Evaluating Performance of Neural Networks

Mean-squared Error

When a neural network is trained, the measure of performance that is optimized is usually the mean square error of the outputs.

The best measure for a particular network depends on the job duties of that network.

It should go without saying that the performance of a network must be evaluated by testing it on a different data set than the one on which it was trained.

MSE: It is easily computed by summing the squared differences between what a predicted variable should be versus what it actually is, then dividing by the number of components that went into that sum.

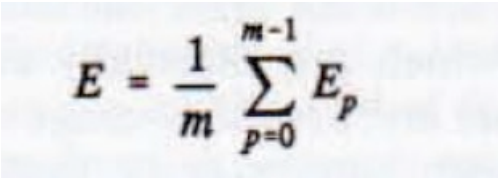
If there are n output neurons, the error for that single

presentation is →

$$E_p = \frac{1}{n} \sum_{j=0}^{n-1} (t_{pj} - o_{pj})^2$$

Problems with Mean-squared Error

If there are m presentations in the epoch, the error for that epoch is →



If the network is attempting to determine the presence of a particular signal pattern in a time series, the mean square error says nothing about the likelihood missing the pattern if it is present, or falsely detectings it when it is not present.

If the task is to classify a pattern into one of several categories, the mean square error tells us nothing about expected frequency of misclassification, let alone the nature of potential misclassifications.

Mean Absolute Error

The mean absolute error is the average magnitude of the error. It is identical to the mean square error except that the individual errors are not squared. Only their absolute value is taken. This eliminates the emphasis given to large errors.

Maximum Absolute Error

Maximum absolute error is the most intuitively informative error measure in this family. It can be useful to have an upper bound on the error.

Median Error

The median error, whether it be absolute or squared, is more robust than measures based on the mean. A few exceptionally large individual errors will not influence it like they would influence the mean.

Confusion Matrix

If the task of a neural network is to classify cases into one of several categories, the traditional technique of examining a confusion matrix can be highly informative.

“true positive” for correctly predicted event values.

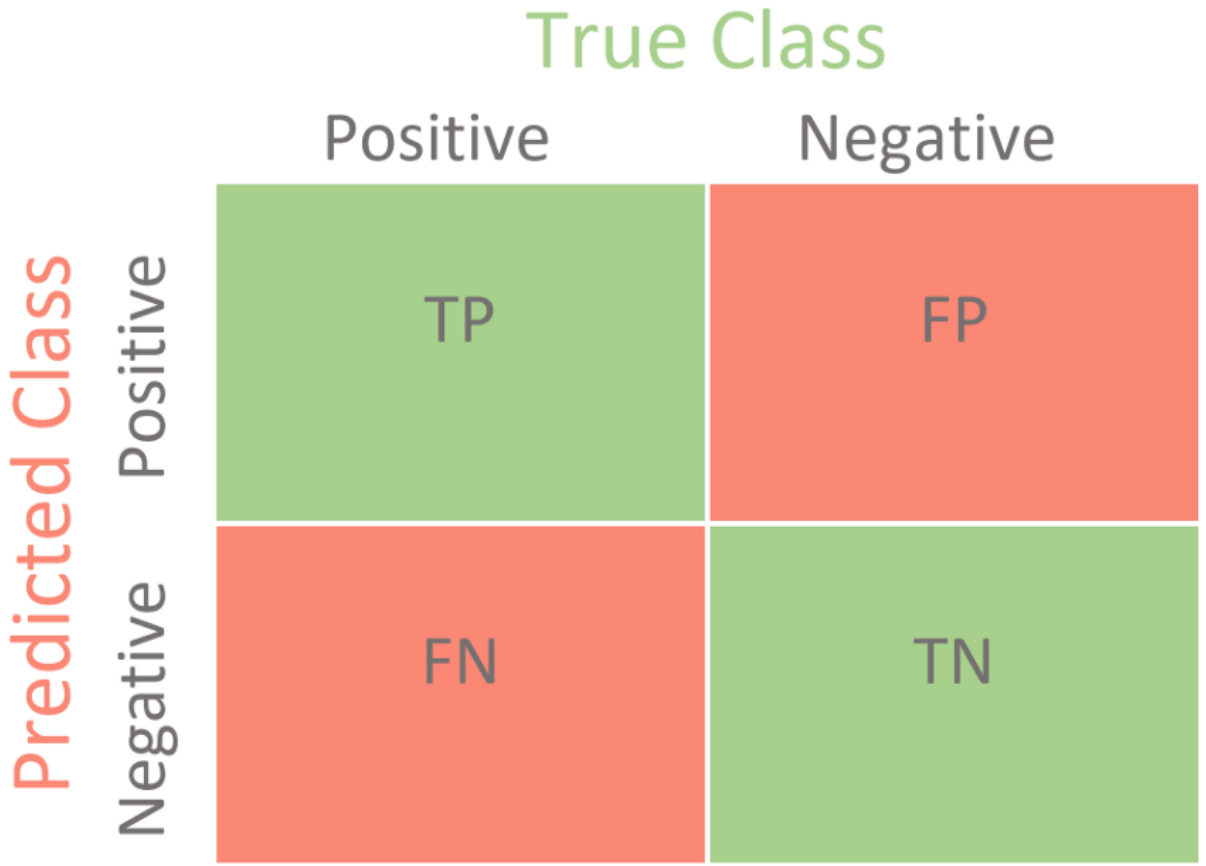
“false positive” for incorrectly predicted event values.

