Evaluation of Machine Learning Systems Review

Akhil Vasvani

April 2019

1 Questions

Exercise 1. What is False negative, False positive, True negative and True positive?

Proof. A **true positive (TP)** is an outcome where the model correctly predicts the positive class. A **true negative (TN)** is an outcome where the model correctly predicts the negative class. A **false positive (FP)** is an outcome where the model incorrectly predicts the positive class. And a **false negative (FN)** is an outcome where the model incorrectly predicts the negative class.

Example. Let's consider a scenario of a fire emergency:

1. **True Positive**: If the alarm goes on in case of a fire.

   Fire is positive and prediction made by the system is true.

2. **False Positive**: If the alarm goes on, and there is no fire.

   System predicted fire to be positive which is a wrong prediction, hence the prediction is false.

3. **False Negative**: If the alarm does not ring but there was a fire.

   System predicted fire to be negative which was false since there was fire.

4. **True Negative**: If the alarm does not ring and there was no fire.

   The fire is negative and this prediction was true.

□

Exercise 2. What are accuracy, sensitivity, specificity, ROC?

Proof. **Receiver Operating Characteristic** curve (ROC curve) is a fundamental tool for diagnostic test evaluation and is a plot of the true positive rate (**Sensitivity** or hit rate) against the false positive rate (**Specificity**) for the different possible cut-off points of a diagnostic test in binary classification.

It shows the trade-off between sensitivity and specificity How to interpret the model:

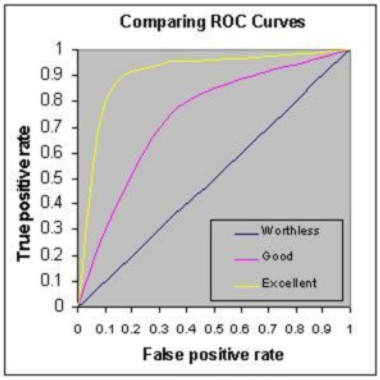

Figure 1: ROC curve

- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. In laymen's terms, the model is able to accurately predict class 1 as 1 and class 0 as 0. This means the model has a *good measure of separability*.
- The closer the curve comes to the 45° diagonal of the ROC space, the less accurate the test. This means the model has the *no class separation capacity* whatsoever.
- The slope of the tangent line at a cut-point gives the (positive) likelihood ratio (LR) ($\frac{TPR}{FPR}$) for that value of the test.
- The area under the curve is a measure of test accuracy, so larger the area under the curve the better the model is at differentiating between two classes.
- Note: If the area under the curve is near to 0, then the model is reciprocating the result. In order words, it is predicting class 1 as 0 and class 0 as 1. Therefore, this is a poor model which has the *worst measure of separability*.

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{Sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

Note. False Positive Rate (FPR)

$$= \frac{FN}{TN + TP} = 1 - \text{Specificity}.$$

Sensitivity and specificity are inversely proportional to each other, so an increase in sensitivity will be accompanied by a decrease in specificity. If you increase the threshold, there will be more negative values → specificity and lower sensitivity. Via versa, if we decrease the threshold, there will be more positive values → increase sensitivity and decrease specificity.

True Positive Rate (TPR) =

$$\frac{TP}{TP + FN}.$$

TPR and FPR are proportional.

Accuracy is another metric for evaluating classification models, which is the fraction of predictions the model got right over all the total number of number of predictions. Formally, for binary classification accuracy has the following definition:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

□

Exercise 3. What are precision and recall?

Proof. **Precision** is the fraction of detections reported by the model that were correct, while **recall** is the fraction of true events that were detected.

Example. Imagine that, your girlfriend gave you a birthday surprise every year for the last 10 years. One day, your girlfriend asks you: 'Sweetie, do you remember all the birthday surprises from me?' To stay on good terms with your girlfriend, you need to recall all the 10 events from your memory. Therefore, **recall is the ratio of the number of events you can correctly recall, to the total number of events**. If you can recall all 10 events correctly, then, your recall ratio is 1.0 (100%) and if you can recall 7 events correctly, your recall ratio is 0.7 (70%)

However, you might be wrong in some answers. For example, let's assume that you took 15 guesses out of which 10 were correct and 5 were wrong. This means that you can recall all events, but not so precisely. Therefore, **precision is the ratio of a number of events you can correctly recall, to the total number of events you can recall (mix of correct and wrong recalls)**. From the above example (10 real events, 15 answers: 10 correct, 5 wrong), you get 100% recall but your precision is only 66.67% (10 / 15).

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Note. When using precision and recall, it is common to plot a **PR curve**, with precision on the y-axis and recall on the x-axis. The classifier generates a score that is higher if the event to be detected occurred. For example, a feedforward network designed to detect a disease outputs $\hat{y} = P(y = 1|\boldsymbol{x})$ estimating the probability that a person whose medical results are described by features $\boldsymbol{x}$ has the disease. We choose to report a detection whenever this score exceeds some threshold. By varying the threshold, we can trade precision for recall.

