Assignment 18: Accuracy Assessment Metrics

75 Points scaled to 20 Points

Introduction

This exercise will explore metrics for assessing and comparing multiclass classification, binary classification, probabilistic, and regression models. These datasets were derived from some of my own research projects. However, they are meant to be generic here, so you don't need to worry about the context or use case.

Objectives

- *Calculate assessment metrics from predictions made to withheld test or validation data*
- *Interpret metrics to assess model performance*
- *Interpret metrics to compare models*
- *Determine when specific metrics are appropriate measures for assessing certain models*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and results embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and answered within Markdown cells.*

Background Questions

Question 1: Why is it important to use withheld validation data when assessing models as opposed to the data used to train the model? (5 Points)

Question 2. Explain the difference between Overall Accuracy and Kappa for assessing a multiclass classification. (5 Points)

Question 3. Explain the difference between Precision, Recall, F1-Score, Specificity, and Negative Predictive Value for assessing a binary classification. (5 Points)

Question 4. Explain the difference between the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve for assessing a probabilistic prediction. (5 Points)

Question 5. Explain the meaning and interpretation of the area under the curve (AUC) measure calculated from the Receiver Operating Characteristic Curve. (5 Points)

Question 6. Explain how R-squared is calculated. What is the unit of measurement of R-squared? (5 Points)

Question 7. Explain how Root Mean Square Error (RMSE) is calculated. What is the unit of measurement? (5 Points)

Tasks and Questions

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following analyses and answer the stated questions.

Multiclass Classification

Task 1. The "classification_data.csv" dataset contains three columns and relates to a multiclass classification where five classes are differentiated. Each class has been assigned a numeric value (1 through 5), but these values represent categories or nominal data. The "truth" column represents the correct or reference class while the "m1" and "m2" columns represent separate predictions. Generate Confusion Matrices from the reference data and predictions obtained using the different models. Next, calculate the Overall Accuracy and Kappa statistic for both models. Based on these results, which model provided the best performance? What classes were most difficult to differentiate? (10 Points)

Task 2. The "classification_data2.csv" dataset represents classification results for differentiating six forest community types across the state of West Virginia using variables derived from satellite data and digital elevation data. The "Fall_T" models used only imagery collected in the fall along with digital terrain variables; the "Spr_T" models were produced using spring imagery and digital terrain variables; and the "Sum_T" models were produced using summer imagery and digital terrain variables. The "H_T" model used imagery from multiple seasons, which were summarized using a process called harmonic regression, alongside the digital terrain variables. The "set" column defined which dataset was used while the "OA" column provides the Overall Accuracy, and the "k" column provided the Kappa statistic. Each of the for sets of variables were used to train 50 different models using different training and testing data partitions in order to capture the variability in model performance when given different input samples. Create two grouped boxplots. The first plot should show the distribution of Overall Accuracy for each of the replicate sets as four separate boxplots. The second should show Kappa as opposed to Overall Accuracy. Based on these results, discuss the relative performance of the different sets of feature and also the variability in performance when different data partitions were used. (10 Points)

Binary Classification and Probabilistic Prediction

Task 3. The "binary_data.csv" dataset represents results for predicting the likelihood of landslide occurrence as predicted using a random forest model. The "truth" column represents the correct classification ("not" = not a landslide; "slopeD" = a landslide). The "predicted" column represents the predicted class. The "prob_not" column represents the predicted probability of the location not being a landslide while the "prob_fail" column represented the predicted probability of the site being a landslide. There are an equal number of "not landslide" and "landslide" examples based on the reference classification. The "slopeD" class should be treated as the positive outcome. From these data and using the classification results and reference classification calculate the following metrics: Precision, Recall, F1-Score, Specificity, and Negative Predictive Value. From these data and using the reference classification and predicted class probabilities, calculate the Area Under the ROC and PR Curves. Discuss the results in regards to the model performance. What was the dominant source of error in the model: false positives or false negatives? (10 Points)

Binary Classification and Probabilistic Prediction

Task 4. The "regression_data.csv" dataset contains 156 values ("truth") and associated predictions using three different models ("m1", "m2", and "m3"). Calculate the R-Squared and RMSE metrics for each of the three models. How do these models compare in regards to accuracy and the amount of variability in the value of interest explained? (10 Points)