Machine Learning — An Error by Any Other Name...

Kendall Fortney Follow
Jan 9, 2018 · 9 min read

It is easy to forget in a time of huge advances in Computer Science and Artificial Intelligence that by it's very nature models are not perfect. One of the greatest headaches is how to tackle measuring how accurate a model is in contrast to the know truth.

The first step is always understanding the cost of errors. Would you rather err on guessing something is true when it is not, or not guessing it at all. Sometimes it may be cheaper to lose a customer then spend the hours retaining them, or conversely the cost of vaccinations for those that may not get the disease versus the potential spreading of a sickness. This kind of cost/benefit analysis will inform the methodology used to identify the probability of being wrong.

The Importance of the Hypothesis

The most import part of modelling is starting with a meaningful question. Instinctively we form a question as a declaring something true, like "the space between airline seats is shrinking." However, the best statistical practice is to create a null hypothesis(also often referred to as H0) like "the distance between seats has remained the same" and then seek to prove it wrong.

This provides a statistical version of innocent until proven guilty and can help to eliminate the affects of randomness. If the null hypothesis is disproved to a statistically significant degree, then the alternative hypothesis (referred to as H1) is reviewed and the airline's shrinking tendencies is now considered as a possible but not definite explanation.

It is important to note that the null hypothesis is NOT the exact opposite of the alternative hypothesis but rather validation that the cause of the observations is not just random chance.

Probability and P Values

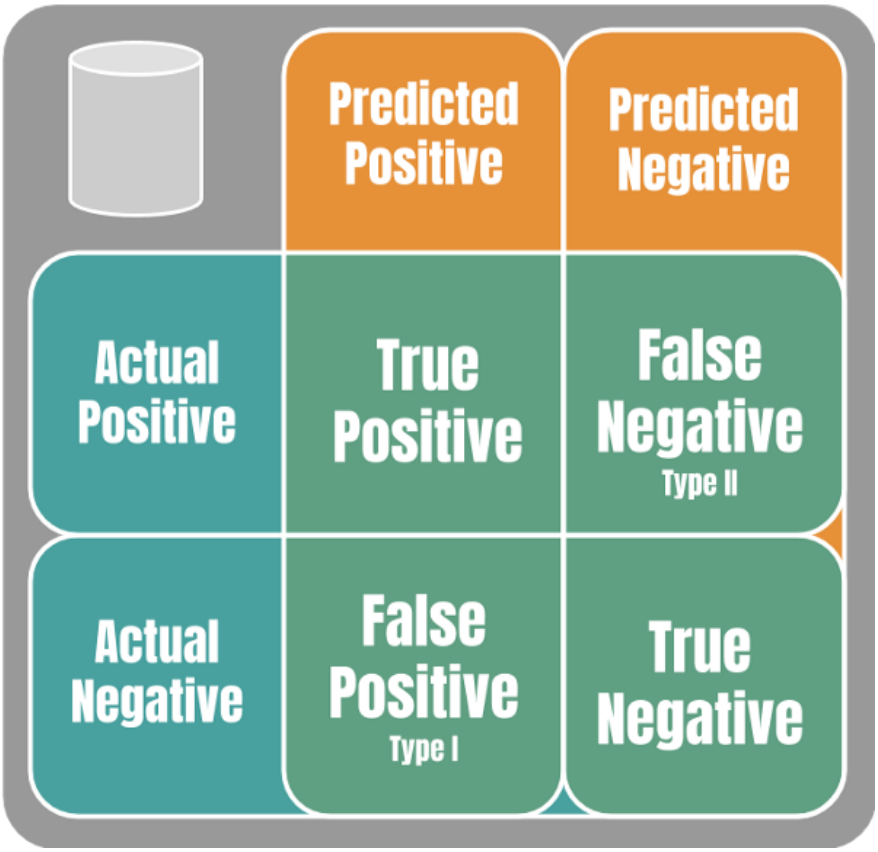
Probability provides a common way to interpret the statistical strength of a model. Called the p value, it can range from 0 to 1 and represents how likely it is to get a result if the null hypothesis (H1) is true. This means the lower the value the better indication that the alternative hypothesis (H1) is actually true.

