeper insight into model performance, especially in scenarios dealing with imbalanced data.