



R news and tutorials contributed by (573) R bloggers

- [Home](#)
- [About](#)
- [RSS](#)
- [add your blog!](#)
- [R jobs](#) ♦♦♦♦
- [Contact us](#)

Welcome!

Follow @rbloggers 25K f

Here you will find daily **news and tutorials about R**, contributed by over 573 bloggers. There are many ways to **follow us** -

[By e-mail:](#)

Your e-mail here

 21,242 readers

BY FEEDBURNER

[On Facebook:](#)

R blog...

 28k likes

 Be the first of your friends to like this

If you are an R blogger yourself you are invited to [add your own R content feed to this site](#) (Non-English R bloggers should add themselves-[here](#))

[Jobs for R-users](#)

- [Looking for a Freelance R-Developer with Shiny experience](#)
- [Data Analyst @ New York](#)
- [R/Shiny App with d3 \(small job, quick turnaround, \\$250 < 4hrs\)](#)
- [Postdoctoral position \(Duke's River center\) in data visualization and ecosystem science @ Durham, North Carolina, United States](#)
- [Software Engineer for The Computational Biology Program at Oregon Health and Science University @ Portland, Oregon, United States](#)

Popular Searches

- [heatmap](#)
- [web scraping](#)
- [maps](#)
- [hadoop](#)
- [twitter](#)
- [alt=](#)
- [boxplot](#)
- [shiny](#)
- [time series](#)
- [animation](#)
- [ggplot2](#)
- [ggplot](#)
- [how to import image file to R](#)
- [latex](#)
- [trading](#)
- [finance](#)
- [PCA](#)
- [excel](#)
- [googlevis](#)
- [eclipse](#)
- [quantmod](#)
- [rstudio](#)
- [market research](#)
- [rattle](#)
- [Tutorial](#)
- [coplot](#)
- [rcmdr](#)
- [knitr](#)
- [title=](#)
- [rbloggers](#)

Recent Posts

- [Partoos, Recommender Systems and More](#)
- [Using htmlwidgets with knitr and Jekyll](#)
- [Wind in Netherlands](#)
- [Bioenergetics in R Workshop](#)
- [Correlation and Linear Regression](#)
- [Linear model with time series random component](#)
- [James Bond movies](#)
- [What it means to be a US Veteran Today](#)
- [Blog Post at Pluralsight](#)
- [The Lady Loves Statistics](#)
- [Annotables: R data package for annotating/convertig Gene IDs](#)
- [Szkolenie z analizy sieciowej](#)
- [Applied Statistical Theory: Quantile Regression](#)
- [Let's meet on SatRdays: the link between RUGs and conferences](#)
- [importance sampling with infinite variance](#)

Other sites

- [Statistics of Israel](#)
- [SAS blogs](#)
- [Jobs for R-users](#)

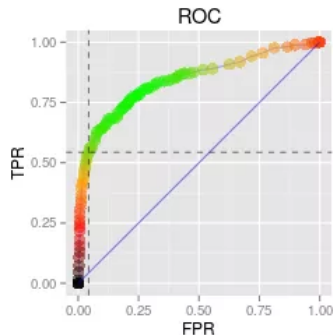
Illustrated Guide to ROC and AUC

June 23, 2015

By [Raffael Vogler](#)

Like Share 335 Tweet 117

(This article was first published on [joy of data > R](#), and kindly contributed to [R-bloggers](#))



(In a past job interview I failed at explaining how to [calculate and interpret ROC curves](#) – so here goes my attempt to fill this knowledge gap.) Think of a [regression model](#) mapping a number of features onto a real number (potentially a probability). The resulting real number can then be mapped on one of two classes, depending on whether this predicted number is greater or lower than some choosable threshold. Let's take for example a logistic regression and [data on the survivorship of the Titanic accident](#) to introduce the relevant concepts which will lead naturally to the ROC (Receiver Operating Characteristic) and its AUC or AUROC (Area Under ROC Curve).