Note. In many cases, we wish to summarize the performance of the classifier with a single number rather than a curve. To do so, we can convert precision p and recall r into an **F-score** (or **F1-score**) given by

$$F = \frac{2pr}{p + r}.$$

□

Exercise 4. What is a Confusion Matrix?

Proof. A confusion matrix or an error matrix is a table which is used for summarizing the performance of a classification algorithm.

	Predicated: YES	Predicted: NO	
Actual: YES	TP	FN (β)	$\longleftrightarrow Sensitivity$
Actual: NO	FP (α)	TN	$\longleftrightarrow Specificity$
	$\downarrow$ PPV	$\downarrow$ NPV	

Note. Positive Predictive value (PPV) is the precision and Negative Predictive value (NPV) is

For general confusion matrix metric, our "class of interest" is the positive class–the item of interest. We can see how along the rows, one can find sensitivity and specificity, and along the columns one can find the PPV and the NPV. □

Exercise 5. What is the difference between Type I and Type II error?

Proof. **Type I error**, also known as a **"false positive"**: the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. Plainly speaking, it occurs when we are observing a difference when in truth there is none (or more specifically—no statistically significant difference). So the probability of making a type I error in a test with rejection region R is $P(R|H_0$ is true).

Type II error, also known as a **"false negative"**: the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. Plainly speaking, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is $1 - P(R|H_a$ is true). The power of the test can be $P(R|H_a$ is true). Let's show an example to illustrate this concept.

Example. Plotted below are two distributions where the red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class (patients with no disease). Now, this is an ideal situation because the two curves do not overlap at all. This means that the model has an ideal measure of separability, so it is perfectly able to distinguish between positive class and negative class. (See Figure 2)

However, when two distributions overlap, we introduce type I and type II error. (See Figure 3). Depending upon the threshold, we can minimize or maximize them. The threshold (commonly donated at η is the solid vertical line that we can vary in the overlap)

Note. The 7 Step Process of Statistical Hypothesis Testing
Power of a Hypothesis Test
Multiple Hypothesis Testing and False Discovery Rate

□

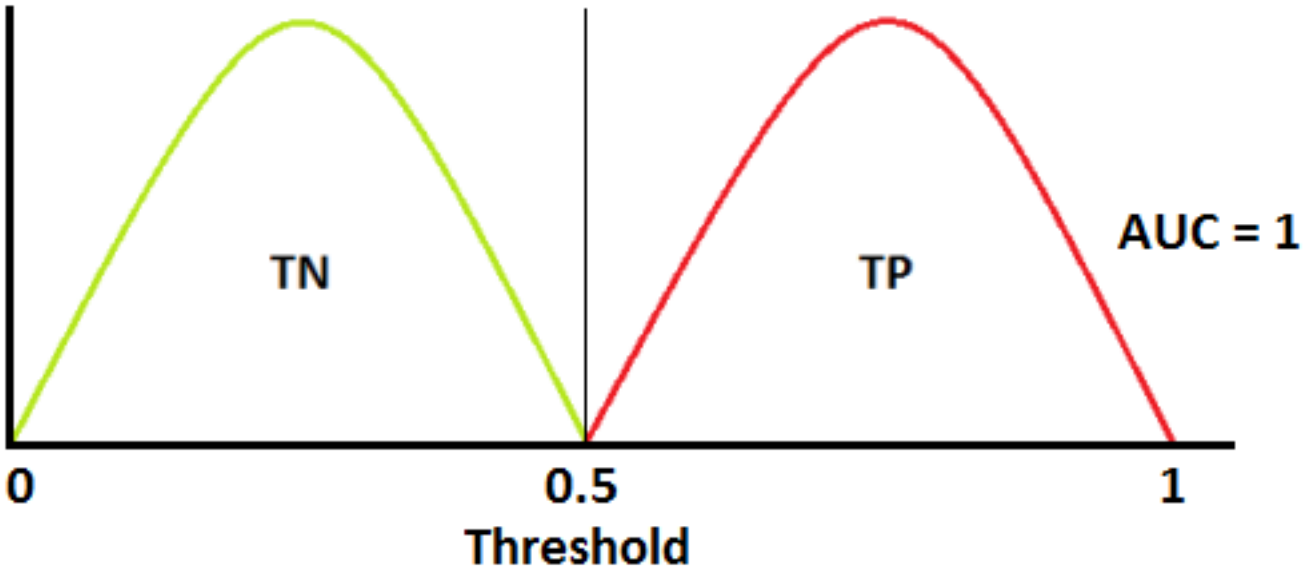

Figure 2: Two Distributions of Patients with disease and Patients with no disease

