Analyzing Performance of Classifiers in Machine Learning

Prof. Buckmire (ron@oxy.edu), Prof. Basu (basu@oxy.edu)

Name (Print) :

Help (Name Individuals or Websites) :

Homework

1. The outcomes True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are the primary building blocks of the metrics that are used to evaluate the performance of a classification model. **Define them below.**

   True Positive:

   True Negative:

   False Positive:

   False Negative:

2. **Restate each of the above outcomes in the context of a classification model that separates emails into two categories: "spam" or "not spam".** Assume "spam" emails to be the positive class and "not spam" emails to be the negative class.

   True Positive:

   True Negative:

   False Positive:

   False Negative:

3. For a balanced data set the standard metric for evaluating classification models is *accuracy*. Accuracy is defined as the is the fraction of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Suppose a classifier is trained using 100 emails of which some are "spam" and others are "not spam". **Using the confusion matrix presented as Table 1, below calculate the accuracy of the model.**

Table 1: Confusion matrix

	Predicted "spam"	**Predicted "not spam"**
"spam"	1	8
"not spam"	1	90

In your opinion is the model doing a great job of identifying "spam" emails?

Let's do a closer analysis of positives and negatives to gain more insight into our model's performance.

4. **Of the 100 emails, how many emails are actually "spam" and how many are actually "not spam"?**

5. **Of the emails that are actually "not spam", how many does the model correctly identify as "not spam"?**

6. **What is your opinion regarding the performance of the model as it pertains to identifying emails that are "not spam"?**

7. **Of the emails that are actually "spam", how many emails does the model correctly identify as "spam"?**

8. **What is your opinion regarding the performance of the model as it pertains to identifying emails that are "spam?**

As demonstrated through Questions 1-8, accuracy alone doesn't tell the full story when you are working with a **class-imbalanced data set**, like this one, where there is a significant disparity between the number of positive and negative labels. You will now proceed to look at two better metrics for evaluating class-imbalanced problems: precision and recall.

9. **Precision** attempts to answer the question: What proportion of positive identifications was correct? Mathematically, precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Compute the precision of the email classification model and write a sentence concluding the significance of your finding.

10. **Recall** attempts to answer the question: What proportion of actual positives was correctly identified? Mathematically, recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Compute the recall of the email classification model and write a sentence describing the significance of your finding.

To fully evaluate the effectiveness of a model, we must examine **BOTH precision and recall**. However, **precision and recall are often in tension**. That means, improving precision typically reduces recall and vice versa. You will explore this notion below by looking at the figure below, which shows 30 predictions made by an email classification model. Those to the right of the classification threshold are classified as "spam", while those to the left are classified as "not spam".

Note that a classification threshold is a scalar-value criterion that is applied to a model's predicted score in order to separate the positive class from the negative class.

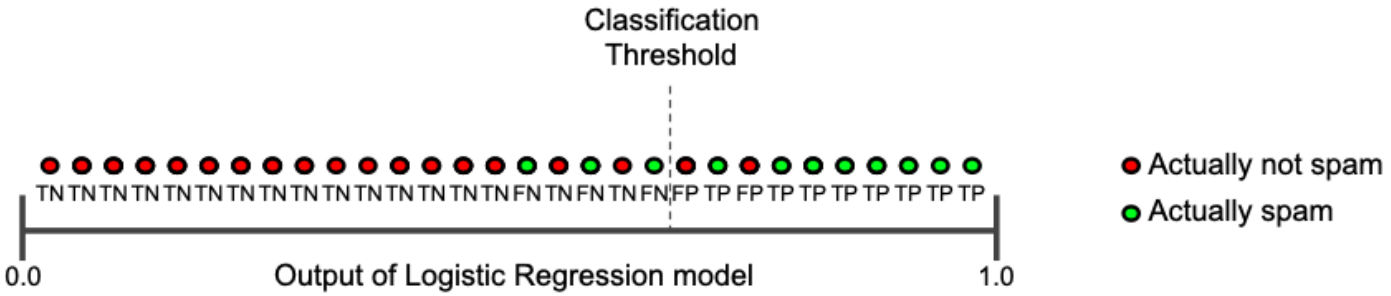

Figure 1: Classifying email messages as spam or not spam

11. **Based on the model represented by Figure 1, complete the confusion matrix below.**