Titanic Data Set and the Logistic Regression Model

Every record in the data set represents a passenger – providing information on her/his age, gender, class, number of siblings/spouses aboard (sibsp), number of parents/children aboard (parch) and, of course, whether s/he survived the accident.

```
# https://github.com/joyofdata/joyofdata-articles/blob/master/roc-auc/read_and_prepare_titanic_dataset.R
> df <- read_and_prepare_titanic_dataset("~/Downloads/titanic3.csv")
> str(df)

'data.frame':  1046 obs. of  6 variables:
 $ survived: Factor w/ 2 levels "0","1": 2 2 1 1 2 2 1 2 1 ...
 $ pclass  : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
 $ age     : num  29 0.92 2 30 25 48 63 39 53 71 ...
 $ sibsp   : int  0 1 1 1 1 0 0 1 0 2 0 ...
 $ parch   : int  0 2 2 2 2 0 0 0 0 0 ...
```

The logistic regression model is tested on batches of 10 cases with a model trained on the remaining N-10 cases – the test batches form a partition of the data. In short, [Leave-10-out CV](#) has been applied to arrive at more accurate estimation of the out-of-sample error rates.

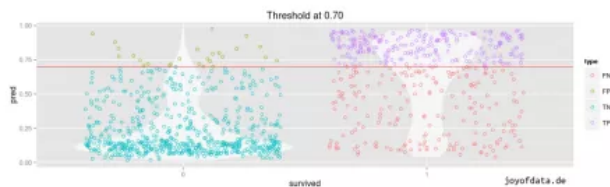
```
# https://github.com/joyofdata/joyofdata-articles/blob/master/roc-auc/log_reg.R
> predictions <- log_reg(df, size=10)
> str(predictions)

'data.frame':  1046 obs. of  2 variables:
 $ survived: Factor w/ 2 levels "0","1": 1 2 1 1 2 2 1 2 1 2 ...
 $ pred     : num  0.114 0.854 0.176 0.117 0.524 ...
```

Distribution of the Predictions

Now let's first have a look at the distribution of survival and death cases on the predicted survival probabilities.

```
# https://github.com/joyofdata/joyofdata-articles/blob/master/roc-auc/plot_pred_type_distribution.R
> plot_pred_type_distribution(predictions, 0.7)
```



If we consider survival as a positive (1) and death due to the accident as a negative (0) result, then the above plot illustrates the tradeoff we face upon choosing a reasonable threshold. If we increase the threshold the number of false positive (FP) results is lowered, while the number of false negative (FN) results increases.

117

85

28

2

Receiver Operating Characteristic

This question of how to balance false positives and false negatives (depending on the cost/consequences of either mistake) arose on a major scale during World War II in context of interpretation of radar signals for identification of enemy air planes. For the purpose of visualizing and quantifying the impact of a threshold on the FP/FN-tradeoff the ROC curve was introduced. The ROC curve is the interpolated curve made of points whose coordinates are functions of the threshold:

threshold = $\theta \in \mathbb{R}$, here $\theta \in [0, 1]$

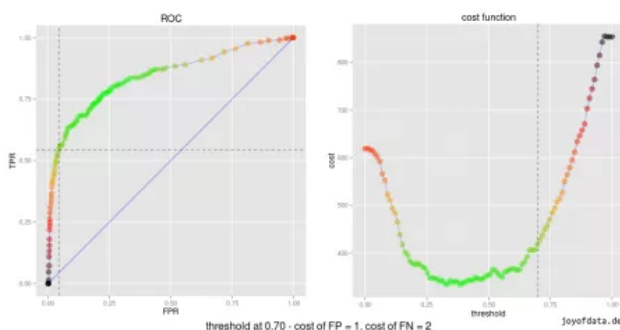
$$ROC_x(\theta) = FPR(\theta) = \frac{FP(\theta)}{FP(\theta) + TN(\theta)} = \frac{FP(\theta)}{\#N}$$

$$ROC_y(\theta) = TPR(\theta) = \frac{TP(\theta)}{FN(\theta) + TP(\theta)} = \frac{FP(\theta)}{\#P} = 1 - \frac{FN(\theta)}{\#P} = 1 - FNR(\theta)$$

[In terms of hypothesis tests](#) where rejecting the null hypothesis is considered a positive result the FPR (false positive rate) corresponds to the Type I error, the FNR (false negative rate) to the Type II error and $(1 - FNR)$ to the power. So the ROC for above distribution of predictions would be:

```
# https://github.com/joyofdata/joyofdata-articles/blob/master/roc-auc/calculate_roc.R
roc <- calculate_roc(predictions, 1, 2, n = 100)
```

```
# https://github.com/joyofdata/joyofdata-articles/blob/master/roc-auc/plot_roc.R
plot_roc(roc, 0.7, 1, 2)
```



The dashed lines indicate the location of the (FPR, TPR) corresponding to a threshold of 0.7. Note that the low corner (0,0) is associated with a threshold of 1 and the top corner (1,1) with a threshold of 0.

