CS60020 : Foundations of Algorithm Design and Machine Learning

Performance Evaluation

Scribed by: Sayali Malshikare (24BM6JP50)

25 March 2025

1 Introduction

In Supervised Learning, we have studied multiple classifier algorithms such as Decision Tree, Artificial Neural Network , Support Vector Machines. Each of these approaches will give certain kind of decision boundaries, among which we need to choose which is the best hypothesis among all.

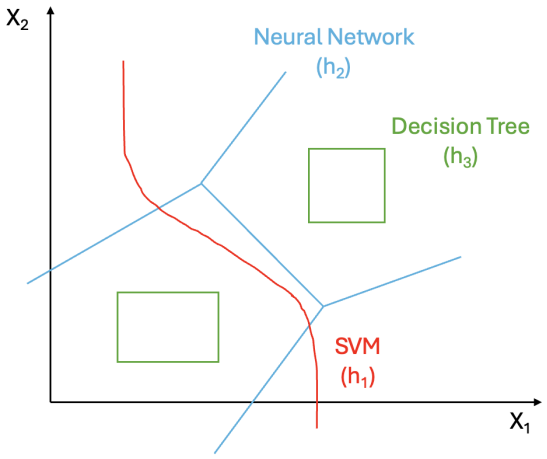

Figure 1: Classifier Hypothesis Set

2 Evaluation of Hypothesis

1. Metrics : How accurately the unknown training test data is classified
2. Methods : How to estimate such metrics
3. Comparison of Models : How do we compare the models with respect to these estimated metrics

2.1 Performance Metrics

2.1.1 Confusion Matrix

One of the performance metrics used is Confusion Matrix constructed from the results of the test data. Confusion matrix for a 2-Class Classification is given below:

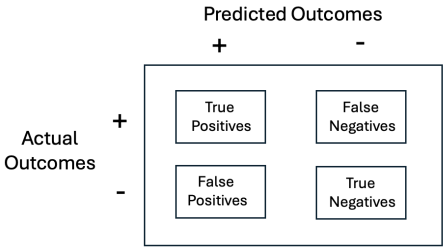

Figure 2: Confusion Matrix

For a k-class classification, we will observe that the resulting matrix will be of the order of k * k, and the diagonal of the matrix will be the set of correct predictions.

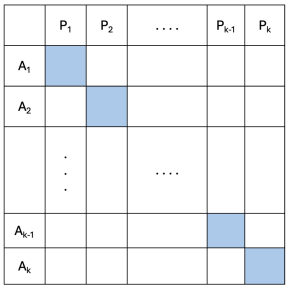

Figure 3: Confusion matrix for K- Class

2.1.2 Metrics

1. **Accuracy**: Out of the total predictions, it represents how many predictions were actually correct. It is the ratio of correctly predicted observations to the total number of predictions made.

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{1}$$

   Fallacy of this metric: For example, if the test data is biased towards a particular class. If the test data has 9990 positives and 10 negatives. Suppose that the classifier function blindly classifies all cases as positive, it will still be 99.9% accurate. This metric does not take into consideration the deviation ratios. Hence, Precision and Recall used.

2. **Precision** : It represents how precise the predictions are. It measures the proportion of positive predictions that were actually correct.

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{2}$$

3. **Recall** : also known as Sensitivity or True Positive Rate, measures the proportion of actual positive instances that were correctly identified by the model. It represents how well we are recalling the prediction with respect to actual class.

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{3}$$

4. **F Score** : Since Precision and Recall have their own merits and demerits when we use them individually, we define F-Score as a combination of Precision and Recall. It provides a single metric that balances both false positives and false negatives. F-Score (F) is taken as a Harmonic Mean of Precision and Recall.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

5. **Weighted Accuracy** : Precision is biased towards TP and FP values. Recall is biased towards TP and FN values. F Score is biased towards TP, FN and FP values. TN is still neglected. To overcome the biases in the above Metrics, we use a Weighted Average, where we give weights to all the parameters, as

$$WeightedAccuracy = \frac{w_1 \cdot |TP| + w_4 \cdot |TN|}{w_1 \cdot |TP| + w_2 \cdot |FP| + w_3 \cdot |FN| + w_34 \cdot |TN|} \tag{5}$$

These weights are hyper-parameters and can be optimized during cross validation. All the metrics as discussed above can also be represented using the Weighted Accuracy by adjusting the weights.

(a) Accuracy, when W1 = W2 = W3 = W4 = 1

(b) Precision, when W1 = W3 = 1 and W2 = W4 = 0

(c) Recall, when W1 = W2 = 1 and W3 = W4 = 0

(d) F-Score,when W1 = W2 = W3 = 1 and W4 = 0

2.2 Methods of Estimation

If we plot accuracy with respect to the Sample Size, we observe that the accuracy increases with increase in Sample size and gets saturated at higher levels (95%). This curve is called the Learning Curve. The method of estimation is hence dependent on

1. Class Distribution
2. Size of Training and Test Sets - If the training set distribution is thin, the training set will either bias the estimate, or if test set is very thin we might get a high variance of the observe metrics and it will not represent all the scenarios.
3. Cost of Misclassification

2.2.1 Holdout Method:

Usually we keep 2/3rd data for training and 1/3rd data as Validation.

