Accuracy measures for machine learning

Introduction

This article is a brief introduction to some of the ideas behind measuring the accuracy of supervised machine learning tools. The first half deals with classification algorithms; those that decide which of several classes a sample belongs to. The chief measure of accuracy there is the confusion matrix, from which a whole host of other statistics may be extracted. The second half of this article deals with regression algorithms, which give a predicted numerical value for some outputs from known inputs.

Confusion matrix

A confusion matrix is a method for measuring the accuracy of a classification algorithm. It is most easily understood through an example: imagine that a classification algorithm has been trained to distinguish between carrots, bananas, and apples. We can then draw a table of the results it obtained in testing:

		Actual class		
		Carrot	Banana	Apple
Predicted class	Carrot	5	2	0
	Banana	3	3	2
	Apple	0	1	11

The numbers in the matrix represent the number of tests that returned a particular result: for example, for the 8 test cases that were actually carrots, the classifier correctly thought that 5 of them were carrots, and incorrectly thought 3 of them were bananas.

The entries on the diagonal of the matrix are correctly classified, and the entries off the diagonal are incorrectly classified. The matrix provides more information than simply the proportion of results that are correctly classified: in the above example, it can be inferred that the classifier is good at identifying apples (only one false positive apple, and two false negative apples, against 11 correctly identified apples), whilst the

classification for bananas is much less accurate. The classifier especially had trouble telling apart carrots and bananas: the corresponding sub-matrix does not appear especially strongly peaked around the diagonal.

A whole range of statistics may be extracted directly from the confusion matrix. These statistics generally refer to the accuracy of classification of one class (binary classification): in this example, bananas.

The condition positive (P) is the number of actual bananas: here, 6. The condition negative (N) is the number of actual not-bananas: here, 21. True positives (TP, 3 here), true negatives (TN, 16 here), false positives (FP, 5 here) and false negatives (FN, 3 here) may be combined to yield information about the accuracy of the classifier. Some particular examples of important statistics are:

- Sensitivity or true positive rate $\frac{TP}{P}$ (0.5)
- Specificity or true negative rate $\frac{TN}{N}$ (0.762)
- Precision $\frac{TP}{TP+FP}$ (0.375)
- False discovery rate $\frac{FP}{FP+TP}$ (0.625)
- False omission rate $\frac{FN}{TN+FN}$ (0.158)
- Accuracy $\frac{TP+TN}{P+N}$ (0.704)
- F1 score, the harmonic mean of precision and sensitivity $\frac{2TP}{2TP+FP+FN}$ (0.429)

All these statistics may be compared between classifiers. They may also be judged on their own merits: all have a possible range of [0,1], and a required standard may be specified ahead of testing.

In problems where there are very imbalanced numbers of data points in the different classes (here, we have many more not-bananas than we do bananas), none of the statistics above are entirely reliable. A modified form of the F1 score, known as the F$_\beta$ score, can be used to weight precision and sensitivity differently, prioritising either the classifier identifying every banana (high sensitivity, $\beta > 1$), or only identifying things it is certain are bananas (high precision, $\beta < 1$). The F$_\beta$ score can be expressed as

$$\frac{(1+\beta^2)TP}{(1+\beta^2)TP+FP+\beta^2 FN},$$

but the value of β (and hence relative importance of sensitivity and precision) need to be set ahead of time, and hence knowledge about the relative sizes of the classes is required.

Matthews correlation coefficient

One further statistic that may be extracted from confusion matrices is the Matthews correlation coefficient. Again, this measure is primarily designed for examining binary classifications, although it has been extended to cover the multiclass case. For binary classification, the Matthews correlation coefficient is given by

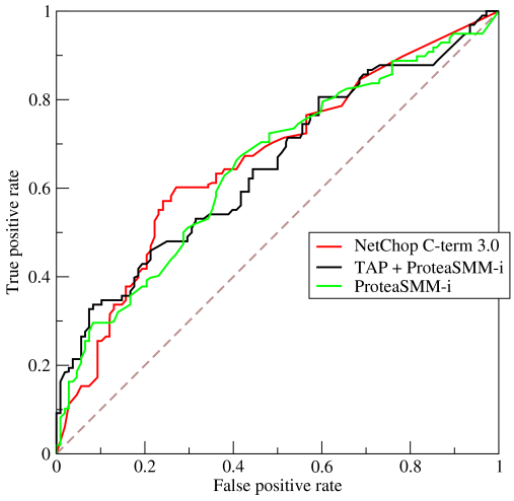
$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

and determines a correlation coefficient that is the geometric mean of the regression coefficients of the problem and its dual. It has been described as one of the best ways of encapsulating the full confusion matrix in a single number, and does not have any problem with systems with imbalanced numbers of data points.