The cost function and the corresponding coloring of the ROC points illustrate that an optimal FPR and TPR combination is determined by the associated cost. Depending on the use case false negatives might be more costly than false positive or vice versa. Here I assumed a cost of 1 for FP cases and a cost of 2 for FN cases.

Area Under (ROC) Curve

The optimal point on the ROC curve is $(FPR, TPR) = (0,1)$. No false positives and all true positives. So the closer we get there the better. The second essential observation is that the curve is by definition monotonically increasing.

$$FPR(\theta) < FPR(\theta') \implies \theta > \theta' \implies TPR(\theta) \leq TPR(\theta')$$

This inequation can be easily checked by looking at the first plot by mentally pushing the threshold (red line) up and down; it implies the monotonicity. Furthermore any reasonable model's ROC is located above the identity line as a point below it would imply a prediction performance worse than

random (in that case, simply inverting the predicted classes would bring us to the sunny side of the ROC space).

All those features combined make it apparently reasonable to summarize the ROC into a single value by calculating the area of the convex shape below the ROC curve – this is the AUC. The closer the ROC gets to the optimal point of perfect prediction the closer the AUC gets to 1.

AUC for the example

```
> library(pROC)
> auc(predictions$survived, predictions$pred)
```

Area under the curve: 0.8421

ROC and AUC for Comparison of Classifiers

Mainly two reasons are responsible for why an ROC curve is a potentially powerful metric for comparison of different classifiers. One is that the resulting ROC is invariant against class skew of the applied data set – that means a data set featuring 60% positive labels will yield the same (statistically expected) ROC as a data set featuring 45% positive labels (though this will affect the cost associated with a given point of the ROC). The other is that the ROC is invariant against the evaluated score – which means that we could compare a model giving non-calibrated scores like a regular linear regression with a logistic regression or a random forest model whose scores can be considered as class probabilities.

The AUC furthermore offers interesting interpretations:

The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. *[Fawcett]*

[The AUC] also has various natural intuitive interpretations, one of which is that it is the average sensitivity of a classifier under the assumption that one is equally likely to choose any value of the specificity — under the assumption of a uniform distribution over specificity. *[Hand]*

As the ROC itself is variable with respect to a given data set it is necessary to average multiple ROCs derived from different data sets to arrive at a good estimation of a classifier's true ROC function.

stay tuned -->



Criticism of the AUC

It seems problematic, in the first place, to absolutely measure and compare the performance of classifiers with something as simple as a scalar between 0 and 1. The main fundamental reason of this is that problem specific cost functions hurt the assumption of points in the ROC space being homogenous in that regard and by that comparable across classifiers. This non-uniformity of the cost function causes ambiguities if ROC curves of different classifiers cross and on itself when the ROC curve is compressed into the AUC by means of integration over the false positive rate.

However, the AUC also has a much more serious deficiency, and one which appears not to have been previously recognised. This is that it is fundamentally incoherent in terms of misclassification costs: the AUC uses different misclassification cost distributions for different classifiers. This means that using the AUC is equivalent to using different metrics to evaluate different classification rules. It is equivalent to saying that, using one classifier, misclassifying a class 1 point is p times as serious as misclassifying a class 0 point, but, using another classifier, misclassifying a class 1 point is P times as

serious, where $p = P$. This is nonsensical because the relative severities of different kinds of misclassifications of individual points is a property of the problem, not the classifiers which happen to have been chosen. [Hand]

[David J. Hand](#) gives a statistically profound reasoning for the dubiousness of the AUC.

Sources

[Fawcett]: [“An introduction to ROC analysis” by Tom Fawcett](#)

[Hand]: [“Measuring classifier performance: a coherent alternative to the area under the ROC curve” by David J. Hand](#)

(original article published on www.joyofdata.de)

Like Share 335 Tweet 117

335 117
Like Tweet
Share

To **leave a comment** for the author, please follow the link and comment on their blog:
[joy of data » R](#).

