Open in app ↗

Search Medium

Get unlimited access to the best of Medium for less than $1/week. Become a member

CRPS — A Scoring Function for Bayesian Machine Learning Models

The Continuous Ranked Probability Score is a statistical metric that compares distributional predictions to ground-truth values

Itamar Faran · Follow

Published in Towards Data Science

5 min read · Jan 28

Listen Share More

An important part of the machine learning workflow is the model evaluation. The process itself can be considered common knowledge: split the data into train and test sets, train the model on the train set, and evaluate its performance on the test set using a score function.

The score function (or metric) is a mapping of the ground truth values and their predictions into a single and comparable value [1]. For example, for continuous predictions one could use score functions such as the RMSE, MAE, MAPE or R-squared. But what if the prediction is not a point-wise estimate, but a distribution?

In Bayesian machine learning, the predictions are often not point-wise estimates but distributions of values. For example, the prediction could be estimated parameters of a distribution, or, in the non-parametric case—an array of samples from an MCMC method.

In these cases, traditional score functions do not suit the statistical design; one could aggregate the predicted distributions into their mean or median values, but

that would result with a great loss of information regarding the dispersion and shape of the predicted distribution.

The Continuous Ranked Probability Score

The CRPS — Continuous Ranked Probability Score — is a score function that compares a single ground truth value to a Cumulative Distribution Function (CDF):

$$CRPS\left(F, y\right) = \int \left(F\left(x\right) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx$$

Definition of the CRPS [1]. Image by author.

First introduced in the 70's [4] and primarily used in weather forecasts, it is now gaining renewed attention in the literature and industry [1] [6]. It can be used as a metric to evaluate a model's performance when the target variable is continuous and the model predicts the target's distribution; Examples include Bayesian Regression or Bayesian Time Series models [5].

The fact that the theoretical definition includes the CDF makes the CRPS useful for both parametric and non-parametric predictions: for many distributions there is an analytic expression for the CRPS [3], and for non-parametric predictions, one could use the CRPS with the Empirical Cumulative Distribution Function (eCDF).

After computing the CRPS for each observation in our test set, we are left to aggregate the results into a single value. Similarly to the RMSE and MAE, we'll aggregate them using a (possibly weighted) average:

$$\sum_i w_i \cdot \int \left(\hat{F}_i\left(x\right) - \mathbf{1}_{\{x \geq y_i\}}\right)^2 dx; \quad \sum_i w_i = 1$$

Aggregation of CRPS over the test set, with empirical CDFs. Image by author.