2.2.2 Random Sub-Sampling:

Smapling in which randomly choose which sample we would take in the training set and which in the testing set. From the random sampling, came the concept of **Stratified Sampling**. Suppose we have some Ys and some Ns in the test data, we segregate our test data into 2 Stratas - Ys and Ns. We call each of these as a Strata. Now we apply random sampling on these Stratas to make the test data.

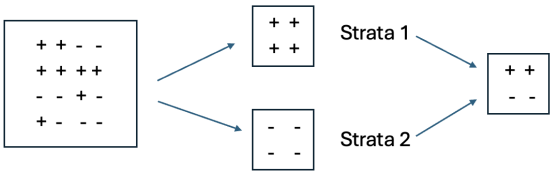

Figure 4: Stratified Sampling

Bootstrapping is an extension of stratified sampling that generates new samples by drawing instances from the stratas with replacement.

2.2.3 K-fold Cross Validation:

In K-fold cross-validation, the dataset X is divided randomly into K equalcross-validation sized parts, Xi, i = 1,...,K. To generate each pair, we keep one of the K parts out as the validation set and combine the remaining K - 1 parts to form the training set. Doing this K times, each time leaving out another one of the K parts out, we get K pairs.

2.3 Model Comparison

2.3.1 ROC Curve (Receiver Operating Characteristic)

The ROC Curve (Receiver Operating Characteristic Curve) is a graphical plot used to evaluate the performance of a binary classifier as its decision threshold is varied. It illustrates the trade-off between the **True Positive Rate (Recall)** and the **False Positive Rate**.

$$TPR = \frac{|TP|}{|TP| + |FN|} \tag{6}$$

$$FPR = \frac{|FP|}{|FP| + |TN|} \tag{7}$$

In the ideal scenario, we would desire a TPR of 1.0 and an FPR of 0.0

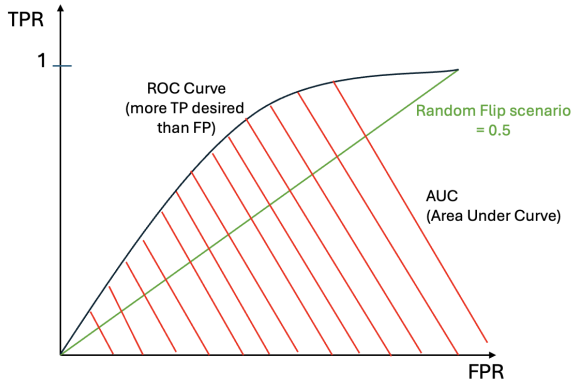

Figure 5: ROC Curve

- AUC = 1.0 indicates a perfect classifier. (ideal)
- AUC = 0.5 suggests no discriminative power (random guessing).
- The higher the AUC, the better the model distinguishes between classes.

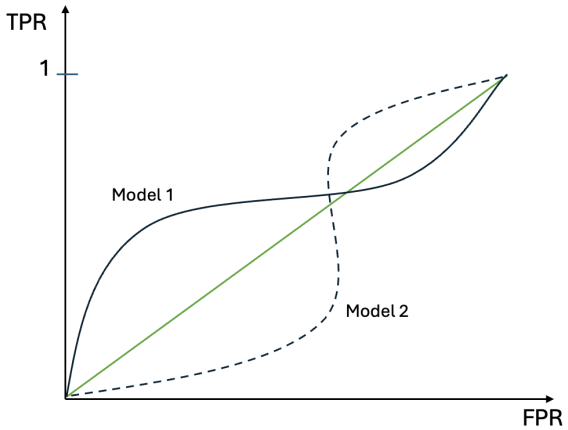

Figure 6: Model Comparison

From Figure 6, we observe that M2 is better when the FPR is high while M1 is better when FPR is low. Thus There is a tradeoff between TPR and FPR.

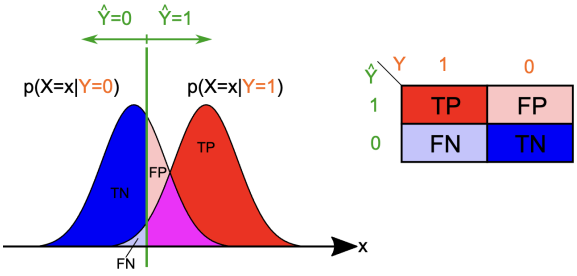

Figure 7: ROC Curve Analysis

Adjusting the threshold directly influences both the True Positive Rate (TPR) and the False Positive Rate (FPR).
When the threshold is lowered, more samples are classified as positive. As a result, the TPR increases because more true positive instances are correctly identified. However, this also leads to an increase in the FPR, since more negative instances are mistakenly classified as positive. The classifier becomes more sensitive but less specific in this case.
Conversely, when the threshold is increased, fewer samples are classified as positive. This reduces the FPR, as fewer negative instances are misclassified. However, the TPR also decreases, since more positive instances are missed. The classifier becomes more conservative, favoring specificity over sensitivity.
Choosing an appropriate threshold involves balancing the trade-off between TPR and FPR based on the application context. In situations where false negatives are costly, a lower threshold may be preferred, while in cases where false positives must be minimized, a higher threshold is more appropriate.
The ROC curve is a useful tool for visualizing this trade-off and identifying the threshold that offers the best compromise between sensitivity and specificity.