Probability calibration for boosted trees

Manish Chablani · Following
Published in Towards Data Science · 3 min read · Jul 22, 2017

Background: Seminal paper on probability calibration: http://www.cs.cornell.edu/~caruana/niculescu.scldbst.crc.rev4.pdf

Boosted decision trees typically yield good accuracy, precision, and ROC area. However, because the outputs from boosting are not well calibrated posterior probabilities, boosting yields poor squared error and cross-entropy. It tends to predict probabilities conservatively, meaning closer to mid-range than to extremes. Here is the effect of boosting on predicted probability:

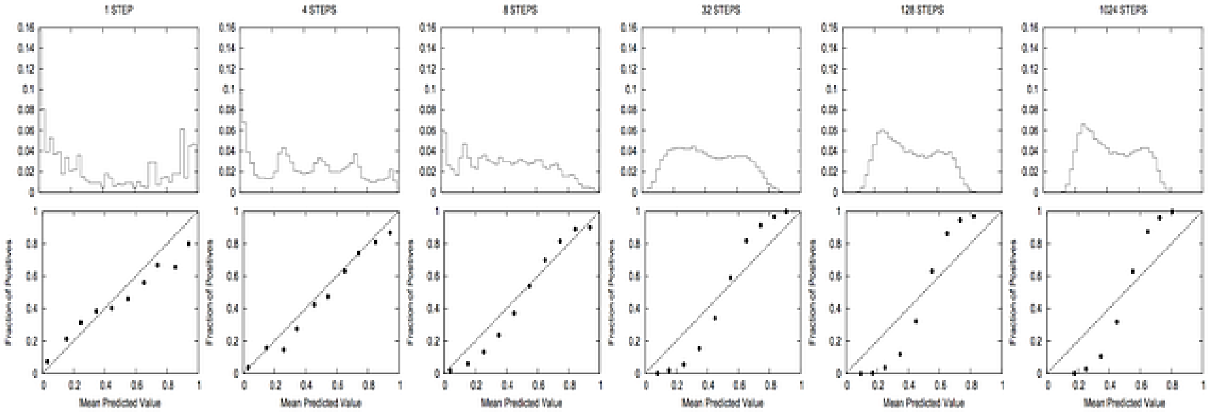
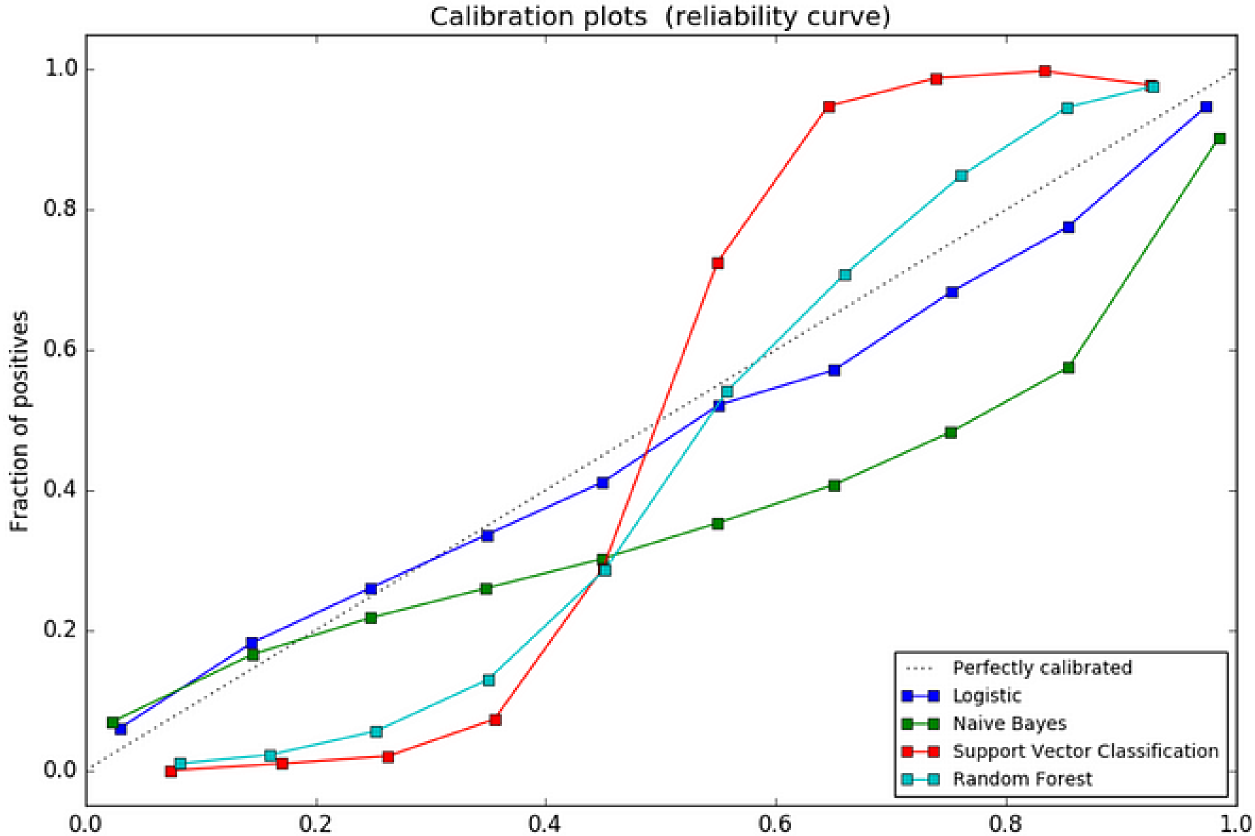


