Separability metric

Tom Rochette <tom.rochette@coreteks.org>

July 2, 2020 — 837eeaf7

0.1 Context

Given a tabular dataset, what is the theoretical limit on the number of entries we can accurately predict? Many real world problems are not functions (i.e., they are non-functional relations), that is to say that the same inputs may produce different outputs. Given a problem that should return the same output given the same inputs, what should be returned when an input has a set of different possible outputs?

0.2 Learned in this study

0.3 Things to explore

1 Overview

In machine learning, we try to minimize the concept of loss. When an input is known to generate different potential outputs, such a system will tend to choose the output with the most probability.

1.1 Setup

We are given a dataset of features, where we want to predict a feature Y given a set of features X. For some X, we may have different values of Y.

X	Y
1	3
1	4
1	3

1.2 Cases

In order to reflect on this problem, we use an iterative approach.

1.2.1 1 point

When a dataset is composed of 1 point, then we have 100% separability.

1.2.2 2 points

If X provides two different values, then we can separate the dataset into two distinct set, which leads us to 100% separability.

However, if X contains the same value, then the separability will depend on the target value:

- If the target value is the same for both points, then we have 100% separability, since both X lead to the same Y

- If the target value is different, then we only have 50% separability. This is due to the fact that, given no additional information, the best we can do is to randomly pick one of the two options for Y.

2 Notes

Number of unique input X/Number of points/rows
Given a tabular dataset, compute the separability metric as follow:

- 1 point: 100% separable
- 2 points: given the target feature to predict, if the other attributes can separate two distinct targets, then 100%, if not, 50%
- x points: if the target is different for all points, yet the attributes are all the same, we should expect the metric to be 1/x
- x points with y, z similar targets: in the case that we have x points with similar attributes, but for which there are y and z similar targets (two groups with the same target), we can at best hope for ... For a dataset with 1 B and 3 C as output values, we expect the metric to be between 0.25 and 0.75 (1/4 and 3/4)
- sum(count(points for target x with attributes X)/count(points with attributes X))/count(points)
- It may make sense to have a separability metric per target output (when those are categorical)
- In the case of numerical targets, minimizing for distance between all the targets (clustering) would turn the problem into the categorical form
- Datasets with defined inputs and target are often not functions (i.e, the same inputs may produce different outputs)

3 See also

4 References

- https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598
- https://github.com/8080labs/ppscore