Data Preprocessing – rescaling, normalization, standardization

Eric Chio “Log0”

12 September 2013

http://www.chioka.in/

Overview

- Definition
- Motivation
- Rescaling
- Normalization
- Standardization

WHAT IS DATA PREPROCESSING

Definitions

- An important step before using the data for machine learning.
- Get rids of missing values, fixing erroneous records, converting data to another format for suitable for machine learning algorithms, etc…
- We will talk about how to convert input data to suit a machine learning algorithm – **rescaling, normalization, and standardization**

WHY PREPROCESS THE DATA?

Motivation

- Some machine learning algorithms make assumptions on the input data.
- If the data does not exhibit these assumptions, the algorithm could behave badly.
- Typical operations: rescaling, normalization, and standardization

RESCALING

Rescaling

- The process of converting a vector by add/subtract and then multiply/divide by a constant.

- E.g. Converting Celsius to Fahrenheit

- Note, this is used interchangeably as standardization in ML literature sometimes…

NORMALIZATION

Normalization

- The process of converting a vector to have unit norm, defined as L^2-norm (Euclidean distance)

$$|\mathbf{x}| = \sqrt{\sum_{k=1}^{n} |x_k|^2},$$

Motivation

- Some algorithms assume this property in the input data, such as the vector space model.

- The vector space model is used in text classification in calculating the distance between two document vectors:

$$\text{sim}(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

When not to do it?

- Normalization still discards some information of the input data, should only be used unless necessary.

- e.g. scale of the data has no significance

STANDARDIZATION

Standardization

- The process could be converting a vector:
  1. to have mean = 0 and standard deviation = 1, which is a Gaussian distribution
  2. to be in the range of [-1, +1], or [0, 1], or even [a, b] where a, b are arbitrary ranges.
- Pick one that is appropriate to the algorithm (which has different assumptions)
- Also termed as scaling/rescaling in practice, so it is very confusing.

Motivation

- As mentioned, algorithms have assumptions on data.

- Concretely, for algorithms that uses the distance could be affected, if the range of the different features differ by too much.

Example

- Suppose there are two data points (1), (2) with two features A, B:

Feature	Min	Max	Point 1	Point 2	Difference
A	0	100000	10000	90000	80000
B	0	10	1	9	8

- The difference between these two points in feature A is 80000 and feature B is 8.

- Looks like A is so much further apart!

Example

- Suppose we put it on a relative scale of [0, 1] instead…

Feature	Min	Max	Point 1	Point 2	Difference
A	0	1	0.1	0.9	**0.8**
B	0	1	0.1	0.9	**0.8**

- We can see the difference is actually the same, relative to a standardized scale!
- Without standardization, the algorithm **will think the difference between these two features are not the same, in reality they are the same**, leading to a suboptimal classifier.

When to do it?

- Use it when the concept of distance is used by an algorithm.

- For instance, SVM uses distance to calculate the largest margin to separate data with a hyperplane. Great numeric values could dominate other smaller numeric values, causing suboptimal behavior.

When not to do it?

- When the range of the feature is very unclear, that is, you do not know the min and max range.

- Also when the concept of distance is not used. For instance, multilayer perceptrons are a linear combination of the input data multiplied with weights, the scale will be accounted and scaled up/down by the weights.

Summary

- Data preprocessing is a crucial step of machine learning
- Some machine learning algorithms have assumptions about the input data
- Normalization and standardization converts input data into a format suitable for machine learning algorithms

References

- http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html
- http://mathworld.wolfram.com/L2-Norm.html
- http://en.wikipedia.org/wiki/Vector_Space_Model