Figure 1: Effect of boosting on the predicted values. Histograms of the predicted values (top) and reliability diagrams (bottom) on the test set for boosted trees at different steps of boosting on the COV_TYPE problem.

Well calibrated classifiers are probabilistic classifiers for which the output of the model can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a prediction value close to 0.8, approximately 80% actually belong to the positive class. The following plot compares how well the probabilistic predictions of different classifiers are



LogisticRegression returns well calibrated predictions by default as it directly optimizes log-loss. In contrast, the other methods return biased probabilities; with different biases.

Two approaches for performing calibration of probabilistic predictions are: a parametric approach based on Platt's sigmoid model and a non-parametric approach based on isotonic regression. Probability calibration should be done using test/validation data that is not used for model fitting.

Platt's scaling:

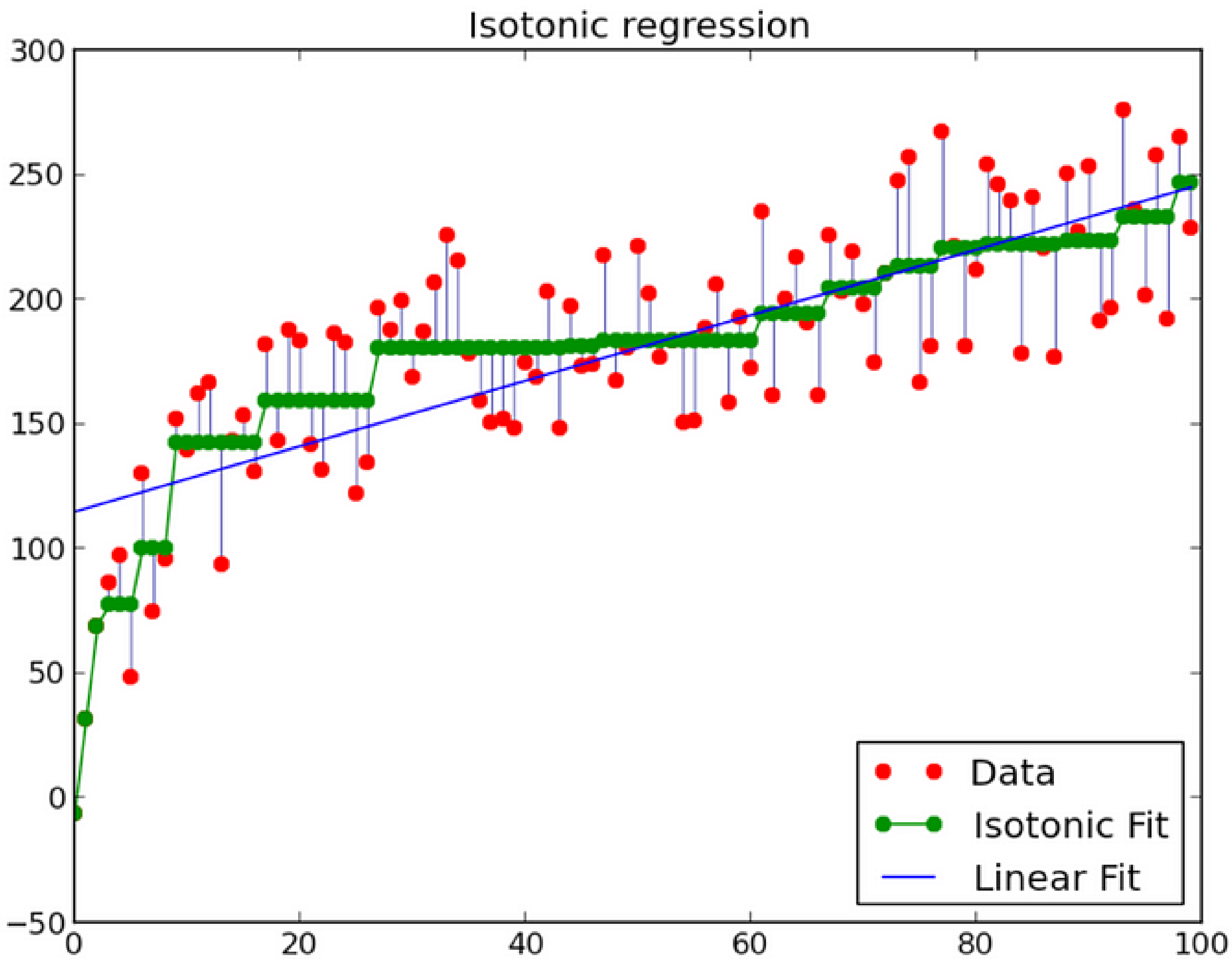
Platt's scaling amounts to training a logistic regression model on the classifier outputs.

You essentially create a new data set that has the same labels, but with one dimension (the probability output of the uncalibrated classifier. You then train on this new data set, and feed the probability output of the uncalibrated classifier as the input to this calibration method, which returns a calibrated probability. In Platt's case, we are essentially just performing logistic regression on the probability output of the uncalibrated classifier with respect to the true class labels.

You logistic regression model is f(x) = y. y is true label of input data and x is predicted probability of your base classifier. Now you use the predicted probability from logistic regression as your true probability from you calibrated classifier.

Isotonic regression:

The idea is to fit a piecewise-constant non-decreasing function instead of logistic regression. Piecewise-constant non-decreasing means stair-step shaped. Implemented via Pool Adjacent Violators Algorithm (PAVA). PAVA is a linear time (and linear memory) algorithm for linear ordering isotonic regression.



The way you train a isotonic regression is similar:

f(x) = y, y is true label of input data and x is predicted probability of your base classifier. Isotonic model is going to sort data by x(predicted probability of your base classifier) and then fit a step function to give probability of true label.

