Title: Phi coefficient

URL: https://en.wikipedia.org/wiki/Phi_coefficient

PageID: 23240139

Categories: Category:Bioinformatics, Category:Cheminformatics, Category:Computational chemistry, Category:Information retrieval evaluation, Category:Machine learning, Category:Statistical classification, Category:Statistical ratios, Category:Summary statistics for contingency tables

---

In statistics , the phi coefficient , or mean square contingency coefficient , denoted by ϕ or r_ϕ , is a measure of association for two binary variables .

In machine learning , it is known as the Matthews correlation coefficient (MCC) and used as a measure of the quality of binary (two-class) classifications , introduced by biochemist Brian W. Matthews in 1975. [1]

Introduced by Karl Pearson , [2] and also known as the Yule phi coefficient from its introduction by Udny Yule in 1912 [3] this measure is similar to the Pearson correlation coefficient in its interpretation.

In meteorology , the phi coefficient, [4] or its square (the latter aligning with M. H. Doolittle's original proposition from 1885 [5]), is referred to as the Doolittle Skill Score or the Doolittle Measure of Association.

Definition

A Pearson correlation coefficient estimated for two binary variables will return the phi coefficient. [6]

Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal.

If we have a 2×2 table for two random variables x and y

where n 11 , n 10 , n 01 , n 00 , are non-negative counts of numbers of observations that sum to n , the total number of observations. The phi coefficient that describes the association of x and y is

Phi is related to the point-biserial correlation coefficient and Cohen's d and estimates the extent of the relationship between two variables (2×2). [7]

The phi coefficient can also be expressed using only n {\displaystyle n} , n 11 {\displaystyle n_{11}} , n 1 ■ {\displaystyle n_{1\bullet }} , and n ■ 1 {\displaystyle n_{\bullet 1}} , as

Maximum values

Although computationally the Pearson correlation coefficient reduces to the phi coefficient in the 2×2 case, they are not in general the same. The Pearson correlation coefficient ranges from −1 to +1, where ±1 indicates perfect agreement or disagreement, and 0 indicates no relationship. The phi coefficient has a maximum value that is determined by the distribution of the two variables if one or both variables can take on more than two values. [further explanation needed] See Davenport and El-Sanhury (1991) [8] for a thorough discussion.

Machine learning

The MCC is defined identically to phi coefficient, introduced by Karl Pearson , [2] [9] also known as the Yule phi coefficient from its introduction by Udny Yule in 1912. [3] Despite these antecedents which predate Matthews's use by several decades, the term MCC is widely used in the field of bioinformatics and machine learning.

The coefficient accounts for true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. [10] The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. However, if MCC equals neither −1, 0, or +1, it is not a reliable indicator of how similar a predictor is to random guessing because MCC is dependent on the dataset. [11] MCC is closely related to the chi-square statistic for a 2×2 contingency table

where n is the total number of observations.

While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best such measures. [12] Other measures, such as the proportion of correct predictions (also termed accuracy), are not useful when the two classes are of very different sizes. For example, assigning every object to the larger set achieves a high proportion of correct predictions, but is not generally a useful classification.

The MCC can be calculated directly from the confusion matrix using the formula:

In this equation, TP is the number of true positives , TN the number of true negatives , FP the number of false positives and FN the number of false negatives . If exactly one of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews correlation coefficient of zero, which can be shown to be the correct limiting value. In case two or more sums are zero (e.g. both labels and model predictions are all positive or negative), the limit does not exist.

The MCC can be calculated with the formula:

using the positive predictive value, the true positive rate, the true negative rate, the negative predictive value, the false discovery rate, the false negative rate, the false positive rate, and the false omission rate.

The original formula as given by Matthews was: [1]

This is equal to the formula given above. As a correlation coefficient , the Matthews correlation coefficient is the geometric mean of the regression coefficients of the problem and its dual . The component regression coefficients of the Matthews correlation coefficient are markedness (Δ p) and Youden's J statistic (informedness or Δ p ′). [12] [13] Markedness and informedness correspond to different directions of information flow and generalize Youden's J statistic , the δ p {\displaystyle \delta p} statistics, while their geometric mean generalizes the Matthews correlation coefficient to more than two classes. [12]

Some scientists claim the Matthews correlation coefficient to be the most informative single score to establish the quality of a binary classifier prediction in a confusion matrix context. [14] [15]

Example

Given a sample of 12 pictures, 8 of cats and 4 of dogs, where cats belong to class 1 and dogs belong to class 0,

assume that a classifier that distinguishes between cats and dogs is trained, and we take the 12 pictures and run them through the classifier, and the classifier makes 9 accurate predictions and misses 3: 2 cats wrongly predicted as dogs (first 2 predictions) and 1 dog wrongly predicted as a cat (last prediction).

