CSCI316: Week 5 - Classification Model Evaluation

📊Overview

Evaluating classification models is essential to ensure that a model not only performs well on training data but also generalizes to unseen data. This week's topic covers how to measure a classifier's performance using a variety of **evaluation metrics**, **validation techniques**, and **model comparison tools**.

🔗1. Evaluation Concepts

- **Positive tuples (P)**: Main class of interest (e.g., `buys_computer = yes`)
- **Negative tuples (N)**: All other classes
- **True Positives (TP)**: Positives correctly classified
- **True Negatives (TN)**: Negatives correctly classified
- **False Positives (FP)**: Negatives wrongly classified as positives
- **False Negatives (FN)**: Positives wrongly classified as negatives

Confusion Matrix: | | Predicted Positive | Predicted Negative | |----------|----------|----------| | Actual Positive| TP | FN | | Actual Negative| FP | TN |

📊2. Evaluation Metrics

- **Accuracy**: Proportion of correctly classified tuples.
- `Accuracy = (TP + TN) / (TP + TN + FP + FN)`
- **Error Rate**: Complement of accuracy.
- `Error Rate = 1 - Accuracy = (FP + FN) / Total`
- **Sensitivity (Recall)**: Ability to identify positive tuples.
- `Recall = TP / P`
- **Specificity**: Ability to identify negative tuples.
- `Specificity = TN / N`
- **Precision**: Proportion of predicted positives that are actually positive.
- `Precision = TP / (TP + FP)`
- **F1 Score**: Harmonic mean of precision and recall.
- `F1 = 2 * (Precision * Recall) / (Precision + Recall)`
- **Fβ Score**: Weighted F-measure for precision-recall trade-off.
- `Fβ = (1 + β^2) * (Precision * Recall) / ((β^2 * Precision) + Recall)`

⚖️3. Model Validation Techniques

- **Holdout Method**:
- Split data into training (e.g., 2/3) and test (e.g., 1/3) sets.
- **Random Subsampling**:

• Repeat holdout multiple times and average results.
• **K-Fold Cross Validation**:
• Partition data into `k` equal parts. Train on `k-1` folds, test on 1.
• Repeat `k` times; average performance.
• **Stratified K-Fold**: Maintain class distribution across folds.
• **Leave-One-Out (LOO)**:
• Special case where `k = number of tuples`.

🥌4. Comparing Classifiers

• **Why Compare?**: Determine if one model significantly outperforms another.
• **T-test for Statistical Significance**:
• Use k-fold cross-validation to compute error differences `d1, d2, ..., dk`
• Compute t-statistic using:
`t = mean(d) / (std(d) / sqrt(k))`
• Null Hypothesis (H0): Models are the same.
• Reject H0 if `t` falls in the rejection region (based on t-distribution).

🕐5. ROC Curve (Receiver Operating Characteristic)

• **ROC Curve**: Plots True Positive Rate (TPR = Sensitivity) vs. False Positive Rate (FPR = 1 - Specificity).
• **AUC (Area Under Curve)**:
• AUC = 1.0: Perfect classifier
• AUC = 0.5: Random guessing (diagonal line)
• AUC closer to 1 is better

🗜️6. Other Model Considerations

• **Speed**:
• Training time
• Inference time
• **Robustness**: Ability to handle noise or missing data
• **Scalability**: Efficiency on large datasets
• **Interpretability**: How well the model can be understood by humans
• **Model Compactness**: Simplicity of rules (e.g., tree size)

Let me know if you'd like this as a downloadable PDF or want quizzes and exercises!