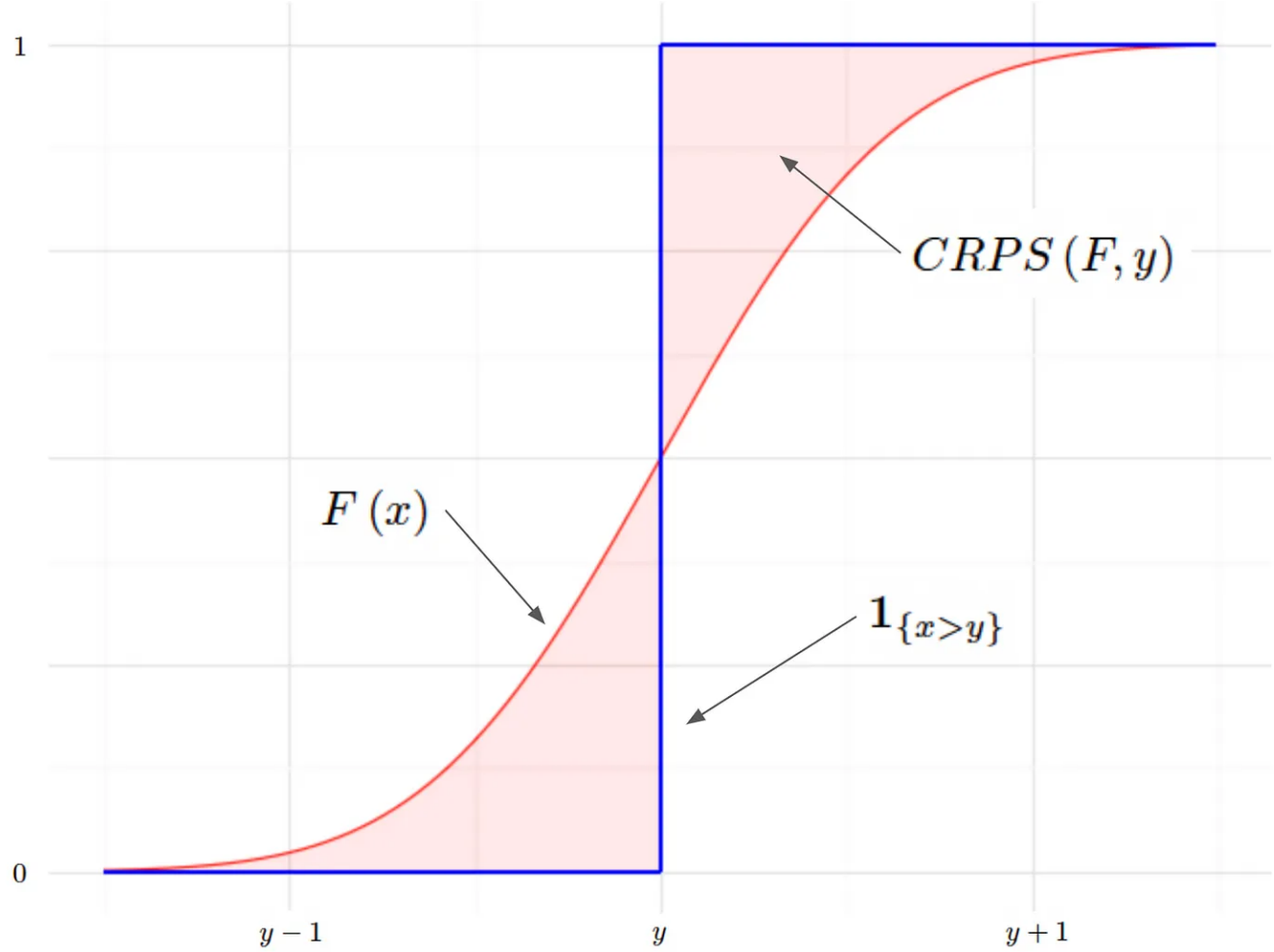
Intuition

The main challenge of comparing a single value to a distribution is how to translate the single value into the domain of distributions. The CRPS deals with that by translating the ground truth value into a degenerate distribution with the indicator function. For example, if our ground truth value is 7, we can translate it with:

$$P\left(7 \leq y\right) = \mathbf{1}_{\{y \geq 7\}} = \begin{cases} 0 & if\ y < 7 \\ 1 & else \end{cases}$$

Example of a degenerate distribution with an indicator function. Image by author.

The indicator function is a valid CDF answering all the requirements of a CDF. Now we are left with comparing the predicted distribution to the degenerate distribution of the ground truth value. Clearly, we want the predicted distribution to be as close as possible to the ground truth; this is expressed mathematically by measuring the (squared) area trapped between these two CDFs:



$$CRPS\left(F,y\right) = \int \left(F\left(x\right) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx$$

Visualization of the CRPS. The predicted distribution is marked in red, and the ground truth's degenerate distribution is marked in blue. The CRPS is the (squared) area trapped between the two CDFs. Image by author.

Relation to the MAE

The CRPS is closely related to the well-known MAE (Mean Absolute Error). If we take a point-wise prediction, treat it as a degenerate CDF and inject it into to the CRPS equation, we get:

$$
\begin{aligned}
&CRPS\left(\mathbf{1}_{\{x \geq \hat{y}\}}, y\right) \\
&= \int \left(\mathbf{1}_{\{x \geq \hat{y}\}} - \mathbf{1}_{\{x \geq y\}}\right)^2 dx \\
&= \begin{cases} \int_{\hat{y}}^{y} 1 dx & y > \hat{y} \\ \int_{y}^{\hat{y}} 1 dx & else \end{cases} \\
&= |y - \hat{y}|
\end{aligned}
$$

Relation between the CRPS and the MAE. Image by author.

So, if the predicted distribution is a degenerate distribution (e.g. a point-wise estimate), the CRPS reduces to the MAE. This helps to get another intuition for the CRPS: it can be viewed as a **generalization of the MAE into distributional predictions: The MAE is a special case of the CRPS** when the predicted distribution is degenerate.

