International Mathematical Forum, Vol. 11, 2016, no. 17, 809 - 815
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/imf.2016.6680

Semigroups of Data Normalization Functions

Matthijs J. Warrens

GION, University of Groningen, The Netherlands

Abstract

Variable centering and scaling are functions that are typically used in data normalization. Various properties of centering and scaling functions are presented. It is shown that if we use two centering functions (or scaling functions) successively, the result depends on the order in which the functions are applied: the second function always cancels the centering or scaling of the first function. Furthermore, it is shown that if we use a centering and a scaling function successively, the result does not depend on the order in which the functions are applied. Moreover, certain sets of normalization functions turn out to be semigroups.

Mathematics Subject Classification: 13P25, 20M14, 20M99, 62H99

Keywords: Idempotent semigroup, semilattice, algebraic statistics

1 Introduction

In statistics, data analysis and classification an important step is the normalization of the independent variables or features of the data. The step is used to standardize the range of the independent variables and is usually performed as a data preprocessing step [1,2]. If the range of the raw variables varies widely the solutions obtained with various machine learning and cluster analysis algorithms will be affected. For example, many algorithms consider distances between points. If one variable has a broad range of values, the final distance will be greatly affected by this particular variable. To ensure that each variable contributes approximately proportionally to the final distance, the range of all variables is usually normalized.

We may distinguish two types of normalization functions, namely centering and scaling functions. Centering functions are here defined as transformations that adjust the location of a variable, whereas scaling functions adjust the range of a variable. In this paper we study the algebraic properties of sets of normalization functions. It turns out that the sets can be classified as semigroups. Furthermore, several properties of normalization functions are presented that concern applications of the functions. The results contribute to the field of algebraic statistics.

The paper is organized as follows. Definitions of centering and scaling functions, together with examples from statistics, are presented in the next section. In Section 3 the algebraic structure of the set of all centering functions and the set of all scaling functions is studied. In Section 4 semigroups containing both centering and scaling functions are studied. Section 5 contains a conclusion.

2 Definitions

In statistics a non-zero vector of length n with real numbers is usually called a variable. Definition 2.1 presents a general formulation of a measure of central tendency.

Definition 2.1. *Let $x \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$ with $a, b \neq 0$. A measure of central tendency is a function $\gamma : \mathbb{R}^n \to \mathbb{R}$ such that $\gamma(ax + b) = a\gamma(x) + b$.*

A commonly used measure of central tendency is the arithmetic mean $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. Another example is the median of x, which is the number separating the higher half of x from the lower half. Furthermore, the minimum and maximum of x also satisfy Definition 2.1. We use Definition 2.1 to define a centering function.

Definition 2.2. *Let γ be a measure of central tendency. The centering function associated with γ is $c : \mathbb{R}^n \to \mathbb{R}^n$ with $c(x) = x - \gamma(x)$. Furthermore, let C denote the set of all centering functions.*

Definition 2.3 presents a general formulation of a measure of dispersion.

Definition 2.3. *Let $x \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$. A measure of dispersion is a function $\sigma : \mathbb{R}^n \to \mathbb{R}$ such that $\sigma(ax + b) = a\sigma(x)$ and $\sigma(x) \neq 0$.*

What distinguishes a measure of dispersion (Definition 2.3) from a measure of central tendency (Definition 2.1) is that the former does not change if a real number is added to the variable x. A measure of dispersion that is commonly used in statistics is the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}.$$

Another example is the range$(x) = \max(x) - \min(x)$. We use Definition 2.3 to define a scaling function.

Definition 2.4. *Let σ be a measure of dispersion. The scaling function associated with σ is $s : \mathbb{R}^n \to \mathbb{R}^n$ with $s(x) = x/\sigma(x)$. Furthermore, let S denote the set of all scaling functions.*

In data normalization centering and scaling functions are not always applied separately. If a centering function is followed by a scaling function, or vice versa, this is called a composition. Compositions of a centering and a scaling function that are commonly used in statistics are, feature scaling

$$f(x) = \frac{x - \min(x)}{\max(x) - \min(x)},$$

and standardization

$$z(x) = \frac{x - \bar{x}}{\sigma}.$$

In the remainder of this section we recall several algebraic properties with respect to the operation of composition. Functions from the sets C and S may or may not possess these properties. The composition of $c, d \in C$ will simply be denoted by $c(d(x)) = cd$. The composition of c with itself will be denoted by $c(c(x)) = c^2$. The composition of c, d is commutative if $cd = dc$. The composition of $c, d, e \in C$ is associative if $c(de) = (cd)e$. A function c is said to be idempotent if $c^2 = c$. Finally, a function $c \in C$ is said to be a left zero if $cd = c$ for all $d \in C$.