“true negative” for correctly predicted no-event values.

“false negative” for incorrectly predicted no-event values.

Accuracy: Overall, how often is the classifier correct?

(TP+TN)/total

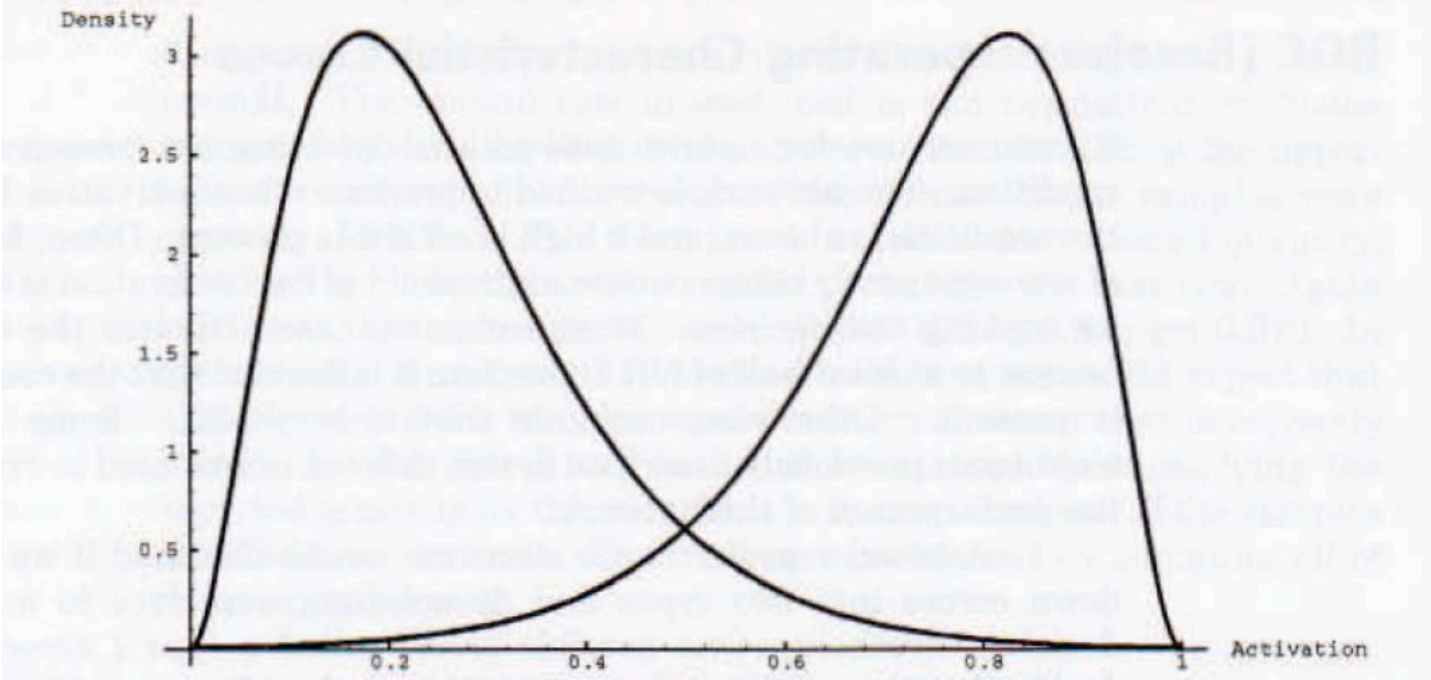


Distributions under null and alternative hypotheses

A type I error is committed if we conclude that the condition is present when it is not. This false positive probability is also frequently called the false alarm rate.

