Summer Undergraduate Research Fellowships

Jaehoon Cha
Research Assistant

Supervisor : Dr. Sanghyuk Lee
Co-Supervisor : Dr. Kyeong Soo (Joseph) Kim

Xi'an Jiaotong-Liverpool University, Suzhou, China.

July 14, 2017

content

(1) Fuzzy sets

(2) Similarity Measure

Fuzzy sets

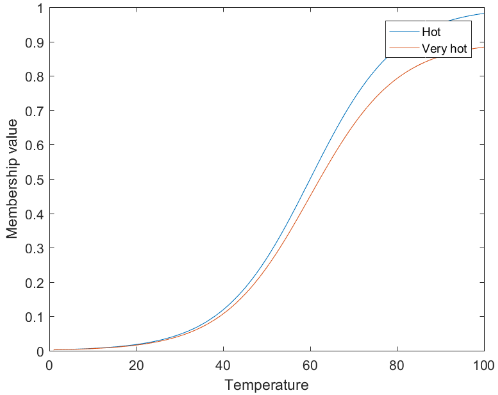
BASIC DEFINITIONS

Fuzzy set

$$100\,^\circ\mathrm{C} > 80\,^\circ\mathrm{C} > 50\,^\circ\mathrm{C}$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$

$$>0.8 \qquad 0.8 \qquad < 0.8$$



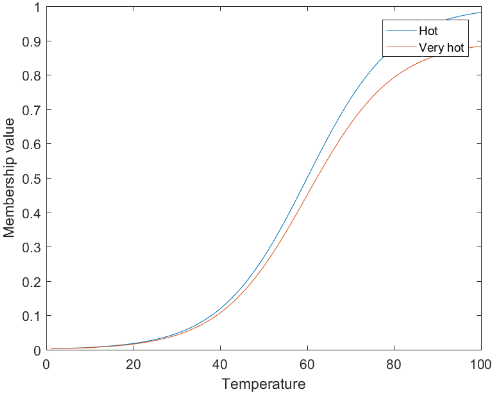
BASIC DEFINITIONS

Fuzzy set

$$100\,^\circ\mathrm{C} > 80\,^\circ\mathrm{C} > 50\,^\circ\mathrm{C}$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$

$$>0.8 \qquad 0.8 \qquad < 0.8$$



Operations on Fuzzy sets.

(1) (Union) $\mu_{A\cup B}(x) = max\{\mu_A(x), \mu_B(x)\}$

(2) (Intersection) $\mu_{A\cap B}(x) = min\{\mu_A(x), \mu_B(x)\}$

(3) (Complement) $\mu_{A^c}(x) = 1 - \mu_A(x)$

- For example, let
  $A = \{(x_1, 0.2), (x_2, 0.7), (x_3, 0.4), (x_4, 1)\}$,
  $B = \{(x_1, 0), (x_2, 0.9), (x_3, 0.5), (x_4, 0.3)\}$
- $A \cup B = \{(x_1, 0.2), (x_2, 0.9), (x_3, 0.5), (x_4, 1)\}$
- $A \cap B = \{(x_1, 0), (x_2, 0.7), (x_3, 0.4), (x_4, 0.3)\}$
- $A^c = \{(x_1, 0.8), (x_2, 0.3), (x_3, 0.6), (x_4, 0)\}$

Operations on Fuzzy sets.

(1) (Union) $\mu_{A\cup B}(x) = max\{\mu_A(x), \mu_B(x)\}$

(2) (Intersection) $\mu_{A\cap B}(x) = min\{\mu_A(x), \mu_B(x)\}$

(3) (Complement) $\mu_{A^c}(x) = 1 - \mu_A(x)$

- For example, let
  $A = \{(x_1, 0.2), (x_2, 0.7), (x_3, 0.4), (x_4, 1)\}$,
  $B = \{(x_1, 0), (x_2, 0.9), (x_3, 0.5), (x_4, 0.3)\}$
- $A \cup B = \{(x_1, 0.2), (x_2, 0.9), (x_3, 0.5), (x_4, 1)\}$
- $A \cap B = \{(x_1, 0), (x_2, 0.7), (x_3, 0.4), (x_4, 0.3)\}$
- $A^c = \{(x_1, 0.8), (x_2, 0.3), (x_3, 0.6), (x_4, 0)\}$

Definition.