Empirical Evaluation

When the model's prediction is a parametric distribution (e.g. the model predicts the distribution's parameters), the CRPS has an analytic expression for some common distributions [3]. For example, if the model predicts the parameters μ & σ of the Normal distribution, the CRPS can be calculated with:

$$
\begin{gathered}
CRPS\left(\mathcal{N}\left(\mu, \sigma^2\right), y\right) = \sigma\left(\omega\left(2 \cdot \Phi\left(\omega\right) - 1\right) + 2 \cdot \phi\left(\omega\right) - \pi^{-1/2}\right) \\
\omega = \left(y - \mu\right) / \sigma
\end{gathered}
$$

Analytical solution of the CRPS for the normal distribution [3]. Image by author.

Analytic solutions are known for distributions such as Beta, Gamma, Logistic, Log-Normal and others [3].

When the prediction is non-parametric, or more specifically — the prediction is an array of simulations, calculating the integral over the eCDF is a hefty task. However, the CRPS can also be analytically expressed by:

$$
\begin{aligned}
&CRPS\left(F, y\right) \\
&= \int \left(F\left(x\right) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx && \text{(INT)} \\
&= E\left[|X - y|\right] - \frac{1}{2} \cdot E\left[|X - X'|\right] && \text{(NRG)} \\
&= E\left[|X - y|\right] + E\left[X\right] - 2 \cdot E\left[X \cdot F\left(X\right)\right] && \text{(PWM)}
\end{aligned}
$$

Different forms of the CRPS and their names [2]. Image by author.

Where X, X' are independently and identically distributed according to F. These expressions, while still a bit computationally intensive, are simpler to estimate:

```python
import numpy as np


# Adapted to numpy from pyro.ops.stats.crps_empirical
# Copyright (c) 2017-2019 Uber Technologies, Inc.
# SPDX-License-Identifier: Apache-2.0
def crps(y_true, y_pred, sample_weight=None):
    num_samples = y_pred.shape[0]
    absolute_error = np.mean(np.abs(y_pred - y_true), axis=0)

    if num_samples == 1:
        return np.average(absolute_error, weights=sample_weight)

    y_pred = np.sort(y_pred, axis=0)
    diff = y_pred[1:] - y_pred[:-1]
    weight = np.arange(1, num_samples) * np.arange(num_samples - 1, 0, -1)
    weight = np.expand_dims(weight, -1)

    per_obs_crps = absolute_error - np.sum(diff * weight, axis=0) / num_samples**2
    return np.average(per_obs_crps, weights=sample_weight)
```

crps_nrg hosted with ❤️ by GitHub view raw

An implementation of the CRPS function according the NRG form [2]. Adapted from pytorch to numpy from pyro-ppl, Uber Technologies © [6]

```python
import numpy as np


def crps(y_true, y_pred, sample_weight=None):
    num_samples = y_pred.shape[0]
    absolute_error = np.mean(np.abs(y_pred - y_true), axis=0)
```

```
7
8       if num_samples == 1:
9           return np.average(absolute_error, weights=sample_weight)
10
11      y_pred = np.sort(y_pred, axis=0)
12      b0 = y_pred.mean(axis=0)
13      b1_values = y_pred * np.arange(num_samples).reshape((num_samples, 1))
14      b1 = b1_values.mean(axis=0) / num_samples
15
16      per_obs_crps = absolute_error + b0 - 2 * b1
17      return np.average(per_obs_crps, weights=sample_weight)
```

crps_pwm hosted with ❤️ by GitHub view raw

An implementation of the CRPS function according the PWM form [2].

You can check out an example on a Bayesian Ridge Regression in a Jupyter notebook here, where I demonstrate the usage of both the parametric and non-parametric CRPS.

Summary

The Continuous Ranked Probability Score (CRPS) is a scoring function that compares a single ground-truth value to its predicted distribution. This property makes it relevant to Bayesian machine learning, where models usually output distributional predictions rather than point-wise estimates. It can be viewed as a generalization of the well known MAE to distributional predictions.

It has analytical expressions for parametric predictions, and can be simply computed for non-parametric predictions. All together, the CRPS emerges as the new standard way to evaluate the performance of Bayesian machine learning models with a continuous target.