Table 2: Confusion Matrix For Figure 1

	Predicted "spam"	**Predicted "not spam"**
"spam"		
"not spam"		

12. **Compute the precision and recall for the above classifier.**

13. Suppose we **increase the classification threshold** (from its original position in Figure 1). **Using Figure 2 below, complete the corresponding confusion matrix.**

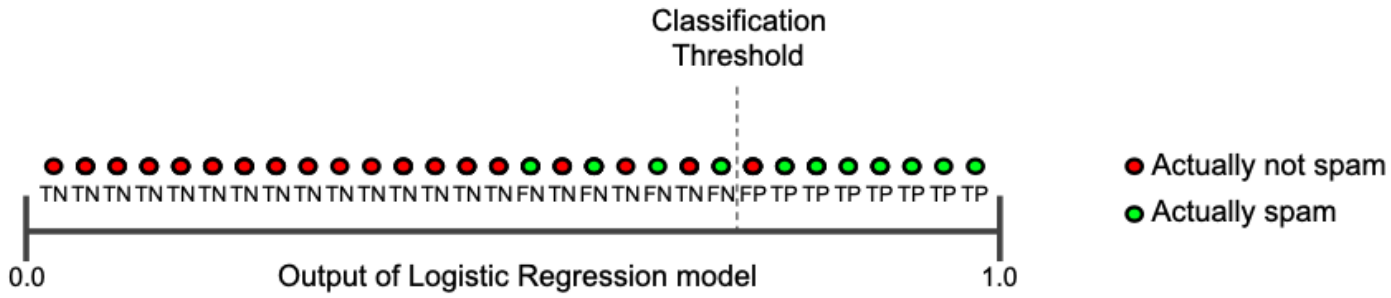

Figure 2: Classifying email messages as spam or not spam

Table 3: Confusion Matrix For Figure 2

	Predicted "spam"	**Predicted "not spam"**
"spam"		
"not spam"		

14. **Compute the precision and recall for the classifier represented in Figure 2. Write a sentence explaining why precision increases but recall decreases.**

15. Let us now examine the effect of **decreasing the classification threshold** (from its original position in Figure 1).

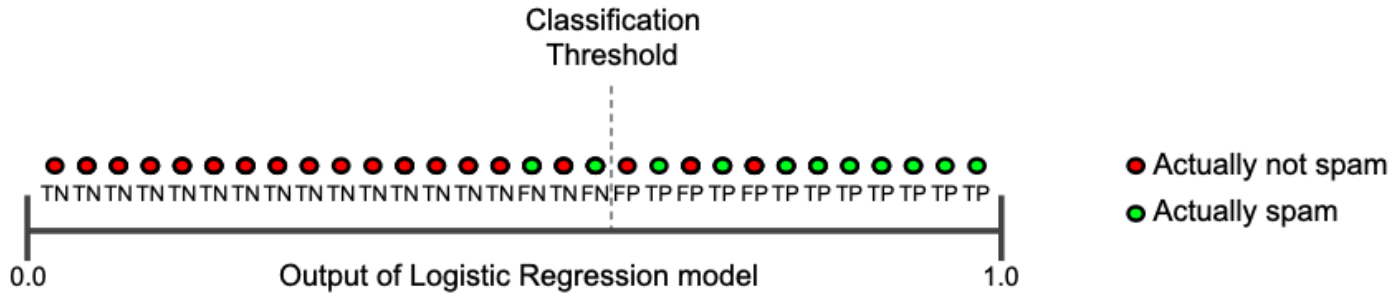

Figure 3: Classifying email messages as spam or not spam

16. **Complete the corresponding confusion matrix.**

Table 4: Confusion Matrix For Figure 3

	Predicted "spam"	**Predicted "not spam"**
"spam"		
"not spam"		