- Let $\mathcal{F}$ be a $\sigma-$*algebra*. m is called a fuzzy measure on $(X, \mathcal{F})$ iff

(F1) $m(\phi) = 0$ when $\phi \in \mathcal{F}$

(F2) For $E \in \mathcal{F}, F \in \mathcal{F}$ and $E \subset F$, it implies $m(E) \leq m(F)$ (monotonicity)

(F3) For $\{E_n\} \subset F, E_1 \subset E_2 \subset \cdots$, and $\cup_{n=1}^{\infty} E_n \in \mathcal{F}$ imply $\lim_{n\to\infty} m(E_n) = m(\cup_{n=1}^{\infty} E_n)$ (continuity from below)

(F4) For $\{E_n\} \subset \mathcal{F}, E_1 \supset E_2 \supset \cdots, m(E_1) < \infty$ and $\cap_{n=1}^{\infty} E_n \in \mathcal{F}$ imply $\lim_{n\to\infty} m(E_n) = m(\cap_{n=1}^{\infty} E_n)$ (continuity from above).

No addition rule!

Definition.

- Let $\mathcal{F}$ be a $\sigma-algebra$ and $A \in \mathcal{F}$. The Sugeno integral of f on A with respect to m, which is denoted by $S\int_A f\,dm$, is defined by

$$S\int_A f\,dm \triangleq \sup_{\alpha\in[0,1]} [\alpha \wedge m(A \cap (f)_\alpha)],$$

where $(f)_\alpha = \{x|f \geq \alpha\}$ and $\vee$ and $\wedge$ denote maximum and minimum operators, respectively, i.e., $a \vee b = max(a, b)$ and $a \wedge b = min(a, b)$.

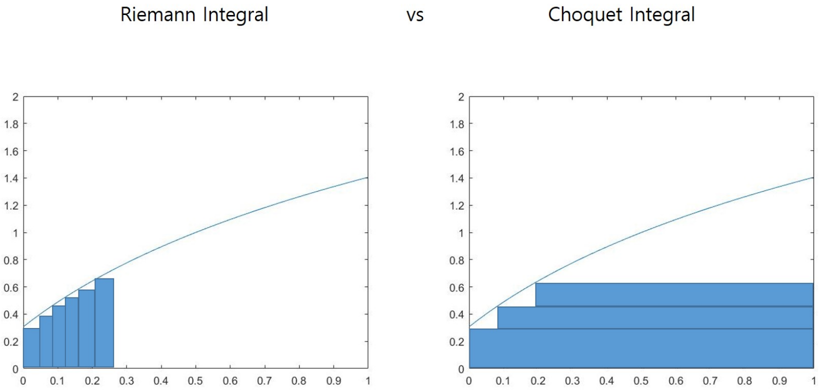
Definition.

- Let $\mathcal{F}$ be a $\sigma - algebra$ and $A \in \mathcal{F}$. The Choquet integral of f on A with respect to m, which is denoted by $C\int_A f dm$, is defined by

$$C\int_A f dm \triangleq \sum_{i=1}^{n}(f(x_i) - f(x_{i-1}))m(A_i),$$

where $\cdot_i$ states that the indices have been permuted so that $0 \leq f(x_1) \leq \cdots \leq (f)(x_n) \leq 1, f(x_0) = 0$ and $A_i \triangleq \{x_i, \cdots, x_n\}$.

Riemann Integral vs Choquet Integral



Similarity measure

Definition.

- A real function $s : \mathcal{F}^2 \to [0, \infty]$ is called a similarity measure, if s has following properties:

(S1) $s(A, B) = s(B, A), \quad \forall A, B \in \mathcal{F}$

(S2) $s(D, D^c) = 0, \quad \forall D \in \mathcal{P}(X)$

(S3) $s(C, C) = \max_{\forall A, B \in \mathcal{F}} s(A, B), \quad \forall C \in \mathcal{F}$

(S4) $A, B, C \in \mathcal{F}$, if $A \subset B \subset C$, then $s(A, B) \geqslant s(A, C)$ and $s(B, C) \geqslant s(A, C)$

Theorem.

