The Coefficient of Determination

- The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information.
- It is the proportion of variability in a data set that is accounted for by the statistical model.
- It provides a measure of how well future outcomes are likely to be predicted by the model.

R-Squared

- R^2 is a statistic that will give some information about the goodness of fit of a model.
- In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points.
- An R^2 of 1.0 indicates that the regression line perfectly fits the data.
- In the case of simple linear regression, the coefficient of determination is equivalent to the squared value of the Pearson correlation coefficient. (Consider this to be co-incidental, rather than a definition)

The Adjusted Coefficient of Determination

- Adjusted R^2 (often written as and pronounced "**R bar squared**") is a modification of R^2 that adjusts for the number of predictor terms in a model.
- Adjusted R^2 is used to compensate for the addition of variables to the model.
- As more independent variables are added to the regression model, unadjusted R^2 will generally increase but there will never be a decrease.
- This will occur even when the additional variables do little to help explain the dependent variable.

- To compensate for this, adjusted R^2 is corrected for the number of independent variables in the model, increases only if the new term improves the model more than would be expected by chance.
- If too many predictor variables are being used, this will be reflected in a reduced adjusted R^2.
- The adjusted R^2 can be negative, and will always be less than or equal to R^2.
- The result is an adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model.
- Adjusted R^2 will always be lower than

- Adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square.
- It has become standard practice to report the adjusted R^2, especially when there are multiple models presented with varying numbers of independent variables.

Confusion Matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Accuracy

Accuracy

Accuracy measures a fraction of the classifier's predictions that are correct.

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

Accuracy

	Predict P	Predict N
True P	100	70
True N	30	9800

$$\text{Accuracy} = \frac{9900}{10,000} = 0.99$$

Accuracy

- However, accuracy is not an informative metric if the proportions of the classes are skewed in the population. (**Class Imbalance**)
- For example, a classifier that predicts whether or not credit card transactions are fraudulent may be more sensitive to false negatives than to false positives.
- To promote customer satisfaction, the credit card company may prefer to risk verifying legitimate transactions than risk ignoring a fraudulent transaction.

Accuracy

- Because most transactions are legitimate, accuracy is not an appropriate metric for this problem.
- A classifier that always predicts that transactions are legitimate could have a high accuracy score, but would not be useful.
- For these reasons, classifiers are often evaluated using two additional measures called **precision** and **recall**.

Precision and Recall

- **Precision** (or positive predictive value (PPV))

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

- **Recall** (or Sensitivity , true positive rate (TPR))

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

- **Specificity** (or true negative rate (TNR))

$$SPC = TN/N = TN/(TN + FP)$$

Some Related Metrics

- Negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

- False positive rate (FPR) also called "fallout"

$$FPR = FP/N = FP/(FP + TN)$$

- False negative rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

- False discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Precision and Recall

- Precision is the fraction of positive predictions that are correct.
- For instance, in our SMS spam classifier, precision is the fraction of messages classified as spam that are actually spam.
- Precision is given by the following ratio:

$$P = \frac{TP}{TP + FP}$$

The Confusion Matrix

- The confusion matrix indicates that there were four true negative predictions, three true positive predictions, two false negative predictions, and one false positive prediction.
- Confusion matrices become more useful in multi-class problems, in which it can be difficult to determine the most frequent types of errors.

Recall

- Recall is the fraction of the truly positive instances that the classifier recognizes. A recall score of one indicates that the classifier did not make any false negative predictions.
- For our SMS spam classifier, recall is the fraction of spam messages that were truly classified as spam.
- Recall is calculated with the following ratio:

$$\text{Recall} = \frac{tp}{tp + fn}$$

- Recall is often called sensitivity in medical contexts.

Precision and Recall

- Individually, precision and recall are seldom informative; they are both incomplete views of a classifier's performance.
- Both precision and recall can fail to distinguish classifiers that perform well from certain types of classifiers that perform poorly.
- A trivial classifier could easily achieve a perfect recall score by predicting positive for every instance.

Precision and Recall

- For example, assume that a test set contains ten positive examples and ten negative examples.
- A classifier that predicts positive for every example will achieve a recall of one, as follows:

10 1 10 0 R = = +

- A classifier that predicts negative for every example, or that makes only false positive and true negative predictions, will achieve a recall score of zero.
- Similarly, a classifier that predicts that only a single instance is positive and happens to be correct will achieve perfect precision.