CLASIFICACIÓN DE LAS MEDIDAS DE DISIMILARIDAD d ENTRE CONGLOMERADOS DE DATOS SEGÚN LA ESTRUCTURA MÉTRICA DEL ESPACIO MUESTRAL DE APLICACIÓN

ISADORE NABI

<table>
<tr><th>Espacio Muestral de d</th><th>Estructura Métrica del Espacio Muestral de d</th><th>Medida de Disimilaridad d</th><th>Expresiones Funcionales de d</th></tr>
<tr><td rowspan="3">Espacios Métricos Completos (Espacios de Cauchy)</td><td rowspan="3">Según (Chiang, 2006, pág. 65),
a) $d(x,y) = 0$, para $x = y$[1]
b) $d(x,y) = d(y,x) > 0$, para $x \neq y$
c) $d(x,y) \leq d(x,z) + d(z,y)$, para $z \neq x, y$[2]
d) $d(x,y) \geq 0$, positivida</td><td rowspan="2">Distancia Euclidiana</td><td>En $\mathbb{R}^2$, según (Weisstein, Euclidian Metric, 2022): $d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$, con $x = x_1, \ldots, x_n$ y $y = y_1, \ldots, y_n$.</td></tr>
<tr><td>En $\mathbb{R}^n$, según (Halabiski, 2022): $d_{ij} = \sum_{v=1}^{n}\left(X_{vi} - X_{vj}\right)^2$, donde $x_{vi} - x_{vj}$ denota el valor de la $v - ésima$ variable de la celda i menos el valor de la $v - ésima$ variable en la celda j.</td></tr>
<tr><td>Coeficiente de Correlación de Pearson</td><td>Según (Gujarati & Porter, 2010, pág. 77): $d(x,y) = \frac{\sum_{i=1}^{n}(X_i - Y_i)}{\sqrt{(\sum_{i=1}^{n} X_i^2)(\sum_{i=1}^{n} Y_i^2)}}$</td></tr>
</table>

[1] También conocida como *identidad de los indiscernibles*.
[2] También conocida como desigualdad triangular, la cual no es más que una generalización del Teorema de Pitágoras para cualesquiera tipos de rectángulos.

	d, deducida de las tres propiedades anteriores.	*Coeficiente de Correlación de Spearman*	Según (Barcelona Field Studies Center, 2022), cuando no hay rangos empatados (del mismo tamaño): $d(x,y) = 1 - \left(\frac{6\sum_{i=1}^{n}(X_i - Y_i)^2}{n^3 - n}\right)$ Según (Laerd Statistics, 2022), cuando hay rangos empatados: $d(x,y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\sum_{i=1}^{n}(Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \bar{X})^2)(\sum_{i=1}^{n}(Y_i - \bar{Y})}}$ donde $\bar{X}$ y $\bar{Y}$ son las medias muestrales de X e Y, respectivamente.
		Coeficiente de Correlación de Kendall	Según (Statology, 2020): $d(x,y) = \frac{C - D}{C + D}$, donde C es el número de pares concordantes y D es el número de pares discordantes[3].

[3] Como se señala en (StatisticsHowTo, 2016), supóngase que dos entrevistadores calificaron a un grupo de doce solicitantes de empleo, datos recopilados en la siguiente tabla:

Espacios Semi-Métricos	Según (McAuley, 1956, pág. 315): a) $d(x,y) = 0, sii\ x = y$ b) La topología del espacio topológico S es invariante con	*Cuasi-Métrica*	a) Según (Stojmirovic, 2005, pág. 23), las cuasi métricas en $\mathbb{R}$ que generan las denominadas topología superior (d^R) e inferior (d^L) son: $d^L(x,y) = u^L = máx\{x-y, 0\}$ $d^R(x,y) = u^R = máx\{y-x, 0\}$ En la expresión anterior, tanto d^L como d^R son mapas de la forma $\mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. Además,

Candidate	Interviewer 1	Interviewer 2
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11