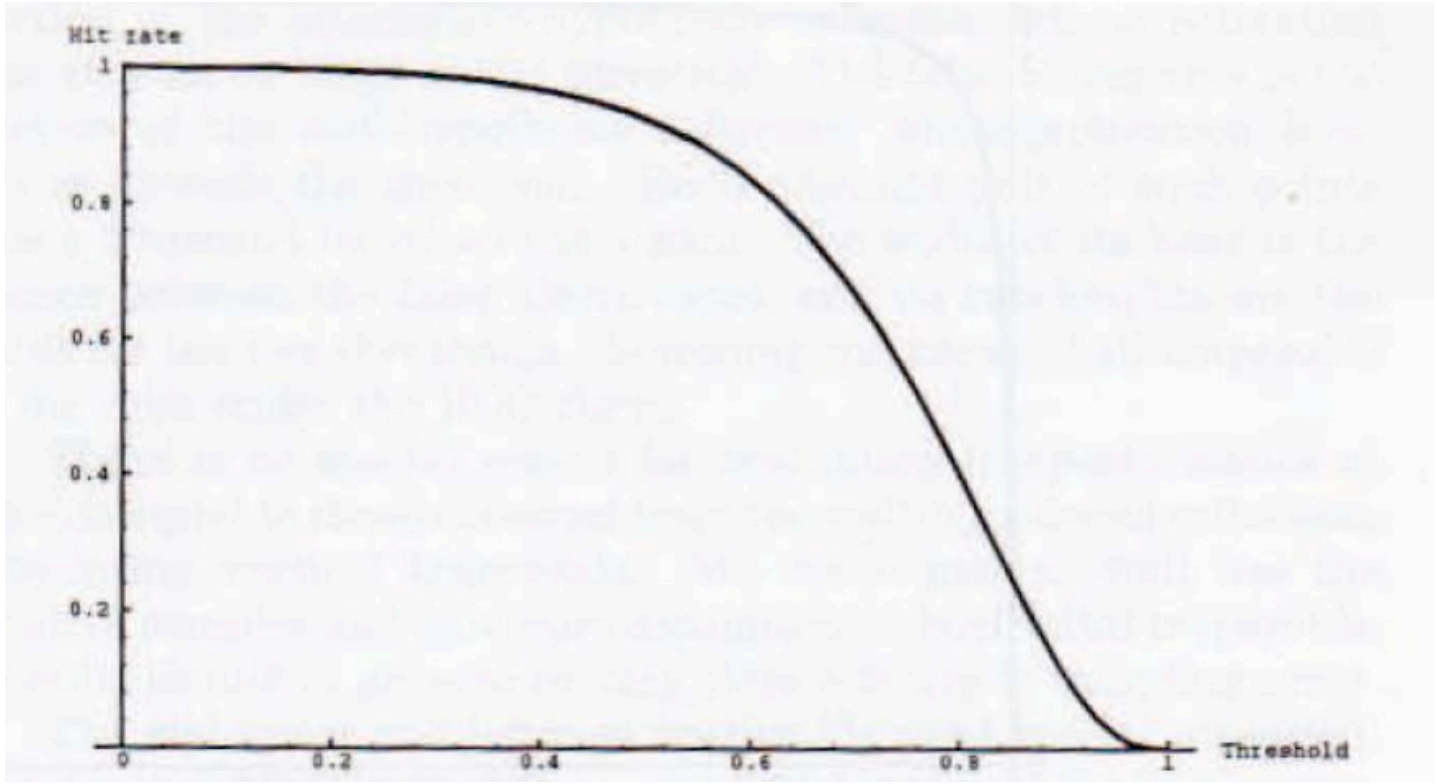
A type II error is committed if we conclude that the condition is absent (null hypothesis is true) when it is really present. The probability of this error is the area under the right curve to the left of the threshold.

One minus this error is the true positive rate. In military applications it is frequently called the hit rate, since it is the probability that we will detect the condition when it is present.