3 Left zero semigroups

We first consider two properties of the functions in C and S. Lemma 3.1 and 3.2 show, respectively, that all elements of C and S are idempotent. In other words, Lemma 3.1 and 3.2 show that a centering or a scaling function can be applied multiple times to a variable without changing the result beyond the initial application.

Lemma 3.1. *We have $c^2 = c$ for all $c \in C$.*

Proof. We have $c(c(x)) = c(x - \gamma(x)) = x - \gamma(x) - \gamma(x - \gamma(x))$. Since $\gamma(x)$ is a real number we have $\gamma(x - \gamma(x)) = \gamma(x) - \gamma(x) = 0$, and thus $c(c(x)) = x - \gamma(x) = c(x)$. □

Lemma 3.2. *We have $s^2 = s$ for all $s \in S$.*

Proof. We have

$$s(s(x)) = s(x/\sigma(x)) = \frac{x/\sigma(x)}{\sigma(x/\sigma(x))}.$$

Since $\sigma(x)$ is a real number we have $\sigma(x/\sigma(x)) = \sigma(x)/\sigma(x) = 1$, and thus $s(s(x)) = x/\sigma(x) = s(x)$. □

Lemma 3.3 and 3.4 show, respectively, that all elements of C and S are left zeros. In other words, Lemma 3.3 and 3.4 show that if we use two centering functions (or scaling functions) successively, the result depends on the order in which the functions are applied. The function that is applied last cancels the result of the function that was applied first.

Lemma 3.3. *We have* $cd = c$ *for all* $c, d \in C$.

Proof. We have $c(d(x)) = c(x - \delta(x)) = x - \delta(x) - \gamma(x - \delta(x))$. Since $\delta(x)$ is a real number we have $\gamma(x - \delta(x)) = \gamma(x) - \delta(x)$, and thus $c(d(x)) = x - \delta(x) - \gamma(x) + \delta(x) = x - \gamma(x) = c(x)$. □

Lemma 3.4. *We have* $st = s$ *for all* $s, t \in S$.

Proof. We have

$$s(t(x)) = s(x/\tau(x)) = \frac{x/\tau(x)}{\sigma(x/\tau(x))}.$$

Since $\tau(x)$ is a real number we have $\sigma(x/\tau(x)) = \sigma(x)/\tau(x)$, and thus

$$s(t(x)) = \frac{x/\tau(x)}{\sigma(x)/\tau(x)} = \frac{x}{\sigma(x)} = s(x).$$

□

Lemmas 3.1 to 3.4 describe the structure of the sets C and S. It follows from Lemma 3.1 and 3.3 that the set C, together with the operation of composition, is associative and contains idempotent elements that are left zeros. In other words, C is a so-called left zero semigroup. For example, let $c, d, e \in C$. The three-element subset $\{c, d, e\}$ is a left zero semigroup under composition. Its Cayley table is as follows.

	c	d	e
c	c	c	c
d	d	d	d
e	e	e	e

Analogously, it follows from Lemma 3.2 and 3.4 that S is also a left zero semigroup.

4 More idempotent semigroups

In this section we study compositions of a centering and a scaling function. Lemma 4.1 is an important result in this respect. Lemma 4.1 shows that a function from C and a function S commute under composition. In other words Lemma 4.1 shows that if we apply a centering and a scaling function successively, the result does not depend on the order in which the functions were applied.

Lemma 4.1. *We have $cs = sc$ for all $c \in C$ and $s \in S$.*

Proof. We have

$$c(s(x)) = c\left(\frac{x}{\sigma(x)}\right) = \frac{x}{\sigma(x)} - \gamma\left(\frac{x}{\sigma(x)}\right).$$

Since $\sigma(x)$ is a real number we have $\gamma(x/\sigma(x)) = \gamma(x)/\sigma(x)$, and thus

$$c(s(x)) = \frac{x - \gamma(x)}{\sigma(x)}.$$

Furthermore, we have

$$s(c(x)) = s(x - \gamma(x)) = \frac{x - \gamma(x)}{\sigma(x - \gamma(x))}.$$

Since $\gamma(x)$ is real number we have $\sigma(x - \gamma(x)) = \sigma(x)$, and thus

$$s(c(x)) = \frac{x - \gamma(x)}{\sigma(x)}.$$

□

With respect to feature scaling and standardization, Lemma 4.1 shows that it does not matter if we first center the variable and then rescale it, or vice versa, because the result will be the same. Lemma 4.2 shows that a composition of a centering and a scaling function is idempotent. In other words, Lemma 4.2 shows that the composition can be applied multiple times to a variable without changing the result beyond the initial application.

