Trust Scores

Sarah Gillespie

12/16/2021

Trust scores' goal

Tell the person who runs an algorithm how reliable a specific prediction made by the model is.
It's somewhat like a machine learning version of a p-score except. . . a little worse. Trust scores are model specific to the model's data and an individual trust score provides little context of how reliable that prediction is: trust scores are

What's the math behind the trust score?

Trust Score Step 1: Preprocess the data

Keep only the points that exist in a high density area. That specific density varies but is defined as α as a general variable. This has the effect of removing outliers and having smooth clusters of points to use for calculating the trust score. The specific distribution of the smooth cluster of points is irrelevant. The trust scores paper does not recommend an α value to define what density counts as a high density set.

Trust Score Step 2: Calculate the trust score

The trust score compares the the distance from the testing sample to the nearest classes similar and different from the testing sample.

The trust score is calculated as:

(distance from the testing sample to the nearest class* different from the predicted class) / (distance from the testing sample to the nearest class the same as the predicted class)

*Note that "class" in this definition describes the grouping categorization of a single point (like, a specific penguin species if the goal is the determine the penguin's species)

Trust Score Step 3: is my prediction correct?

If trust score > 1, then the different class is geographically closer to the testing sample than its predicted class. This is a red flag for the testing sample's categorization potentially being incorrect.

Flexibility of this metric

Trust scores can have variations throughout different categories, just like creating a Build-A-Bear or picking out parts for a custom computer.

1. A specific test point. The trust score is about looking at the categorization of a single input's categorization compared to the training data.
2. A specific classifier model. For example, neural network, random forest, or logistic regression.
3. A specific neighbor metric to use to compare the different classes. For example, k-Nearest Neighbors, a single nearest neighbor, or a centroid.
4. A specific data setting. For example, raw inputs, an unsupervised embedding of the space, or individual layers.

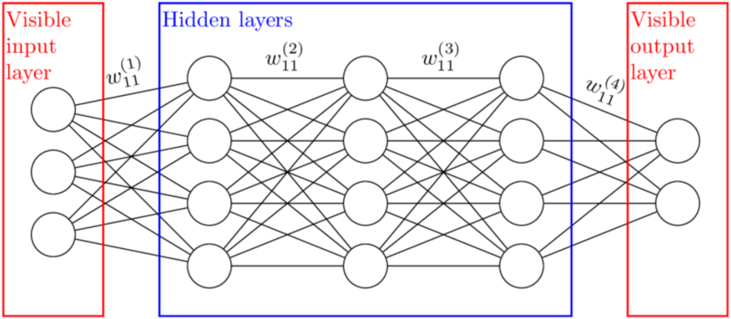
Metrics similar to trust scores:

Confidence and Credibility Scores

These scores are generally done on deep learning algorithms and calculated on each of the deep learning algorithm's layers.

Deep learning refresher

An example of a deep learning neural network with 3 hidden layers. Each layer is specified as a vector of binary components, with the edges between the vectors defined as a matrix of weight values.



Adcock, Jeremy & Allen, Euan & Day, Matthew & Frick, Stefan & Hinchliff, Janna & Johnson, Mack & Morley-Short, Sam & Pallister, Sam & Price, Alasdair & Stanisic, Stasja. (2015). Advances in quantum machine learning.

Confidence Scores

Confidence: quantifies the likelihood of the prediction being correct.

Confidence score = the distance between the test input and the model's nearest neighboring training points, summed for each layer to create a confidence score for each layer and then the summation of layer confidence scores can be added together to get a confidence score for the model as a whole.

The confidence score does not exist in an informational vacuum: an algorithm user most consider the confidence score relative to other inputs' confidence scores.

Confidence Score math

Let I be the model's input that is being analyzed in each hidden layer. The input point is located at (x_I, y_I) in each unique hidden layer.

n be a specific nearest neighbor point located at (x_n, y_n).

φ be the total nearest neighbor points considered in each layer.

We can use the distance formula, $\sqrt{(x_2 - x_1) + (y_2 - y_1)}$, as a base to compute a single layer's confidence score. Note that the (x, y) location is a simplistic representation of the point in a two-dimension Cartesian plane. This can vary but the mathematician must adapt the below formula for another representation of the two points. We sum up the distances between the input point and each nearest neighbor point to find a single hidden layer's confidence score.

$$\sum_{n=1}^{\varphi} \sqrt{(x_n - x_I) + (y_n - y_I)}$$

Credibility Score

Credibility: characterizes how relevant the training data is to the prediction.

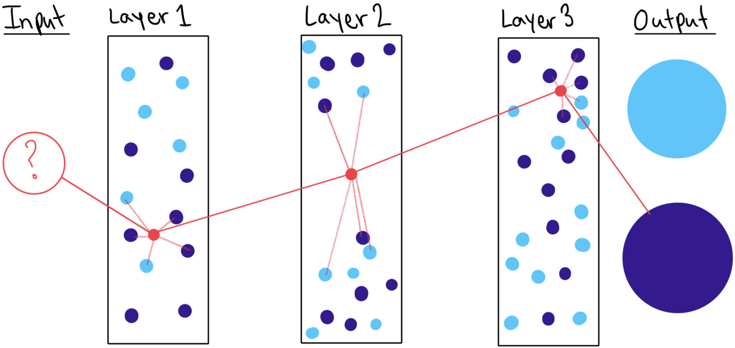
A large credibility score means that it is more likely the input is adversarial or otherwise wrongly classified. The credibility score range is [0, 1]. A score of 0 means that all of the point's nearest neighbors are the same group that the point's final classification is while a score of 1 means that none of the point's nearest neighbors are the same group as the point's final classification. The credibility score exists for an individual hidden layer and the model as a whole.

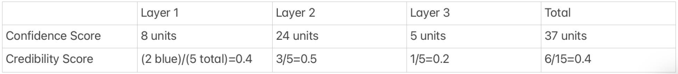
Credibility Score Math

φ: the total nearest neighbor points considered in each layer
λ: a specific hidden layer
Λ: the total number of hidden layers
with the addition of letting
ψ: the number of nearest neighbor points that are a different class than the input point's final classification.
A single layer's credibility score would be $\frac{\psi}{\varphi}$.
The credibility score for an entire model would be

$$\frac{\sum_{\lambda=1}^{\Lambda} \psi_\lambda}{\Lambda\varphi}$$

DkNN visualization





	Layer 1	Layer 2	Layer 3	Total
Confidence Score	8 units	24 units	5 units	37 units
Credibility Score	(2 blue)/(5 total)=0.4	3/5=0.5	1/5=0.2	6/15=0.4

Differences between the papers

During this week's reading, glance through the structure of the two difference papers. Trust scores, confidence scores, and credibility scores are all similar ideas but the trust scores authors did a much more through analysis of how trust scores relate to different probability distributions. The trust scores paper was presented at NeurIPS 2018 while the paper about confidence and credibility scores does not appear to have been presented at a conference.