The Matthews correlation coefficient takes values in the range [-1,1], with 1 being perfect prediction and -1 completely incorrect prediction: 0 indicates the prediction is no better than random. For the example above, the Matthews correlation coefficient gives 0.932, indicating that despite the fairly low precision this classifier has done a reasonably good job of recognising bananas. Again the primary use case for the Matthews correlation coefficient is in comparisons between classifiers.

ROC curve

Another method of analysing the accuracy of a binary classification algorithm is through a Receiver Operating Characteristics (ROC) curve. This curve plots the true positive rate $\frac{TP}{P}$ against the false positive rate $\frac{FP}{N}$, with the curve being plotted parametrically as a function of a sensitivity threshold. Example ROC curves from a biological context are shown in the figure: at



any point on one of the curves, the sensitivity and (1-specificity) can be read off. A perfect classifier's ROC curve would pass through the top left corner of the plot, and a classifier that's no better than random would have a curve along the diagonal; most classifiers, naturally, are somewhere in between. The Area Under a ROC Curve (AUC) is another measure of the accuracy of a classifier: a value of 1 indicates that the classifier correctly identifies every sample, a value of 0.5 indicates it cannot distinguish between the two classes, and a value of 0 indicates it gets every sample incorrect (and one's classifier should be inverted).

Pearson correlation coefficient

The Pearson correlation coefficient is a metric designed for use comparing two variables. In a machine-learning context, these can be the target values for a supervised regression algorithm and the predicted values. Plotting one against the other in a scatter graph, a perfect machine-learning algorithm would give a straight line through the origin ($y = x$), with inaccuracies in the result giving scatter around this line. The Pearson correlation coefficient gives a measure of this scatter. The Pearson correlation coefficient for the relationship between X and Ycan be expressed as

$$\frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)},$$

where $\sigma(X)$ is the standard deviation of the variable X and $\sigma(X,Y)$ is the covariance of the variables X and Y, defined as $\sigma(X)^2 = E[X^2] - (E[X])^2$ with $E[X]$ being the expectation value (mean) of X, and $\sigma(X,Y) = E[XY] - E[X]E[Y]$. The Pearson correlation coefficient takes values in the range [-1,1], with 1 and -1 being perfect correlation and anti-correlation, and 0 indicating the variables are uncorrelated.

The Pearson correlation coefficient is only designed for comparisons where the relationship between variables is expected to be linear. This is perfectly adequate for examining the accuracy of a machine learning regression algorithm, but the concept behind the Pearson correlation coefficient may be easily extended to include non-linear relationships.

r^2 coefficient

In the case of linear regression, as found for the accuracy of a regression algorithm, the r^2 coefficient is just the square of the Pearson correlation coefficient. More generally, we can define

$$r^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - E[Y])^2},$$

where y_i are the target values, and f_i are the predicted values. This measure of least-squares correlations takes values between 0 and 1, as before with 0 indicating no correlation and 1 perfect correlation. This correlation coefficient is also referred to as the coefficient of determination.

There are multiple variations on the r^2 coefficient, including modifications to remove the unwelcome property of the original coefficient spuriously increasing when extra data points are introduced. However the base r^2 coefficient remains a popular choice for analysing the accuracy of models of data.

The choice between the Pearson correlation coefficient and the r^2 coefficient when analysing the accuracy of a machine learning regression algorithm is to some extent a matter of personal preference. For linear regression all the information in the r^2 coefficient is contained in the Pearson correlation coefficient, although the inverse is not true; but for analysing the accuracy of a machine learning algorithm, where a linear relationship is expected and large changes in the evaluated gradient are unlikely, both measures give equivalent information.

Mean squared error

The Mean Squared Error (MSE) is similar to the r^2 coefficient, and used for analysing a *supervised regression* algorithm's accuracy. Using the same notation as above, MSE is expressed as

$$\frac{\sum_i (y_i - f_i)^2}{\sum_i 1},$$

i.e. the sum of the squared errors, divided by the number of samples to obtain the mean. For a set of target values with unit variance, the MSE tends to $1 - r^2$; but for general (dimensionful) data, the MSE is dimensionful (whilst r^2 is dimensionless), and the magnitude of the MSE depends on the magnitude of the data. This makes the MSE less transferable, and more difficult to interpret, than the r^2 coefficient, without prior knowledge of the data.