Lemma 4.2. *Let $c \in C$ and $s \in S$. Then $(cs)^2 = cs$.*

Proof. Since $cs = sc$ (Lemma 4.1) we have $(cs)^2 = cscs = ccss = c^2s^2$. Furthermore, since c and s are idempotent (Lemma 3.1 and 3.2) it follows that $(cs)^2 = c^2s^2 = cs$. □

Lemmas 4.1 and 4.2, together with Lemmas 3.1 and 3.2, can now be used to describe the structure of the three-element set that consists of a centering function, a scaling function and their composition. The result is presented in Theorem 4.3.

Theorem 4.3. *Let $c \in C$ and $s \in S$. The set $\langle c, s\rangle = \{c, s, cs\}$ is a semigroup where cs acts as a zero, that is, an absorbing element. The Cayley table is as follows.*

	c	s	cs
c	c	cs	cs
s	cs	s	cs
cs	cs	cs	cs

Lemma 4.4 shows that if we apply two different compositions successively, the result depends on the order in which the compositions are applied: the composition that is applied last cancels the result of the composition that was applied first.

Lemma 4.4. *Let $c, d \in C$ and $s, t \in S$. We have $(cs)(dt) = cs$.*

Proof. Using Lemmas 3.3, 3.4 and 4.1 we have $csdt = scdt = sct = cst = cs$. □

The above lemmas specify how different elements from the same set (C or S) and two elements from different sets (one from C and one from S) behave under composition. By combining functions from C with functions S we may obtain various different semigroups. The structure of such a semigroup can be made precise using the lemmas in this paper. Theorem 4.5 specifies the total number of elements of a set that is generated by k centering functions and m scaling functions.

Theorem 4.5. *Let $c_1, \ldots, c_k \in C$ and $s_1, \ldots, s_m \in S$. The set $\langle c_1, \ldots, c_k, s_1, \ldots, s_m\rangle$ is a semigroup with $k + m + km$ idempotent elements.*

5 Conclusion

In statistics, data analysis and classification data normalization is a common preprocessing step [1,2]. Functions that are typically used are variable centering and scaling. The set of all centering functions and the set of all scaling functions are both left zero semigroups. Furthermore, the set generated by a centering and a scaling function is a semilattice with three elements, that is not a chain.

The results may contribute to the study of data normalization and statistics by means of algebraic methods. For example, it follows that if we use two centering functions (or scaling functions) successively, the second function always cancels the centering or scaling of the first function. Thus, the result always depends on the order in which the functions are applied. On the other hand, it turns out that a centering and a scaling function always commute, which means that the result does not depend on the order in which the functions are applied.

In statistics and data analysis various functions have been used to normalize certain measures of similarity or association [3-7]. Sets of these normalization functions also form semigroups under function composition [8].

References

[1] S. Aksoy and R. Haralick, Feature normalization and likelihood-based similarity measures for image retrieval, *Pattern Recognition Letters,* **22** (2001), 563 - 582. http://dx.doi.org/10.1016/S0167-8655(00)00112-4

[2] C. Hennig, M. Meila, F. Murtagh and R. Rocci, *Handbook of Cluster Analysis,* CRC Press, Boca Raton, 2015.

[3] M.J. Warrens, On similarity coefficients for 2×2 tables and correction for chance, *Psychometrika,* **73** (2008), 487 - 502.
http://dx.doi.org/10.1007/S11336-008-9059-Y

[4] M.J. Warrens, Conditional inequalities between Cohen's kappa and weighted kappas, *Statistical Methodology,* **10** (2013), 14 - 22.
http://dx.doi.org/10.1016/j.stamet.2012.05.004

[5] M.J. Warrens, A comparison of Cohen's kappa and agreement coefficients by Corrado Gini, *International Journal of Research and Reviews in Applied Sciences,* **16** (2013), 345-351.

[6] M.J. Warrens, Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa, *Journal of Classification,* **31** (2014), 179 - 193.
http://dx.doi.org/10.1007/s00357-014-9156-9

[7] M.J. Warrens, Inequalities between similarities for numerical data, *Journal of Classification,* **33** (2016), 141 - 148.
http://dx.doi.org/10.1007/s00357-016-9200-z

[8] M.J. Warrens, On association coefficients, correction for chance, and correction for maximum value, *Journal of Modern Mathematics Frontier,* **2** (2013), 111 - 119. http://dx.doi.org/10.14355/jmmf.2013.0204.01

Received: July 10, 2016; Published: August 10, 2016