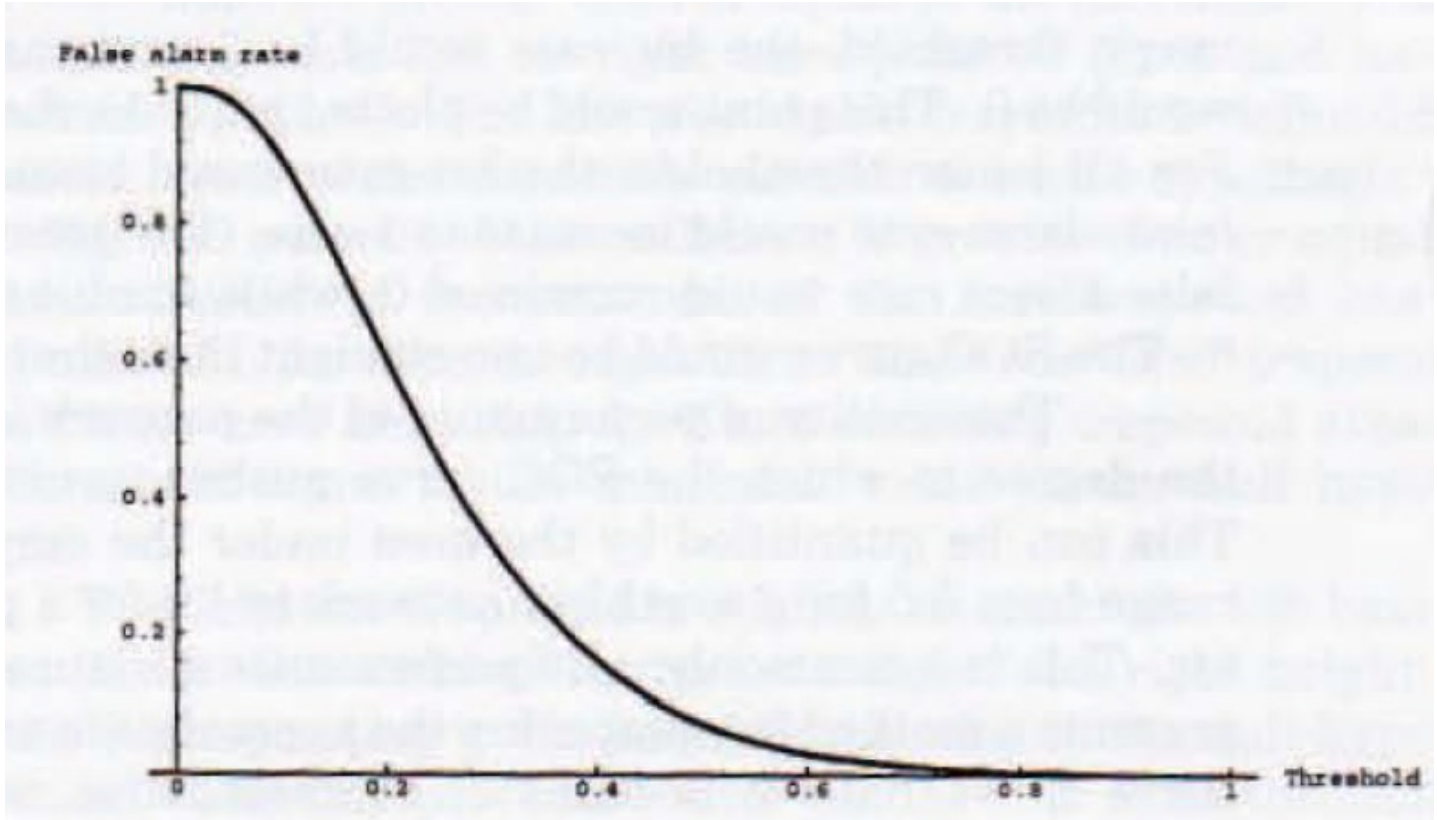
Hit rate versus activation threshold

If hit rate were the only consideration, we would obviously want to set the threshold as low as possible in order to have high probability of detecting the condition when it is present.



False alarm rate versus activation threshold

Not surprisingly. better performance in terms of false alarms is achieved by setting the threshold higher.