References

1. *Strictly Proper Scoring Rules, Prediction, and Estimation,* Gneiting & Raftery (2007)
2. *Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts*, Zamo & Naveau (2017)
3. *Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics*, Taillardat, Zamo & Naveau (2016)
4. *Scoring Rules for Continuous Probability Distributions*, Matheson & Winklers (1976)
5. *Distributional Regression and its Evaluation with the CRPS: Bounds and Convergence of the Minimax Risk,* Pic, Dombry, Naveau & Taillardat (2022)

6. CRPS Implementation in Pyro-PPL, Uber Technologies, Inc.

7. CRPS Implementation in properscoring, The Climate Corporation

Bayesian Machine Learning | Model Evaluation | Bayesian Statistics | Machine Learning | Probability

Follow

Written by Itamar Faran

178 Followers · Writer for Towards Data Science

Data Scientist & Statistician

More from Itamar Faran and Towards Data Science

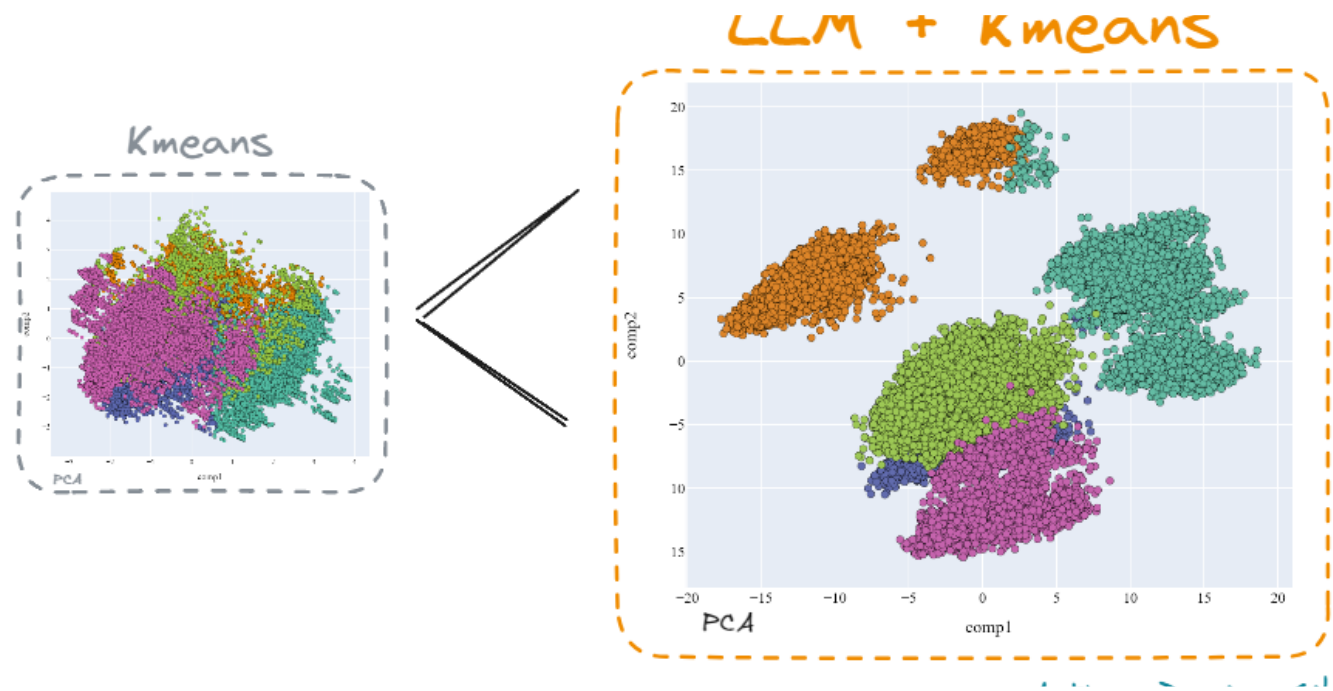
Itamar Faran in Geek Culture

Dask or Spark? A Comparison for Data Scientists

3 Reasons Why Dask is Better Suited for Data Science Projects (And When it is Not)

6 min read · Jun 4, 2022

290 2



Damian Gil in Towards Data Science

Mastering Customer Segmentation with LLM

Unlock advanced customer segmentation techniques using LLMs, and improve your clustering models with advanced techniques