The threshold for the p value is called the Level of Significance. If the probability is equal to or less than .05 (although depending on the use case that may change) then the result is often said to be significant. In simpler terms, we would likely get confirmation of the null hypothesis 5 times out of 100 (or conversely confirm the alternative hypothesis 95 of a 100 times). The higher the p value the closer to random chance and the more the null hypothesis is likely.

. . .

Error Measurement in Classification Problems

Classification problems are usually binary identification determining if an observation is, or is not, a certain condition. Out of any classification model there are four types of results.



True Positive(TP) -A true positive test result is one that detects the condition when the condition is present.

False Positive (FP)-Also know as a **Type I error**, a false positive test result is one that detects the condition when the condition is absent.

False Negative (FN)-Also know as a **Type II error**, a false negative test result is one that does not detect the condition when the condition is present.

True Negative (TN)- A true negative test result is one that does not detect the condition when the condition is absent.

Error is calculated of different ratios and formulas based on these four states. It is easy to see that depending on the cost of a Type I or a Type II the way the error is measured might be adjusted.

Measuring Error

To understand how a model is performing, there are a variety of ways to measure the interplay of the types of conditions. A **Confusion Matrix** (yes, that is really what it is called) is used to present multiple types of error measurements so a data scientist can determine if the

model is performing well or not. Below we will cover the following types of error measurements:

- Specificity or True Negative Rate (TNR)
- Precision, Positive Predictive Value (PPV)
- Recall, Sensitivity, Hit Rate or True Positive Rate (TPR)
- F Measure (F1,F0.5,F2)
- Matthew's Correlation Coefficient (MCC)
- ROC Area (ROC AUC)
- Fallout,False Positive Rate (FPR)
- R^2, Coefficient of Determination (r^2)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

. . .

Specificity or True Negative Rate (TNR)

TNR (ranges from 0 to 1, higher is better) measures the proportion of negatives that are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition).

$$TNR = TN/(TN+FP)$$

A good measurement if the costs of missing a negative value is high.

Precision, Positive Predictive Value (PPV)

PPV (ranges from 0 to 1, higher is better) is the ratio of true positives over all true and false positives:

$$PPV = TP/(TP+FP)$$

High precision means that an algorithm returned substantially more relevant results than irrelevant ones, or in other word the more likely

everything it returns is right, but it does not mean it may get all the right results that are out there.

Likewise this can be done with Negative Predictive Value (NPV) with positive flipped to negative and calculated to determine precision in the negative predictions. The complement of the NPV is the false omission rate (FOR).

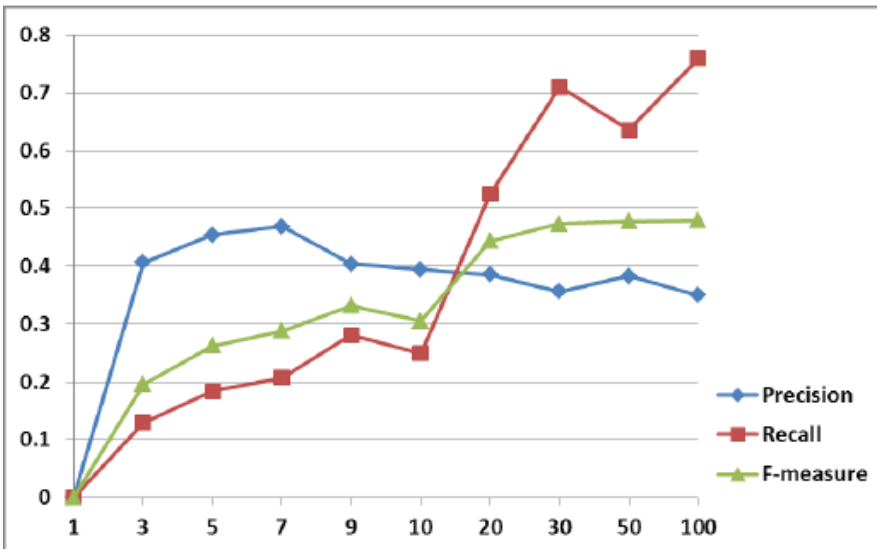
Recall, Sensitivity, Hit Rate or True Positive Rate (TPR)