Other resources:

http://fastml.com/classifier-calibration-with-platts-scaling-and-isotonic-regression/

https://jmetzen.github.io/2015-04-14/calibration.html

https://scikit-learn.org/stable/modules/calibration.html

Machine Learning | Classifcation Models | Random Forest | Probability

More from the list: "ML"

Curated by Himanshu Birla

Kyosuke... in Towards Dat...
Probability Calibration for Imbalanced Dataset
8 min read · Oct 20, 2019

Mattia Ci... in Analytics Vi...
How Probability Calibration Works
6 min read · May 28, 2020

Jason Yo...
Why Calibrat the Series on
7 min read · Oc

View list

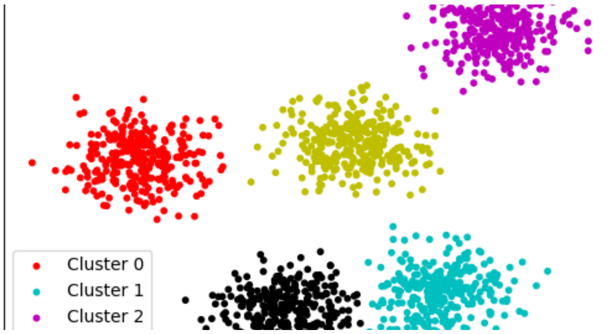
Written by Manish Chablani

Following

1.93K Followers · Writer for Towards Data Science

Head of AI @EightSleep , Marathoner. (Past: AI in healthcare @curaiHQ , DL for self driving cars @cruise , ML @Uber , Early engineer @MicrosoftAzure cloud

More from Manish Chablani and Towards Data Science



Manish Chablani in Towards Data Science

Semantic similarity classifier and clustering sentences based on...

Recently we have been doing some experiments to cluster semantically similar...

· 2 min read · Jun 30, 2019

137 6

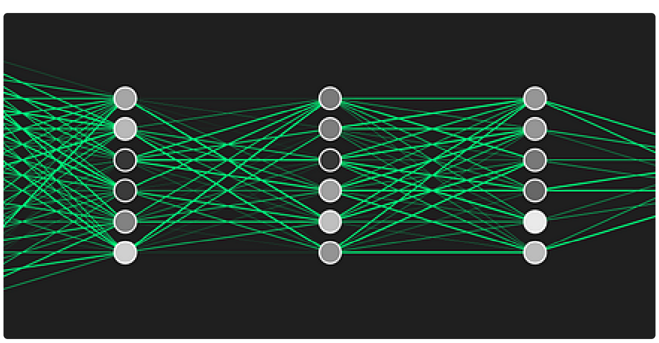
Antonis Makropoulos in Towards Data Science

How to Build a Multi-GPU System for Deep Learning in 2023

This story provides a guide on how to build a multi-GPU system for deep learning and...

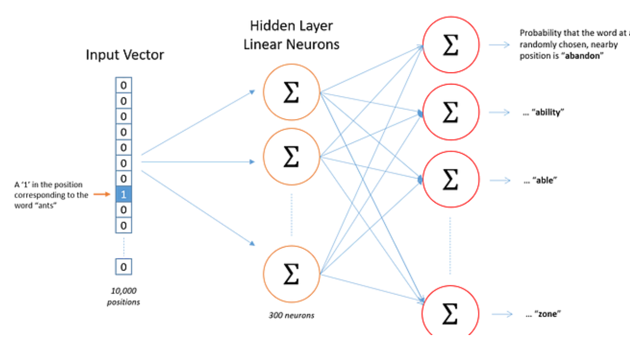
10 min read · Sep 17

552 11



Callum Bruce in Towards Data Science

How to Program a Neural Network



Manish Chablani in Towards Data Science

Word2Vec (skip-gram model): PART 1 - Intuition.

A step-by-step guide to implementing a neural network from scratch

· 14 min read · Sep 24

485 4

Most of the content here is from Chris's blog. I have condensed it and made minor...

