Chapter 1

Evaluation Binary Classifiers

Accuracy, recall, precision, and the F1 score are widely used metrics to measure and compare the performance of binary classifiers. This chapter will delve into these evaluation measures, providing insights into their interpretation and practical applications.

1.0.1 Binary Classification

Before diving into the evaluation metrics, let's clarify the concept of binary classification. In binary classification, we aim to assign each instance in a dataset to one of two mutually exclusive classes. For example, classifying emails as spam or not spam, identifying whether a patient has a specific medical condition or not, or predicting whether a credit card transaction is fraudulent or legitimate are common binary classification tasks.

To evaluate the performance of a binary classifier, we need metrics that can provide insights into how well the classifier performs in distinguishing between the two classes.

1.0.2 Accuracy

Accuracy is a widely used metric for evaluating binary classifiers. It measures the overall correctness of the classifier's predictions by calculating the ratio of correctly classified instances to the total number of instances in the dataset. Mathematically, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

While accuracy provides a general overview of the classifier's performance, it may not be sufficient in certain scenarios. This is especially true when the dataset is imbalanced, meaning that one class significantly outweighs the other in terms of the number of instances.

1.0.3 Precision and Recall

Precision and recall are evaluation metrics that provide insights into the classifier's performance on specific classes, allowing us to identify potential trade-offs between false positives and false negatives.

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). It can be expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). Mathematically, recall can be represented as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision and recall are complementary metrics. Precision focuses on the quality of positive predictions, while recall emphasizes the classifier's ability to identify positive instances. The choice between precision and recall depends on the specific requirements of the problem at hand.

1.0.4 F1 Score

The F1 score combines precision and recall into a single metric, providing a balanced evaluation measure that considers both false positives and false negatives. It is the harmonic mean of precision and recall, and it can be calculated as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score ranges between 0 and 1, where a value of 1 represents perfect precision and recall. It is particularly useful when we want to strike a balance between precision and recall, considering both the false positives and false negatives in the classifier's predictions.

This is a draft chapter from the Kontinua Project. Please see our website (`https://kontinua.org/`) for more details.

Appendix A

Answers to Exercises

Index