TPR (ranges from 0 to 1, higher is better) is the ratio of true positives over the sum of true positives and false negatives:

$$TPR = TP / (TP+FN)$$

High recall means that an algorithm returned most of the relevant results, but it may have a bunch of false returns as well like a drag net that will certainly grab the fish you want but also catch a bunch you don't want.

F Measure



F Measure (ranges from 0–1) is a ratio that describes the balance between Precision (PPV) and Recall (TPR). Using the harmonic mean it can describe how heavily a model may be leaning one way or another.

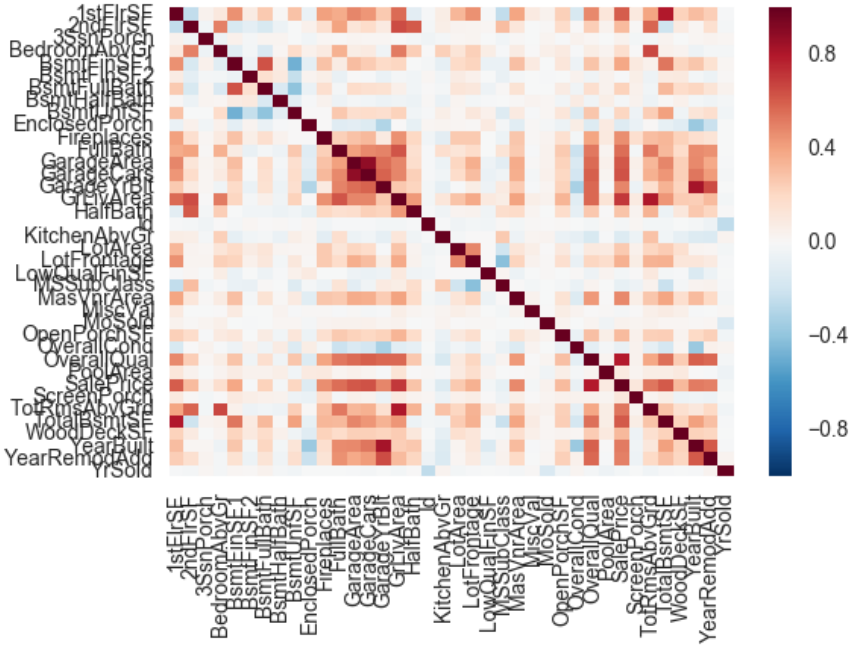
$$F = (PPV*TPR)/(PPV+TPR)$$

or

$$F = 2TP/(2TP+FP+FN)$$

The most common is called F1, while two other commonly used F measures are the F2 measure, which weights recall higher than precision, and the F0.5 measure, which puts more emphasis on precision than recall.

Matthew's Correlation Coefficient (MCC)



The MCC (ranges from -1 to 1) was introduced in 1975 by biochemist Brian W. Matthews. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications. A coefficient of +1 represents a perfect prediction, 0 is equal to no better than random prediction and −1 indicates total disagreement between prediction and observation. This is often represented as a correlation heatmap like it is here, and allows for quick observations about which features are useful (further in depth reading here about all the types of coorelation).

$$MCC = (TP*TN — FP*FN)/Sqr(FT+FP)(TN+FP)(TN+FN)$$

Values of 0.05+ can be useful as features for a model and negative correlation can, in certain situations, be helpful too.

ROC Area (ROC AUC)

The ROC, or "Receiver Operating Characteristic" (ranges from 0 to 1, higher is better) was originally used for radar object detection in World War II. The ROC area is a measure of the area under the curve produced by graphing the ratio between TPR and FPR with values ranging from 1-.8 is great to good, .8-.6 is fair to poor and below that is not better than random chance.