23 min read · Sep 26

2.8K 24

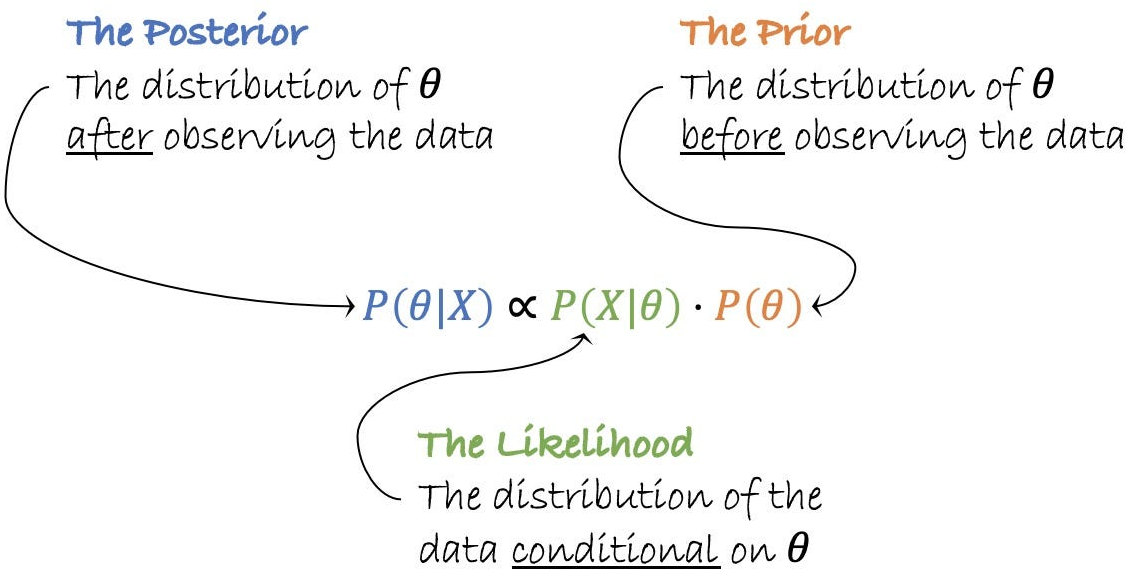
Khouloud El Alami in Towards Data Science

Don’t Start Your Data Science Journey Without These 5 Must-Do Steps From a Spotify Data Scientist

A complete guide to everything I wish I’d done before starting my Data Science journey, here’s to acing your first year with data

· 18 min read · Sep 24

2.1K 21



Itamar Faran in Towards Data Science

How To Do Bayesian A/B Testing, FAST!

Bayesian A/B Testing Without Compromising on Performance

10 min read · Jun 21, 2021

508 4

See all from Itamar Faran

See all from Towards Data Science

Recommended from Medium

Piero Paialunga in Towards Data Science

From Theory to Practice with Bayesian Neural Network, Using Python

Here's how to incorporate uncertainty in your Neural Networks, using a few lines of code

11 min read · Dec 21, 2022

360 11

Charlie Lai

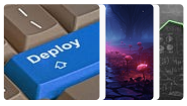
Probability Density Function is Not a Probability

In this article, I'd like to explain why a probability density function(pdf) is not a probability. In fact, a probability density function...