17. **Compute the precision and recall for the classifier represented in Figure 3. Write a sentence explaining why precision decreases but recall increases.**

18. **Precision versus Recall:** For some classification models it is detrimental for false positives to occur but in other models false negatives are more damaging. **In the case of the email classifier model, what is more important to control for or prevent: false positives or false negatives? In other words, which evaluation metric would you prefer to be higher: precision or recall?**

Other Evaluation Metrics

19. The F_β **score** is the weighted harmonic mean of precision and recall, reaching its optimal value at 1 and its worst value at 0.

$$F_\beta \text{ score} = (1+\beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}} \tag{1}$$

The β parameter determines the weight of precision in the combined score, i.e, $\beta < 1$ lends more weight to precision, and $\beta > 1$ favors recall. When $\beta = 1$ the resulting F_1 score is the harmonic mean of precision and recall and thus must lie between them. Further, $\beta = 0$ considers only precision and $\beta \to \infty$ considers only recall. F_β-Score provides a way to combine both precision and recall into a single measure that captures both properties.

$$Precision : F_0 \longleftarrow F_{0.5} \longleftarrow F_1 \longrightarrow F_2 \longrightarrow Recall : F_\infty$$

Figure 4: F_β score diagram.

Compute the $F_{.5}$ score corresponding to the model represented in Figure 2.

20. **AUC Score** is another metric used to evaluate the performance of a classification model.

An **ROC curve (receiver operating characteristic curve)** is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: **True Positive Rate** & **False Positive Rate**.

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$\text{TPR} = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

$$\text{FPR} = \frac{FP}{FP + TN}$$

An ROC curve plots TPR versus FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive (see Figure 3), thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

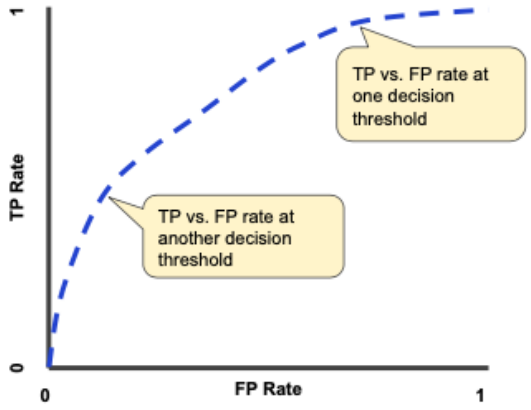

Figure 5: Typical ROC Curve

To compute the points in an ROC curve, one would have to evaluate a classification model many times with different classification thresholds, but this is inefficient. Fortunately, there's an efficient, algorithm that can provide this information for us, called **AUC Score**.

Compute the False Positive Rate corresponding to the email classifiers represented by Figure 1, Figure 2, and Figure 3.

AUC stands for "Area under the ROC Curve." That is, AUC Score measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

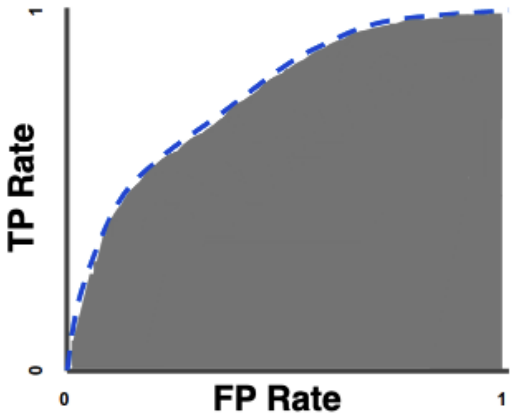

Figure 6: AUC Score

AUC Score provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. A model whose predictions are all wrong has an AUC Score of 0; one whose predictions are always correct has an AUC Score of 1. AUC Score is desirable for the fact that it is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen, whereas the metrics of precision, recall and F_β values depend on the classification threshold as demonstrated in Questions 11-17.

21. **AUC Score by definition involves computing an area under a curve. Explain why one can't use Calculus, specifically the fundamental theorem of Calculus, to compute the AUC Score of a classifier. What methods are used instead?**