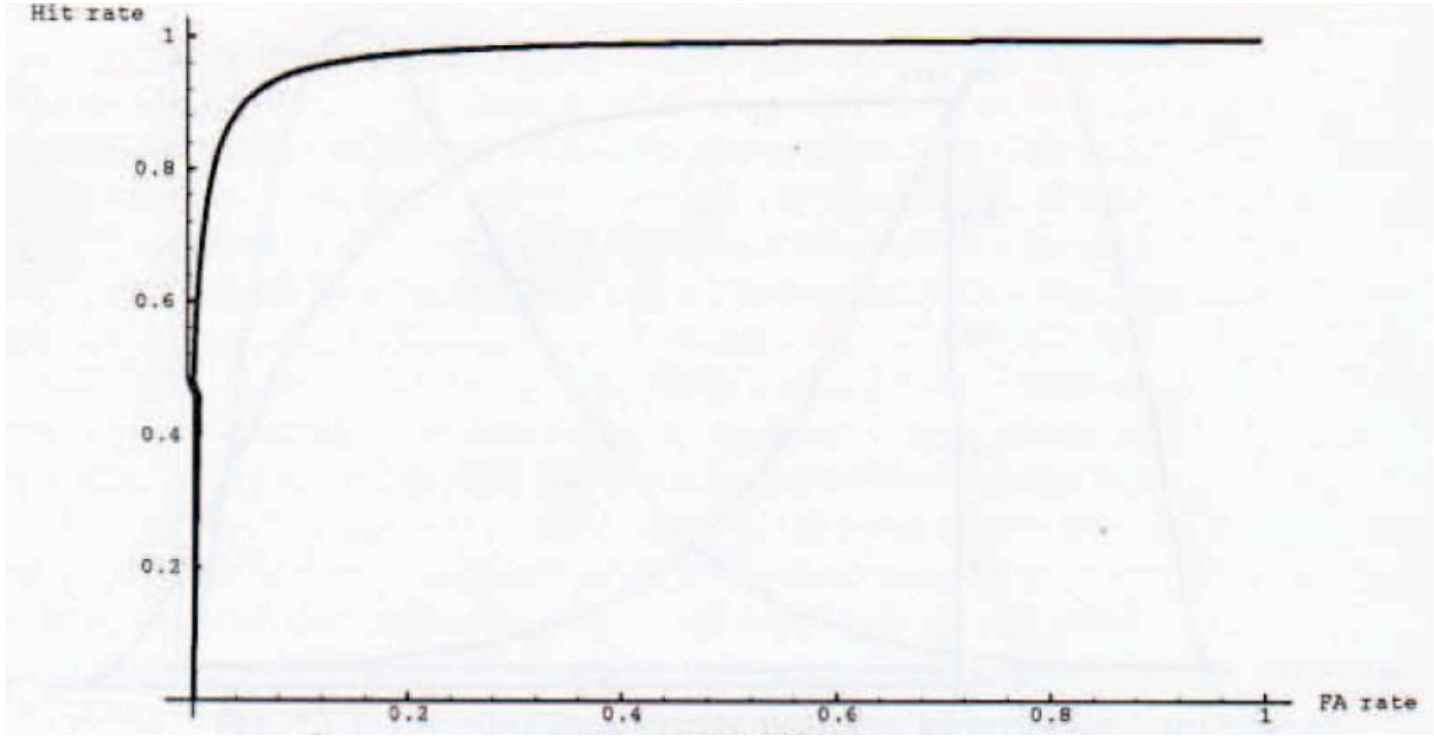


ROC (Receiver Operating Characteristics) Curve

The lower left corner (0, 0) of the ROC curve will always be one endpoint of this curve. At this threshold. a decision in favor of the null hypothesis will always be made. So, both the hit rate and the false alarm rate will be zero.

The upper-right corner (1, 1) will be the other endpoint of the ROC curve. This corresponds to a threshold of zero, which results in all decisions being in favor of the alternative hypothesis. The hit rate and the false alarm rate will both be 100 percent.

IT the network were worthless; the curve would be a straight line along the diagonal.



ROC Curve

If there were some threshold at which the network could perfectly discriminate, the ROC curve would be a right angle. The hit rate would be 1 and the false alarm rate would be 0. This point would be plotted at (0, 1), the upper left corner.

The quality of performance of the network is demonstrated by the degree to which the ROC curve pushes upward and to the left. This can be quantified by the area under the curve. The area will range from 0.5 for a worthless network to 1.0 for a perfect discriminator.

This is a commonly used performance measure.