Teniéndose en cuenta que, en la primera columna, las opciones del entrevistador 1 se ordenaron de menor a mayor, se puede hacer una comparación entre las opciones para el entrevistador 1 y 2. Con pares concordantes o discordantes, básicamente está respondiendo a la pregunta: ¿los jueces/evaluadores clasificaron los pares en el mismo orden? No necesariamente está buscando exactamente el mismo rango, sino más bien si un solicitante de empleo fue clasificado constantemente más alto por ambos entrevistadores. Tres posibles escenarios son posibles para estos pares ordenados: 1. pares empatados: ambos entrevistadores están de acuerdo. Por ejemplo, el candidato A se marcó como primera opción para ambos entrevistadores, por lo que están empatados. 2. Pares concordantes: ambos entrevistadores clasifican a ambos solicitantes en el mismo orden, es decir, ambos se mueven en la misma dirección. Si bien no tienen el mismo rango (es decir, ambos 1° o 2°), cada par se ordena igual de alto o igual de bajo. El entrevistador 1 clasificó a F como 6 y G como 7, mientras que el entrevistador 2 clasificó a F como 5 y G como 8. F y G son concordantes porque F siempre se clasificó más alto que G. 3. Pares discordantes: los candidatos E y F son discordantes porque los entrevistadores clasificaron en direcciones opuestas (uno dijo que E tenía una clasificación más alta que F, mientras que el otro dijo que F tenía una clasificación superior a 6).

	respecto a la función distancia d[4]. c) No se cumple que $d(x,y) = d(y,x)$, aunque las distancias siempre positivas cuando $x \neq y$.		d^L es el conjugado de d^R y viceversa. b) Según (Stojmirovic, 2005, pág. 22), si X es un conjunto de datos y existe una medida d que describe un mapa de la forma $d: X \times X \to \mathbb{R}$, entonces la denominada *métrica discreta* tiene la forma funcional: $d(x,y) = \begin{cases} 0, & si\ x = y \\ 1, & si\ x \neq y \end{cases}$ c) Según (Stojmirovic, 2005, pág. 24), otra cuasi métrica en $\mathbb{R}_+$ es: $d(x,y) = \begin{cases} mín(1, y - x), & si\ x \leq y \\ 1, & de\ otra\ forma \end{cases}$
Espacios Pseudo-Métricos	Posee casi la misma estructura definida para los espacios de		Según (Planethmath, 2013), los siguientes son ejemplos de medidas cuasi-métricas:

[4] Es decir, si existe un punto p que es punto límite (una observación x será punto límite de un conjunto si cumple que para todo número arbitrariamente mayor que cero $\epsilon > 0$ existe un miembro del conjunto y diferente de x tal que $|y - x| < \epsilon$ (Weisstein, Limit Point, 2022); son conocidos también como puntos clúster o puntos de acumulación) de un subconjunto $M \in S$, entonces p es un punto límite distancia de M (*i.e.*, un punto límite que puede ser considerado para calcular la distancia a M) y viceversa. Lo anterior significa que la función distancia d es un invariante topológico (una función, en cuanto invariante topológico, es aquella función que para aplicarse depende únicamente de la topología del espacio, por lo que es aplicable a cualquier espacio homeomórfico, es decir, a cualquier espacio con la misma topología, con independencia de las consideraciones métricas) en el espacio muestral.

	Cauchy, con la única diferencia que existe la posibilidad de $d(x,y)=0$ aun cuando $x \neq y$.	*Pseudo-Métrica*	a) Sea $X=\mathbb{R}^2$ y considérese una función $d: X \times X \to \mathbb{R}_+$ dada por $d[(x_1,x_2),(y_1,y_2)]=\lvert x_1-y_1\rvert$, entonces $d(x,x)=\lvert x_1-x_1\rvert=0$, $d(x,y)=\lvert x_1-y_1\rvert=\lvert y_1-x_1\rvert=d(y,z)$ y la desigualdad triangular se deriva naturalmente de la desigualdad triangular en $\mathbb{R}^1$, por lo que (X,d), donde d es la medida de disimilaridad, satisface las condiciones de un espacio pseudo-métrico. En este espacio puede obtenerse una distancia nula entre dos observaciones diferentes, por ejemplo, $d[(2,3),(2,5)]=\lvert 2-2\rvert=0$ b) Sea X un conjunto, sea $x_0 \in X$ y sea $F(X)$ un conjunto de funciones tal que $X \to R$. Entonces, $d(f,g)=\lvert f(x_0)-g(x_0)\rvert$ es una pseudo-métrica de $F(X)$. c) Si X es un espacio vectorial y p es una