Relative error

The relative error is another measure of accuracy for a regression algorithm, which combines features of both the r^2 coefficient and the MSE. It is expressed as

$$\frac{1}{\sum_i 1}\sum_i \frac{|y_i - f_i|}{y_i},$$

where $|x|$ is the absolute value of x. This measure is dimensionless, like the r^2 coefficient, but suffers from problems when the expected outcome is 0 (as this appears in the denominator), and only makes sense for measurements in units of a ratio scale (one where zero is a definite lower bound on the possible values), as otherwise shifting every output value will change the measured relative error. This makes the relative error a much worse measure of accuracy than r^2 in most cases.

Multi-dimensional regression

The basic expressions for the Pearson correlation coefficient, r^2 coefficient, and MSE above assume that there is only one target variable being optimised in the regression. However, some more advanced machine learning algorithms are capable of mapping inputs to multiple outputs simultaneously. The most obvious way to measure the accuracy of multiple outputs together is by simply summing over the MSE of each individually; however, this only makes sense for outputs with the same dimensionality. Similarly, summing over the relative error is only suitable when all variables are measured on ratio scales.

However, there are alternative measures more suited to multi-dimensional regression. For D dimensions of data, the average relative root mean square error (aRRMSE) takes the form

$$\frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{\sum_i (y_i^{(d)} - f_i^{(d)})^2}{\sum_i (y_i^{(d)} - E[Y^{(d)}])^2}}.$$

Similarly, a multi-dimensional version of the Pearson correlation coefficient can be written

$$\frac{1}{D}\sum_{d=1}^{D}\frac{\sum_i (y_i^{(d)} - E[Y^{(d)}])(f_i^{(d)} - E[F^{(d)}])}{\sqrt{\sum_i (y_i^{(d)} - E[Y^{(d)}])^2}\sqrt{\sum_i (f_i^{(d)} - E[F^{(d)}])^2}},$$

which contains information about the correlation vs anticorrelation of the result as well. Either of these expressions could be squared, to obtain a measure similar to a multi-dimensional r^2, which could more directly be written as

$$1 - \frac{1}{D}\sum_{d=1}^{D} \frac{\sum_i (y_i^{(d)} - f_i^{(d)})^2}{\sum_i (y_i^{(d)} - E[Y^{(d)}])^2}.$$

Note that squaring the previous measures does not give the same result as this version of multi-dimensional r^2, and they cannot be compared, although either separately would serve as a good measure of accuracy. The multi-dimensional r^2, Pearson correlation coefficient, and aRRMSE, weight each dimension of the output identically; the sum over them could be weighted, but this would need to be justified before the analysis was begun.

Conclusions

The measures used for the accuracy of machine learning algorithms can be split into two classes; those for classification and those for regression, just as with the algorithms themselves. The confusion matrix is the fundamental object for analysing the accuracy of classification algorithms; even a simple binary classification can be analysed in depth using a confusion matrix. Extracting statistics from the confusion matrix is not difficult, although the choice of statistic to use is not straightforward: simple measures like the sensitivity and precision lose a lot of information from the confusion matrix, whilst the F1 score is biased in cases with different class sizes. The Matthews correlation coefficient is a balanced measure that gives a good idea of the accuracy of a classification algorithm, although a single value can never capture the full detail of a matrix of accuracies in the confusion matrix.

For regression algorithms, the choice of statistic to use is more straightforward; a scatter plot of predicted value vs actual value can be interpreted in terms of the equivalent Pearson or r^2 correlation coefficients, and these values have absolute scales that can be used a priori to set a required accuracy. The mean square error, whilst similar to the r^2 coefficient, suffers from a lack of transferability between problems, and relative error depends on the measurement units, meaning that a preference choice between the Pearson and r^2 coefficients is the main decision to make when it comes to choosing an accuracy measure for machine learning regression algorithms.

For multi-dimensional regression problems, the Pearson correlation coefficient may be extended to capture information about the accuracy of the regression in all output dimensions. This (or perhaps its square) is probably the most effective measure of multi-dimensional regression accuracy, although as in a single dimension the choice between Pearson and r^2 coefficients is mostly personal preference.