With these two labelled sets (actual and predictions) we can create a confusion matrix that will summarize the results of testing the classifier:

In this confusion matrix, of the 8 cat pictures, the system judged that 2 were dogs, and of the 4 dog pictures, it predicted that 1 was a cat. All correct predictions are located in the diagonal of the table (highlighted in bold), so it is easy to visually inspect the table for prediction errors, as they will be represented by values outside the diagonal.

In abstract terms, the confusion matrix is as follows:

where P = positive; N = negative; TP = truepositive; FP = false positive; TN = true negative; FN = false negative.

Plugging the numbers from the formula:

Confusion matrix

Let us define an experiment from P positive instances and N negative instances for some condition. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix , as follows:

view

talk

edit

Multiclass case

The Matthews correlation coefficient has been generalized to the multiclass case. The generalization called the R K {\displaystyle R_{K}} statistic (for K different classes) was defined in terms of a K × K {\displaystyle K\times K} confusion matrix C {\displaystyle C} [24] . [25]

When there are more than two labels the MCC will no longer range between −1 and +1. Instead the minimum value will be between −1 and 0 depending on the true distribution. The maximum value is always +1.

This formula can be more easily understood by defining intermediate variables: [26]

t k = ∑ i C i k {\displaystyle t_{k}=\sum _{i}C_{ik}} the number of times class k truly occurred,

p k = ∑ i C k i {\displaystyle p_{k}=\sum _{i}C_{ki}} the number of times class k was predicted,

c = ∑ k C k k {\displaystyle c=\sum _{k}C_{kk}} the total number of samples correctly predicted,

s = ∑ i ∑ j C i j {\displaystyle s=\sum _{i}\sum _{j}C_{ij}} the total number of samples. This allows the formula to be expressed as:

Using above formula to compute MCC measure for the dog and cat example discussed above, where the confusion matrix is treated as a 2 × Multiclass example:

An alternative generalization of the Matthews Correlation Coefficient to more than two classes was given by Powers [12] by the definition of Correlation as the geometric mean of Informedness and Markedness .

Several generalizations of the Matthews Correlation Coefficient to more than two classes along with new Multivariate Correlation Metrics for multinary classification have been presented by P Stoica and P Babu. [27]

Advantages over accuracy and F1 score

As explained by Davide Chicco in his paper "Ten quick tips for machine learning in computational biology " [14] (BioData Mining , 2017) and "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" [28] (BMC Genomics , 2020), the Matthews correlation coefficient is more informative than F1 score and accuracy in evaluating binary classification problems, because it takes into account the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives). [14] [28]

The former article explains, for Tip 8 : [excessive quote]

In order to have an overall understanding of your prediction, you decide to take advantage of common statistical scores, such as accuracy, and F1 score.

(Equation 1, accuracy: worst value = 0; best value = 1)

(Equation 2, F1 score: worst value = 0; best value = 1)

However, even if accuracy and F1 score are widely employed in statistics, both can be misleading, since they do not fully consider the size of the four classes of the confusion matrix in their final score computation.

Suppose, for example, you have a very imbalanced validation set made of 100 elements, 95 of which are positive elements, and only 5 are negative elements (as explained in Tip 5). And suppose also you made some mistakes in designing and training your machine learning classifier, and now you have an algorithm which always predicts positive. Imagine that you are not aware of this issue.

By applying your only-positive predictor to your imbalanced validation set, therefore, you obtain values for the confusion matrix categories:

These values lead to the following performance scores: accuracy = 95%, and F1 score = 97.44%. By reading these over-optimistic scores, then you will be very happy and will think that your machine learning algorithm is doing an excellent job. Obviously, you would be on the wrong track.

On the contrary, to avoid these dangerous misleading illusions, there is another performance score that you can exploit: the Matthews correlation coefficient [40] (MCC).

(Equation 3, MCC: worst value = −1; best value = +1).

By considering the proportion of each class of the confusion matrix in its formula, its score is high only if your classifier is doing well on both the negative and the positive elements.