Recently the validity of using ROC has been called into question due to the noisiness and inconsistency of results (one example of several papers is here).

Fallout, False Positive Rate (FPR)

FPR (ranges from 0 to 1, lower is better) is the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events.

$$FDR = FP/(FP+TN)$$

Unlike many of the error rates above, the higher the value the worse as that means there are more proportionally more false negatives identified.

Accuracy (ACC)

Accuracy (ranges from 0 to 1, higher is better) is simply a ratio of correctly predicted observation to the total observations.

$$ACC = (TP+TN)/(TP+FP+FN+TN)$$

Instinctively one would assume accuracy is a great measure but it actually tells you very little about false positives and negatives.

. . .

Regressions and Error Methods

Unlike the classification problems above, regressions don't produce binary absolute values but rather a numeric range. Ideally algorithms should be stable, although what that means is largely dependent on the situation.

R^2, Coefficient of Determination

The Coefficent of Determination (ranges from 0 to 1, higher is better), also more often called R^2 or r^2, is the proportion of how well data fits the regression by using the ordinary least-squares regression.

There are multiple optimized versions of R^2 available for use depending on the case. The nature of R^2 means that the addition of variable, useful or not, will always increase its value. In those cases an Adjusted R^2 can be used.

Root Mean Squared Error (RMSE)

RMSE (ranges from 0 to infinity, lower is better), also called Root Mean Square Deviation (RMSD), is a quadratic-based rule to measure the absolute average magnitude of the error. Technically it is produce by taking residuals (the difference between the regression model and the actual data), squaring it, averaging all the results and then taking the square root of the average. Because of this the product will always be a positive number.

Because values are squared before averages, the affect of larger errors (think of the result of 3^2 compared to 8^2) is greatly amplified and should be used of those kinds of errors are important to identify (This is a great article that explains in a lot more detail).

RMSD can be normalized by mean or range in order to compared between models of different scale. It will usually be expressed as a percentage and notated as NRMSD or NRMSE.

Mean Absolute Error (MAE)

MAE (ranges from 0 to infinity, lower is better) is much like RMSE, but instead of squaring the difference of the residuals and taking the square root of the result, it just averages the absolute difference of the residuals. This produces a positive only numbers, and is less reactive to large errors but can show nuance a bit better. It has also fallen out of favor over time.

Summary

Understanding how to measure errors can help you tell when something is too good to be true or if it actually works. In putting this together I personally realized that the next thing I need to explore is normalization and how it can impact error recognition. Remember to

consider the cost of errors before creating a method to identify them otherwise it can have you chasing butterflies rather than identifying the right optimizations.

Consider that any model is highly optimized for the data it was trained on. It is expected that the error rate on new data will *always* be higher than it was for the training set.

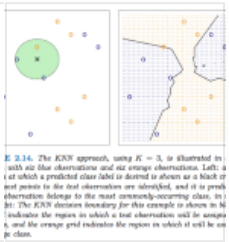
This is far from comprehensive list of error measurement, go and explore!

Additional Links:

Model Accuracy

Model accuracy is typically assessed by measuring the quality of fit (mean squared error, typically).

medium.com



Simple guide to confusion matrix terminology

A confusion matrix is a table that is often used to describe the performance of a classification mod...

www.dataschool.io

Accurately Measuring Model Prediction Error

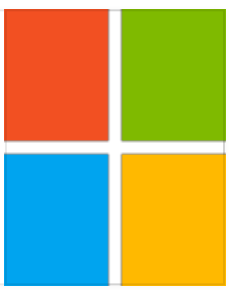
When assessing the quality of a model, being able to accurately measure its prediction error is of k...

scott.fortmann-roe.com

Evaluate model performance in Machine Learning

This article demonstrates how to evaluate the performance of a model in Azure Machine...

docs.microsoft.com



What is a hypothesis test? - Minitab

A hypothesis test examines two opposing hypotheses about a population: the null...

support.minitab.com