			semi-norma[5] sobre el espacio X, entonces $d(x,y) = p(x-y)$ es una pseudo-métrica sobre X. d) La *pseudo-métrica trivial* $d(x,y) = 0$ para todo $x, y \in X$ es una pseudo-métrica.
Variedades Pseudo-Riemannianas[6]	Una variedad $n-dimensional$ es un espacio topológico con la propiedad de que cada observación dentro de dicho espacio tiene un vecindario (un entorno formado partiendo de un centro y una distancia fija) que	*Métrica de Minkowski*	Según (Kolmogórov & Fomin, 1978, pág. 56): $\left(\sum_{k=1}^{n} \lvert x_k + y_k \rvert^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{n} \lvert x_k \rvert\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{n} \lvert y_k \rvert\right)^{\frac{1}{p}}$, donde p expresa el orden de la potencia a la cual las funciones que conforman el espacio son integrables. Un espacio L_p es aquel espacio de funciones f (donde cada coordenada u observación puede ser representada mediante una función) conformado por aquellas f que, elevadas a una

[5] Como señala (Weisstein, Seminorm, 2022), es una función en un espacio vectorial V denotada como $\|v\|$, tal que cumple las siguientes condiciones para todos los vectores v y w que pertenecen a V, así como para todo escalar c:

1. $\|v\| \geq 0$,
2. $\|cv\| = |c|\|v\|$,
3. $\|v + w\| \leq \|v\| + \|w\|$.

Nótese que es posible que $\|v\| = 0$ para vectores v no nulos.

[6] Es una variedad con un tensor métrico no-degenerado en todas partes, lo cual se explicará a continuación.

	es homeomórfico (topológicamente equivalente) a algún conjunto abierto de un espacio euclidiano $n-dimensional$[7]. Una variedad pseudo-riemmaniana posee casi la misma estructura definida para las variedades, con la salvedad de que además es posible obtener un tensor métrico[8] no-		potencia p, son integrables. Por ejemplo, los espacios L_2 son los espacios de funciones que son cuadrado-integrables, es decir, que la integral de f^2 existe (*i.e.*, tiene un valor finito).

[7] Que se diferencia de un espacio métrico (o de un espacio métrico completo) en que el espacio métrico completo es la conjunción del espacio euclidiano y de la métrica euclidiana. Véase (Kolmogórov & Fomin, 1978, pág. 52).

[8] Como se señala en (Wikipedia, 2021), un tensor métrico es un mapa bilineal (función lineal en cada uno de sus argumentos que combina elementos de dos espacios vectoriales -que pueden definirse como matrices de datos- y producen un elemento de un tercer espacio vectorial -que puede definirse también como una matriz de datos-) no degenerado (es decir, el mapa bilineal y su mapa conjugado son isomórficos, *i.e.*, topológicamente equivalentes), suave (sus $n-ésimas$ derivadas evaluados en un punto existen) y simétrico (la distancia de x a y es igual a la distancia de y a x) que asigna un número real a un par de vectores tangentes (relacionados entre sí a través de una estructura diferencial -de derivadas- que se mostrará al definir la estructura de la medida de disimilaridad de las variedades pseudo-riemannianas) en cada espacio tangente (como señala (Weisstein, Tangent Space, 2022), es una copia de un espacio $\mathbb{R}^n$ tangencial a una variedad compacta M, es decir, a una variedad que es compacta en tanto espacio topológico; un espacio

	degenerado En Todas Partes[9]		

REFERENCIAS

Barcelona Field Studies Center. (21 de Abril de 2022). *Spearman's Rank Correlation Coefficient*. Obtenido de Spearman's Rank: https://geographyfieldwork.com/SpearmansRank.htm