In the example above, the MCC score would be undefined (since TN and FN would be 0, therefore the denominator of Equation 3 would be 0). By checking this value, instead of accuracy and F1 score, you would then be able to notice that your classifier is going in the wrong direction, and you would become aware that there are issues you ought to solve before proceeding.

Consider this other example. You ran a classification on the same dataset which led to the following values for the confusion matrix categories:

In this example, the classifier has performed well in classifying positive instances, but was not able to correctly recognize negative data elements. Again, the resulting F1 score and accuracy scores would be extremely high: accuracy = 91%, and F1 score = 95.24%. Similarly to the previous case, if a researcher analyzed only these two score indicators, without considering the MCC, they would wrongly think the algorithm is performing quite well in its task, and would have the illusion of being successful.

On the other hand, checking the Matthews correlation coefficient would be pivotal once again. In this example, the value of the MCC would be 0.14 (Equation 3), indicating that the algorithm is performing similarly to random guessing. Acting as an alarm, the MCC would be able to inform the data mining practitioner that the statistical model is performing poorly.

For these reasons, we strongly encourage to evaluate each test performance through the Matthews correlation coefficient (MCC), instead of the accuracy and the F1 score, for any binary classification problem.

— Davide Chicco, Ten quick tips for machine learning in computational biology [14]

Chicco's passage might be read as endorsing the MCC score in cases with imbalanced data sets. This, however, is contested; in particular, Zhu (2020) offers a strong rebuttal. [29]

Note that the F1 score depends on which class is defined as the positive class. In the first example above, the F1 score is high because the majority class is defined as the positive class. Inverting the positive and negative classes results in the following confusion matrix:

This gives an F1 score = 0%.

The MCC doesn't depend on which class is the positive one, which has the advantage over the F1 score to avoid incorrectly defining the positive class.

See also

Cohen's kappa

Contingency table

Cramér's V , a similar measure of association between nominal variables.

F1 score

Fowlkes–Mallows index

Polychoric correlation (subtype: Tetrachoric correlation), when variables are seen as dichotomized versions of (latent) continuous variables

References

v

t

e

Outline

Index

Mean Arithmetic Arithmetic-Geometric Contraharmonic Cubic Generalized/power Geometric Harmonic Heronian Heinz Lehmer

Arithmetic

Arithmetic-Geometric

Contraharmonic

Cubic

Generalized/power

Geometric

Harmonic

Heronian

Heinz

Lehmer

Median

Mode

Average absolute deviation

Coefficient of variation

Interquartile range

Percentile

Range

Standard deviation

Variance

Central limit theorem

Moments Kurtosis L-moments Skewness

Kurtosis

L-moments

Skewness

Index of dispersion

Contingency table

Frequency distribution

Grouped data

Partial correlation

Pearson product-moment correlation

Rank correlation Kendall's τ Spearman's ρ

Kendall's τ

Spearman's ρ

Scatter plot

Bar chart

Biplot

Box plot

Control chart

Correlogram

Fan chart

Forest plot

Histogram

Pie chart

Q–Q plot

Radar chart

Run chart

Scatter plot

Stem-and-leaf display

Violin plot

Effect size

Missing data

Optimal design

Population

Replication

Sample size determination

Statistic

Statistical power

Sampling Cluster Stratified

Cluster

Stratified

Opinion poll

Questionnaire

Standard error

Blocking
Factorial experiment
Interaction
Random assignment
Randomized controlled trial
Randomized experiment
Scientific control
Adaptive clinical trial
Stochastic approximation
Up-and-down designs
Cohort study
Cross-sectional study
Natural experiment
Quasi-experiment
Population
Statistic
Probability distribution
Sampling distribution Order statistic
Order statistic
Empirical distribution Density estimation
Density estimation
Statistical model Model specification L p space
Model specification
L p space
Parameter location scale shape
location
scale
shape
Parametric family Likelihood (monotone) Location–scale family Exponential family
Likelihood (monotone)
Location–scale family
Exponential family
Completeness
Sufficiency
Statistical functional Bootstrap U V
Bootstrap
U
V

Optimal decision loss function

loss function

Efficiency

Statistical distance divergence

divergence

Asymptotics

Robustness

Estimating equations Maximum likelihood Method of moments M-estimator Minimum distance

Maximum likelihood

Method of moments

M-estimator

Minimum distance

Unbiased estimators Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem Median unbiased

Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem

Rao–Blackwellization

Lehmann–Scheffé theorem

Median unbiased

Plug-in

Confidence interval

Pivot

Likelihood interval

Prediction interval

Tolerance interval

Resampling Bootstrap Jackknife

Bootstrap

Jackknife

1- & 2-tails

Power Uniformly most powerful test

Uniformly most powerful test

Permutation test Randomization test

Randomization test

Multiple comparisons

Likelihood-ratio

Score/Lagrange multiplier

Wald

Z -test (normal)

Student's t -test

F -test

Chi-squared

G -test

Kolmogorov–Smirnov

Anderson–Darling

Lilliefors

Jarque–Bera

Normality (Shapiro–Wilk)

Likelihood-ratio test

Model selection Cross validation AIC BIC

Cross validation

AIC

BIC

Sign Sample median

Sample median

Signed rank (Wilcoxon) Hodges–Lehmann estimator

Hodges–Lehmann estimator

Rank sum (Mann–Whitney)

Nonparametric anova 1-way (Kruskal–Wallis) 2-way (Friedman) Ordered alternative (Jonckheere–Terpstra)

1-way (Kruskal–Wallis)

2-way (Friedman)

Ordered alternative (Jonckheere–Terpstra)

Van der Waerden test

Bayesian probability prior posterior

prior

posterior

Credible interval

Bayes factor

Bayesian estimator Maximum posterior estimator

Maximum posterior estimator

Correlation

Regression analysis

Pearson product-moment

Partial correlation

Confounding variable

Coefficient of determination

Errors and residuals

Regression validation
Mixed effects models
Simultaneous equations models
Multivariate adaptive regression splines (MARS)
Simple linear regression
Ordinary least squares
General linear model
Bayesian regression
Nonlinear regression
Nonparametric
Semiparametric
Isotonic
Robust
Homoscedasticity and Heteroscedasticity
Exponential families
Logistic (Bernoulli) / Binomial / Poisson regressions
Analysis of variance (ANOVA, anova)
Analysis of covariance
Multivariate ANOVA
Degrees of freedom
Cohen's kappa
Contingency table
Graphical model
Log-linear model
McNemar's test
Cochran–Mantel–Haenszel statistics
Regression
Manova
Principal components
Canonical correlation
Discriminant analysis
Cluster analysis
Classification
Structural equation model Factor analysis
Factor analysis
Multivariate distributions Elliptical distributions Normal
Elliptical distributions Normal
Normal

Decomposition

Trend

Stationarity

Seasonal adjustment

Exponential smoothing

Cointegration

Structural break

Granger causality

Dickey–Fuller

Johansen

Q-statistic (Ljung–Box)

Durbin–Watson

Breusch–Godfrey

Autocorrelation (ACF) partial (PACF)

partial (PACF)

Cross-correlation (XCF)

ARMA model

ARIMA model (Box–Jenkins)

Autoregressive conditional heteroskedasticity (ARCH)

Vector autoregression (VAR) (Autoregressive model (AR))

Spectral density estimation

Fourier analysis

Least-squares spectral analysis

Wavelet

Whittle likelihood

Kaplan–Meier estimator (product limit)

Proportional hazards models

Accelerated failure time (AFT) model

First hitting time

Nelson–Aalen estimator

Log-rank test

Bioinformatics

Clinical trials / studies

Epidemiology

Medical statistics

Chemometrics

Methods engineering

Probabilistic design

Process / quality control
Reliability
System identification
Actuarial science
Census
Crime statistics
Demography
Econometrics
Jurimetrics
National accounts
Official statistics
Population statistics
Psychometrics
Cartography
Environmental statistics
Geographic information system
Geostatistics
Kriging
Category
Mathematics portal
Commons
WikiProject
v
t
e
MSE
MAE
sMAPE
MAPE
MASE
MSPE
RMS
RMSE/RMSD
R 2
MDA
MAD
F-score
P4

Accuracy
Precision
Recall
Kappa
MCC
AUC
ROC
Sensitivity and specificity
Logarithmic loss
Silhouette
Calinski–Harabasz index
Davies–Bouldin index
Dunn index
Hopkins statistic
Jaccard index
Rand index
Similarity measure
SMC
DBCV index
MRR
NDCG
AP
PSNR
SSIM
IoU
Perplexity
BLEU
Inception score
FID
Coverage
Intra-list similarity
Cosine similarity
Euclidean distance
Pearson correlation coefficient
Confusion matrix