· 7 min read · Jun 14, 2017

1.3K

See all from Manish Chablani

See all from Towards Data Science

Recommended from Medium

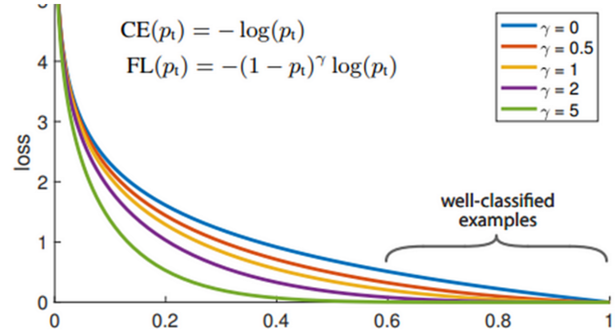
Chandra Prakash Bathula

Machine Learning Concept 68: Platt's Scaling

Platt's Scaling:

3 min read · Apr 13

5



Uman Niyaz in Data Science @ Ecom Express

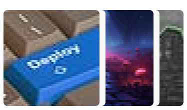
Focal loss for handling the issue of class imbalance

Text classification is widely used in various industries to tackle business challenges by...

9 min read · Jun 12

53

Lists



Predictive Modeling w/ Python

20 stories · 473 saves

Practical Guides to Machine Learning

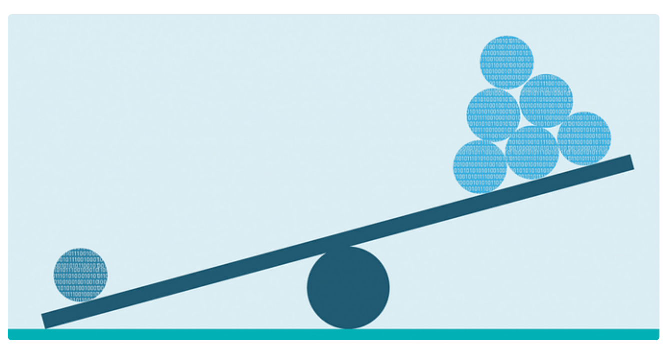
10 stories · 544 saves

Natural Language Processing

689 stories · 304 saves

The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 137 saves



Biman Chakraborty

Two-Sample t Tests, Power, Effect Size and Sample Size Calculator i...

I was going over my daily dose of coffee while redaing the newspaper in the morning. A...

13 min read · Apr 30

8

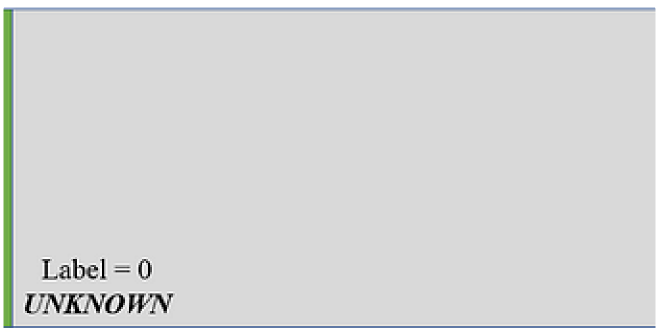
The Data Beast

Handling imbalanced data with XGboost:: Class_weight parameter

In XGBoost, the class_weight the parameter is used to adjust the weights of different classe...

2 min read · Apr 12

12



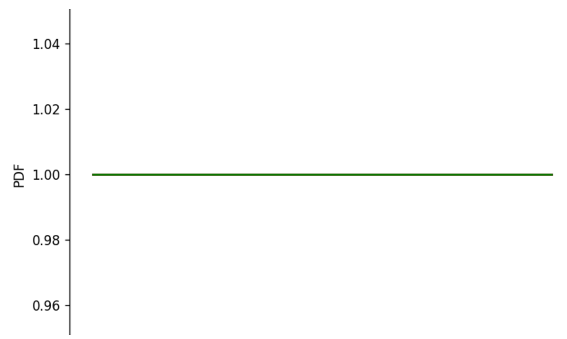
Aniruddha Mitra

Decoding PU Learning: A Fusion of Statistics and Machine Learning

Let's start with few blunt questions:

8 min read · Apr 18

9



Ashish Ranjan Karn

Thompson Sampling — Python Implementation

Thompson Sampling is a popular probabilistic algorithm used in decision-making under...

5 min read · Jul 20

5

See more recommendations