Chiang, A. (2006). *Métodos Fundamentales de Economía Matemática.* México, D.F.: McGraw-Hill. Obtenido de https://elvisjgblog.files.wordpress.com/2018/02/mc3a9todos-fundamentales-de-economc3ada-matemc3a1tica-4ta-edicic3b3n-alpha-c-chiang-freelibros-org.pdf

Gujarati, D. N., & Porter, D. C. (2010). *Econometría.* México D.F.: McGraw-Hill Educación. Obtenido de https://fvela.files.wordpress.com/2012/10/econometria-damodar-n-gujarati-5ta-ed.pdf

Halabiski, A. (23 de Febrero de 2022). *Euclidean Distance In 'n'-Dimensional Space*. Obtenido de Stanford University: https://hlab.stanford.edu/brian/euclidean_distance_in.html

Kaplan, W. (1985). *CÁLCULO AVANZADO.* MÉXICO, D.F.: COMPAÑÍA EDITORIAL CONTINENTAL, S.A. DE C.V., MÉXICO.

Kolmogórov, A. N., & Fomin, S. V. (1978). *Elementos de la Teoría de Funciones y del Análisis Funcional* (Tercera ed.). (q. e.-m. Traducido del ruso por Carlos Vega, Trad.) Moscú: MIR.

---

topológico es compacto si todo cubrimiento abierto -colección de conjuntos abiertos de un espacio topológico cuya unión contiene algún conjunto dado- posee un subcubrimiento finito -un cubrimiento que es subconjunto de otro cubrimiento-) de la variedad generada por el conjunto de datos. Como señala (Kaplan, 1985, pág. 305), un tensor de orden dos (bidimensional) posee la forma $g_{ij} = \frac{\partial \xi^r}{\partial x^i}\frac{\partial \xi^r}{\partial x^j}$, donde ξ expresa el sistema de coordenadas estándar (del cual se generan todos los nuevos sistemas de coordenadas), x^i representa la $i-$ésima variable, x^j la $j-$ésima variable de la matriz de datos (puesto que una matriz bidimensional como las utilizadas en el álgebra lineal son equivalentes a tensores de orden 2) y el símbolo ∂ denota la derivada parcial del numerador respecto del denominador.

[9] Es decir, en cualquier localidad (subregión) del espacio en que se tomen las medidas de disimilaridad.

Laerd Statistics. (1 de Febrero de 2022). *Spearman's Rank-Order Correlation*. Obtenido de Statistical Guides: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

McAuley, L. F. (Diciembre de 1956). A Relation Between Perfect Separability, Completeness, and Normality in Semi-Metric Spaces. *Pacific Journal of Mathematics, VI*(2), 315-326.

Planethmath. (22 de Marzo de 2013). *example of pseudometric space*. Obtenido de planetmath.org: https://planetmath.org/exampleofpseudometricspace

StatisticsHowTo. (20 de Agosto de 2016). *Concordant Pairs and Discordant Pairs*. Obtenido de Statistics Definitions: https://www.statisticshowto.com/concordant-pairs-discordant-pairs/

Statology. (26 de Febrero de 2020). *Kendall's Tau: Definition + Example*. Obtenido de Statistics. Simplified: https://www.statology.org/kendalls-tau/

Stojmirovic, A. (2005). *Quasi-metrics, Similarities and Searches: aspects of geometry of protein datasets.* Wellington}: Victoria University of Wellington. Obtenido de https://arxiv.org/pdf/0810.5407.pdf

Weisstein, E. W. (19 de Abril de 2022). *Euclidian Metric*. Obtenido de MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/EuclideanMetric.html

Weisstein, E. W. (19 de Abril de 2022). *Limit Point*. Obtenido de MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/LimitPoint.html

Weisstein, E. W. (19 de Abril de 2022). *Seminorm*. Obtenido de MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/Seminorm.html

Weisstein, E. W. (19 de Abril de 2022). *Tangent Space*. Obtenido de MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/TangentSpace.html

Wikipedia. (10 de Noviembre de 2021). *Pseudo-Riemannian manifold*. Obtenido de Smooth Manifolds: https://en.wikipedia.org/wiki/Pseudo-Riemannian_manifold