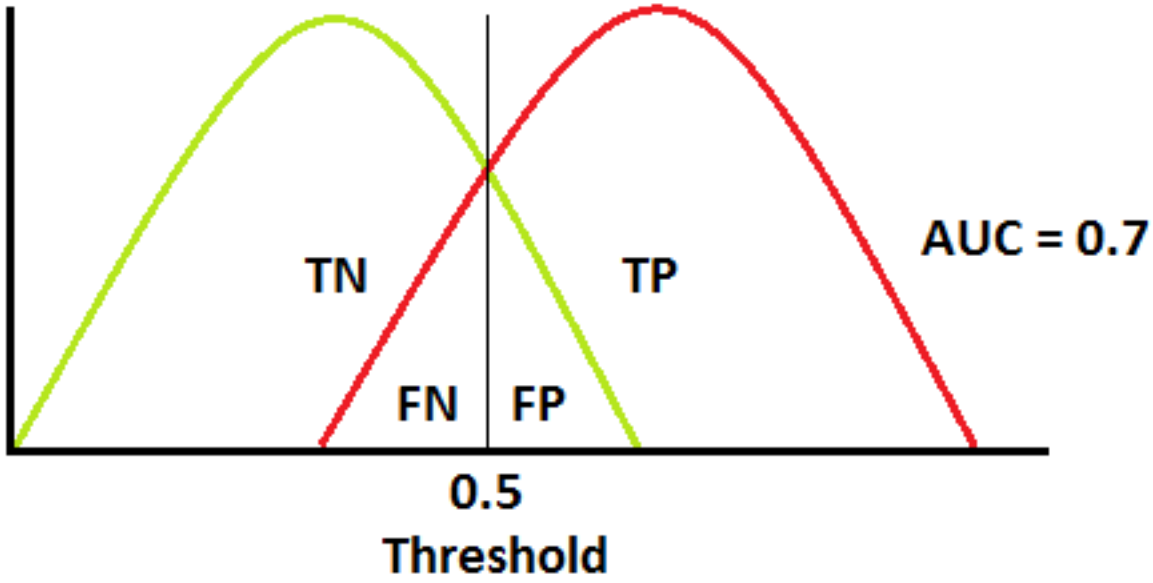

Figure 3: Type I and Type II error

Exercise 6. Describe t-test in the context of Machine Learning.

Proof. (Two sample) T-tests are used to assess the probability for two samples to originate from the same distribution. T-test (or any other type of noninferiority test), in the context of Machine Learning, attempts to provide some justification whether one model significantly outperforms another. That said, a t-test is ill suited for such comparisons (though not uncommon), as the normality assumption underlying the test will not hold.

Note. How to perform a T-test and follow this post on how to implement T-tests with Machine Learning algorithms. Another link on how to perform paired sample T-tests.

□

2 Extras

Exercise 7. When to use ROC vs. Precision-Recall Curves?

Proof. Generally, the use of ROC curves and precision-recall curves are as follows:

- ROC curves should be used when there are roughly equal numbers of observations for each class.
- Precision-Recall curves should be used when there is a moderate to large class imbalance.

The reason for this recommendation is that ROC curves present an optimistic picture of the model on datasets with a class imbalance. The main reason for this optimistic picture is because of the use of true negatives in the False Positive Rate in the ROC Curve and the careful avoidance of this rate in the Precision-Recall curve. Some go further and suggest that using a ROC curve with an imbalanced dataset might be deceptive and lead to incorrect interpretations of the model skill.

Note. More can be found here.

□

Exercise 8. How would you use ROC curve for multi-class model?

Proof. In a multi-class model, we can plot n number of AUC ROC Curves for n number classes using One vs ALL methodology. So, for example, of you have three classes named $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, you will have one ROC for $\boldsymbol{X}$ classified against $\boldsymbol{Y}$ and $\boldsymbol{Z}$, another ROC for $\boldsymbol{Y}$ classified against $\boldsymbol{X}$ and $\boldsymbol{Z}$, and a third one of $\boldsymbol{Z}$ classified against $\boldsymbol{Y}$ and $\boldsymbol{X}$.

Thanks Sarang Narkhede □

Exercise 9. Is it better to have too many false positives or too many false negatives? Explain.

Proof. It depends on the question as well as on the domain for which we are trying to solve the problem. If you are using Machine Learning in the domain of medical testing, then a false negative is very risky, since the report will not show any health problem when a person is actually unwell. Similarly, if Machine Learning is used in spam detection, then a false positive is very risky because the algorithm may classify an important email as spam. □

Exercise 10. Which is more important to you—model accuracy or model performance?

Proof. Well, model accuracy is only a subset of model performance. The accuracy of the model and performance of the model are directly proportional and hence better the performance of the model, more accurate are the predictions. □