- Let $A, B \in \mathcal{F}$ and m be a fuzzy measure on $X = \{x_1, x_2, \cdots, x_n\}$. Then

(I1) $S_S^d(A, B) = 1 - S\int f dm^d$,

(I2) $S_C^d(A, B) = 1 - C\int f dm^d$,

where $f(x) = |\mu_A(x) - \mu_B(x)|$, $m^d(E) = \frac{1}{n}\sum_{x\in E} x^{\frac{1}{d}}$ and $d \in \{1, 2, \cdots\}$, are similarity measure between A and B.

Theorem.

- Similarity measure based on distance measure. The nearer two sets are, the more similar they are.

(L1) $\sum_{i=1}^{n} |x_i - y_i|$

(L2) $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

- Let $A = (1, 0, 0, 0)$, $B = (1, 1, 1, 1)$, and $C = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, then
- $D_{L1}(A, C) = 2$
- $D_{L1}(B, C) = 2$

Theorem.

- Similarity measure based on distance measure. The nearer two sets are, the more similar they are.

(L1) $\sum_{i=1}^{n} |x_i - y_i|$

(L2) $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

- Let $A = (1, 0, 0, 0), B = (1, 1, 1, 1)$, and $C = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, then
- $D_{L1}(A, C) = 2$
- $D_{L1}(B, C) = 2$

Theorem.

- The first one is the correlation-based similarity measure. In this case, we calculate similarity between two items i and j, which is denoted by $w_{i,j}$.The correlation between item i and j will be
$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{i \in I}(r_{u,j} - \bar{r}_j)^2}},$$
where $r_{u,i}$ is the rating of user u on item i, $\bar{r}_i$ is the average rating of the ith item by those users.

Theorem.

- The second one is vector cosine-based similarity. Let i and j be vectors of items which is purchased by users u and v. Then the vector cosine similarity between two users will be
$$w_{i,j} = cos(\overrightarrow{i}, \overrightarrow{j}) = \frac{\overrightarrow{i} \cdot \overrightarrow{j}}{\| \overrightarrow{i} \| \| \overrightarrow{j} \|},$$
where " $\cdot$ " denotes the dot-product of the two vectors.

Example

	i1	i2	i3	i4	i5	i6
A	5	5		4		
B		2	3		4	
C	4			4		2
D		3	3		3	

- $w_{A,C} = cos(\overrightarrow{a}, \overrightarrow{c}) = \frac{5\times4+4\times4}{\sqrt{5^2+5^2+4^2}\times\sqrt{4^2+4^2+2^2}} = 0.739$
- $w_{A,B} = cos(\overrightarrow{a}, \overrightarrow{b}) = \frac{5\times2}{\sqrt{5^2+5^2+4^2}\times\sqrt{2^2+3^2+4^2}} = 0.229$

Example

	i1	i2	i3	i4	i5	i6
A	1/3	1/3		-2/3		
B		-1	0		1	
C	2/3			2/3		-4/3
D	0	0	0		0	

- $w_{A,C} = cos(\overrightarrow{a}, \overrightarrow{c}) = \frac{\frac{1}{3} \times \frac{2}{3} - \frac{2}{3} \times \frac{2}{3}}{\sqrt{\frac{1}{3}^2 + \frac{1}{3}^2 + \frac{2}{3}^2} \times \sqrt{\frac{2}{3}^2 + \frac{1}{3}^2 + \frac{4}{3}^2}} = -0.179$
- $w_{A,B} = cos(\overrightarrow{a}, \overrightarrow{b}) = \frac{\frac{1}{3} \times -1}{\sqrt{\frac{1}{3}^2 + \frac{1}{3}^2 + \frac{2}{3}^2} \times \sqrt{1^2 + 0^2 + 1^2}} = -0.289$
- $w_{B,D} = cos(\overrightarrow{b}, \overrightarrow{d}) = 0.$

- Sanghyuk Lee, Jaehoon Cha, Nipone Theera-Umpon, and Kyeong Soo Kim, "Analysis on similarity measure for non-overlapped data," symmetry, vol. 9, no. 5, pp. 1-11, May 2017.
- Jaehoon Cha, Sanghyuck Lee, Kyeong Soo Kim, and Witold Pedrycz, "On the design of similarity measures based on fuzzy integral," Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computiong and Intelligent Systems, June, Japan, 2017
- John S. Breese, David Heckerman, and Carl Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1998, pp. 43-52.