R-bloggers.com offers [daily e-mail updates](#) about R news and [tutorials](#) on topics such as: [Data science](#), [Big Data](#), [R jobs](#), visualization ([ggplot2](#), [Boxplots](#), [maps](#), [animation](#)), programming ([RStudio](#), [Sweave](#), [LaTeX](#), [SQL](#), [Eclipse](#), [git](#), [hadoop](#), [Web Scraping](#)) statistics ([regression](#), [PCA](#), [time series](#), [trading](#)) and more...

If you got this far, why not **subscribe for updates** from the site?
Choose your flavor: [e-mail](#), [twitter](#), [RSS](#), or [facebook](#)...

Like Share 335 Tweet 117

Comments are closed.

Top 3 Posts from the past 2 days

- [James Bond movies](#)
- [Correlation and Linear Regression](#)
- [Linear model with time series random component](#)

Search & Hit Enter

Top 9 articles of the week

1. [Installing R packages](#)
2. [The Data Science Industry: Who Does What \(Infographic\)](#)
3. [In-depth introduction to machine learning in 15 hours of expert videos](#)
4. [magrittr: The best thing to have ever happened to R?](#)
5. [Using apply, sapply, lapply in R](#)
6. [How to Make a Histogram with Basic R](#)
7. [Introducing Distributed Data-structures in R](#)
8. [James Bond movies](#)
9. [Adding a legend to a plot](#)


Sponsors



R Consulting, Training, Support and
Application Development

Highland Statistics Ltd

Zero Inflated Models & GLMM
 Beginner's Guide to GAM
 Beginner's Guide to GLM & GLMM
 Beginner's Guide to GAMM



R training
 R consulting

QUANTIDE
 knowledge from data

R Studio
 open source & enterprise ready
 professional software for R

REVOLUTION
 ANALYTICS

R for the Enterprise

www.revolutionanalytics.com

Werden Sie zum Expe[R]ten mit der
 R-Akademie von

eoda
 daten • wissen • nutzen

Beratung | Software
 Training | Lösungen

plotly online R graphing

[Plotly: collaborative, publication-quality graphing.](http://plot.ly)

STATISTICS
 VIEWS

Bringing Statistics Together

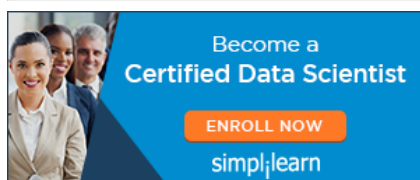
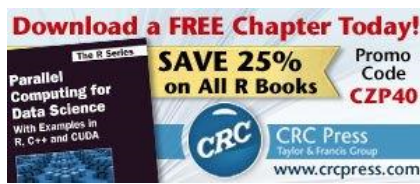
NYC DATA SCIENCE ACADEMY

January 11 - April 1, 2016 | Full Time Program

12 - WEEK DATA SCIENCE BOOTCAMP

Hands-on R, Python & Hadoop Training
 Bootstrap your data science career
 Receive personal and team support in job search
 Become part of an engaged Data Science community

Early Application Deadline: Nov 16, 2015
 Final Application Deadline: Dec 7, 2015



[Contact us](#) if you wish to help support R-bloggers, and place your banner here.

[Jobs for R users](#)

- [Looking for a Freelance R-Developer with Shiny experience](#)

- [Data Analyst @ New York](#)
- [R/Shiny App with d3 \(small job, quick turnaround, \\$250 < 4hrs\)](#)
- [Postdoctoral position \(Duke's River center\) in data visualization and ecosystem science @ Durham, North Carolina, United States](#)
- [Software Engineer for The Computational Biology Program at Oregon Health and Science University @ Portland, Oregon, United States](#)
- [Data Scientist @ Israel](#)
- [Jr Financial Engineer](#)

[Full list of contributing R-bloggers](#)

[R-bloggers](#) was founded by [Tal Galili](#), with gratitude to the [R](#) community.

Is powered by [WordPress](#) using a [bavotasan.com](#) design.

Copyright © 2015 **R-bloggers**. All Rights Reserved. [Terms and Conditions](#) for this website