2 min read · Sep 24

32

Lists

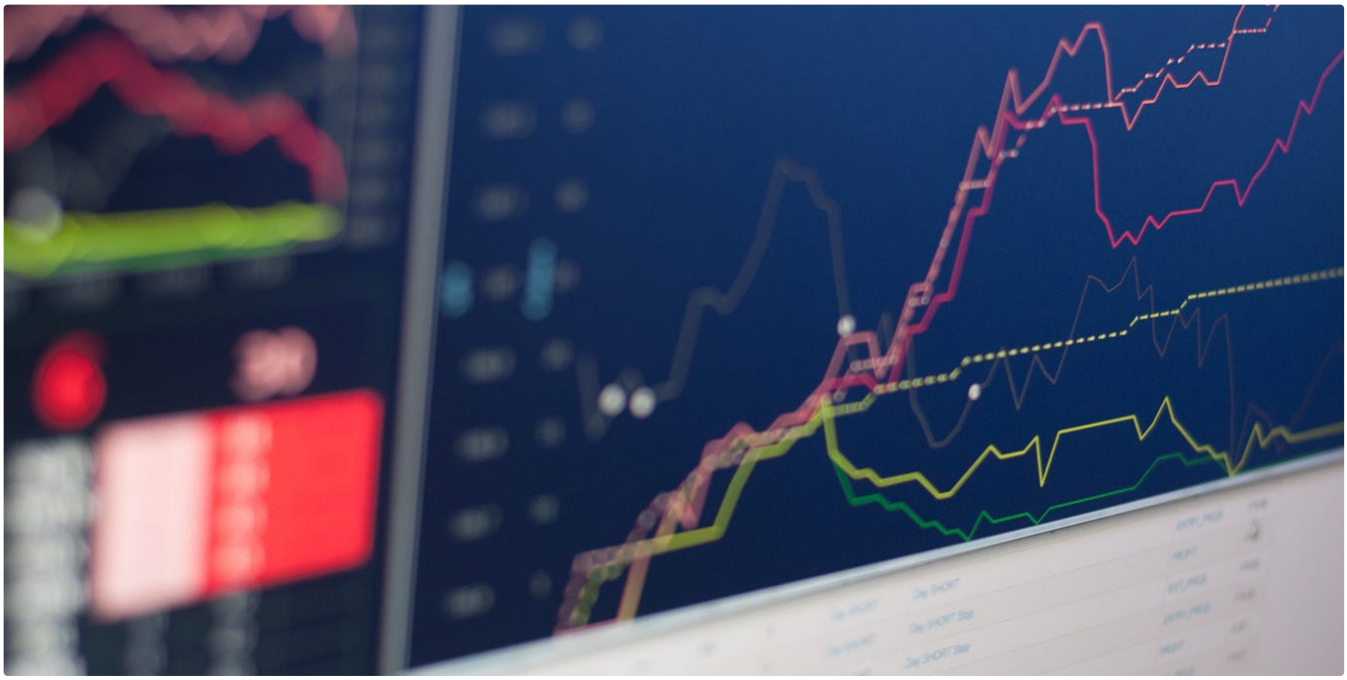


Predictive Modeling w/ Python
20 stories · 475 saves

Practical Guides to Machine Learning
10 stories · 549 saves

Natural Language Processing
689 stories · 305 saves

The New Chatbots: ChatGPT, Bard, and Beyond
13 stories · 138 saves



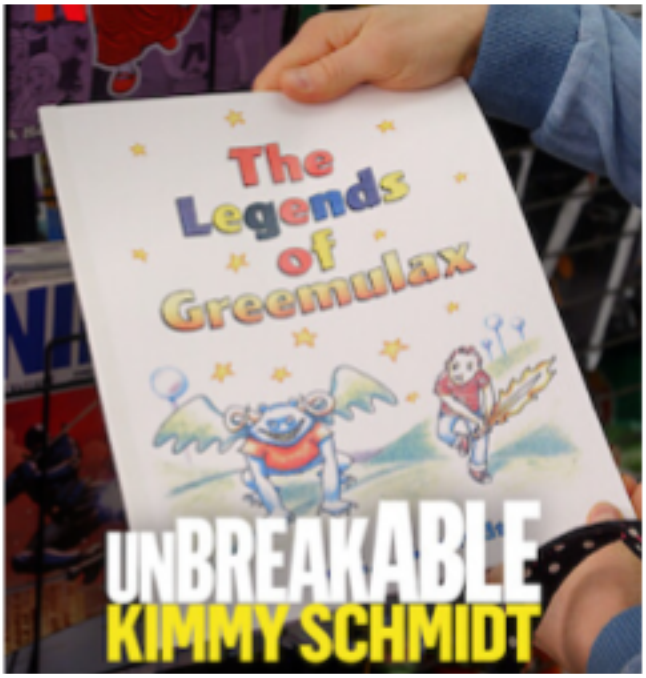
Chandra Prakash Bathula

Machine Learning Concept 68: Platt's Scaling

Platt's Scaling:

3 min read · Apr 12

5



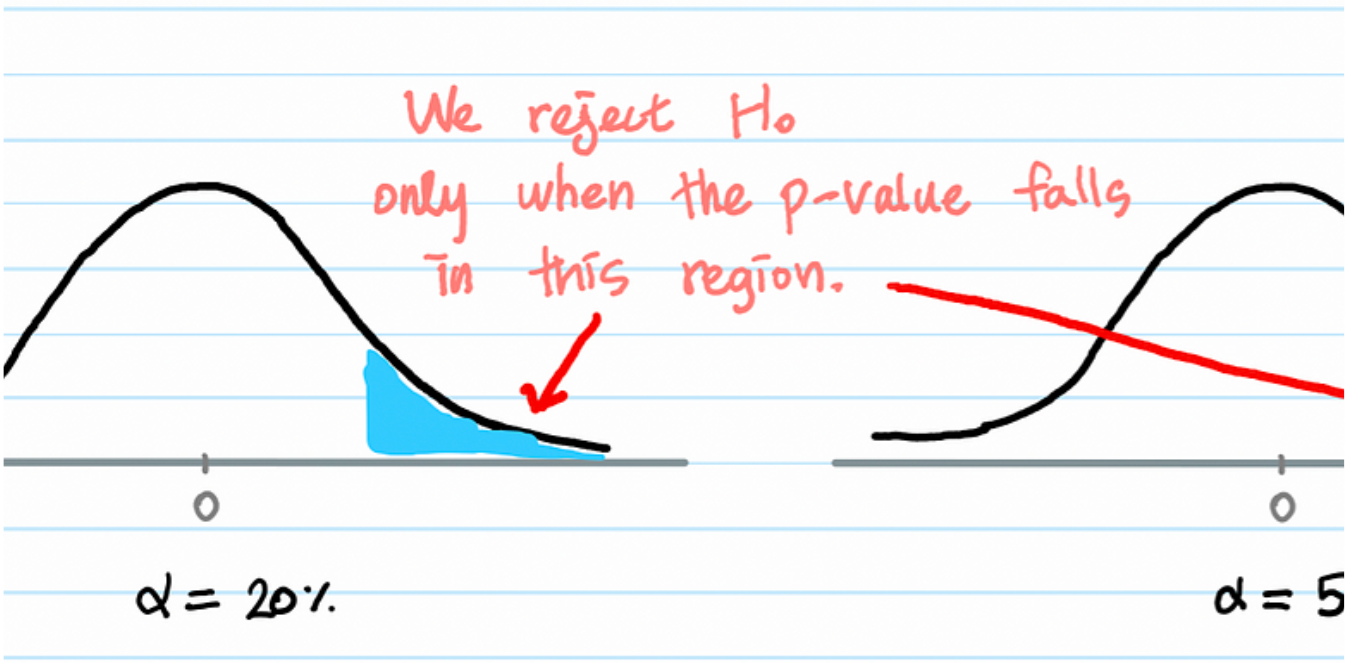
Netflix Technology Blog

Causal Machine Learning for Creative Insights

A framework to identify the causal impact of successful visual components.

13 min read · Jan 11

734 4



Ms Aerin in IntuitionMath

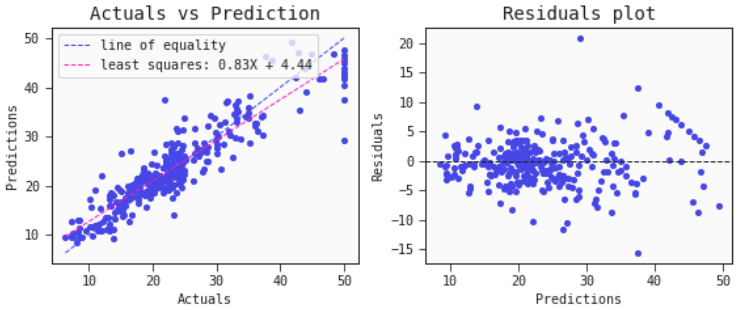
Chi Square Test—Intuition, Examples, and Step-by-Step Calculation

The best way to see if two variables are related.

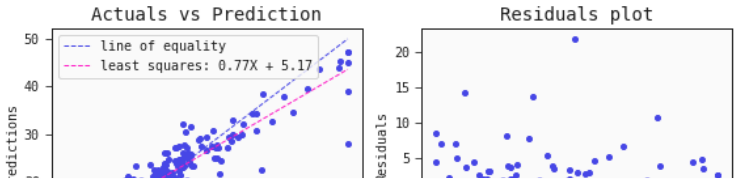
· 15 min read · Feb 12

407 3

Training Metrics



Test



Casper Skern Wilstrup

Symbolic Regression: a Simple and Friendly Introduction

Symbolic Regression is like a treasure hunt for the perfect mathematical equation to describe a dataset. Imagine having a bunch of data...

3 min read · May 5

18 1

See more recommendations