Bagadi, R. (2017). Picking A Least Biased Random Sample Of Size n From A Data Set of N Points {Version 3}. ISSN 1751-3030. *PHILICA.COM Article number 971.*
http://www.philica.com/display_article.php?article_id=971

Picking A Least Biased Random Sample Of Size n From A Data Set of N Points {Version 3}. ISSN 1751-3030

Ramesh Chandra Bagadi *(Physics, Engineering Mechanics, Civil & Environmental Engineering, University of Wisconsin)*

Published in matho.philica.com

Abstract

In this research investigation, a Statistical Algorithm is detailed that enables us to pick a Least Biased Random Sample of Size n, from a Data Set of N Points with n<N

Article body

Bagadi, R. (2017). Picking A Least Biased Random Sample Of Size n From A Data Set of N Points {Version 3}. ISSN 1751-3030. *PHILICA.COM Article number 971*.
http://www.philica.com/display_article.php?article_id=971

Picking A Least Biased Random Sample Of Size n From A Data Set of N Points {Version 3}

ISSN 1751 - 3030

Author:
Ramesh Chandra Bagadi
Data Scientist
International School Of Engineering (INSOFE)
Postal Address: Plot No 63/A, 1st Floor,
Road No 13, Film Nagar, Jubilee Hills,
Hyderabad – 500033, Telengana State, India.
Email: ramesh.bagadi@insofe.edu.in

Abstract

In this research investigation, a Statistical Algorithm is detailed that enables us to pick a Least Biased Random Sample of Size n, from a Data Set of N Points.

Theory

Given a Data Set of N points, if we were to pick a Least Biased Random Sample of Size n, i.e., n Data Points, we can use the following stated Algorithm.

Algorithm

Firstly, we consider all possible Partitions of Size n of the given Data Set of N points. These will be ${}^{N}C_{n} = \frac{N!}{n!(N-n)!}$ in number. Let these be represented by P_i for $i = 1$ to ${}^{N}C_{n}$.

Now, for such Partitions P_i, we find the Average (Arithmetic Mean) $\overline{X}_{P_i}$

We now find, using K-Means Clustering Algorithm, n Clusters using these ${}^{N}C_{n}$ data points called $\overline{X}_{P_i}$ and find their Centroids and let us Label these $\overline{A}_{P_i}$.

We now pick any particular Partition, say P_k, wherein we establish $n!$ Number of One-One Functions between the n Elements of P_k and the aforementioned n Elements of Set $\overline{A}_{P_i}$ and Pick One that Particular Function such that the

a. Differences $\left|\overline{A}_{P_i}(l) - P_k(m)\right|$ are Minimum Possible for $l = 1$ to ${}^N C_n$ and $m = 1$ to ${}^N C_n$.

b. Sum Of the Differences $\sum_{i=1}^{{}^N C_n} \left|\overline{A}_{P_i P_i}(l) - P_k(m)\right|$ are Minimum Possible for $j = 1$ to ${}^N C_n$ and $m = 1$ to ${}^N C_n$.

c. Sum Of the Squares Of the Differences $\sum_{i=1}^{{}^N C_n} \left|\overline{A}_{P_i}(l) - P_k(m)\right|^2$ are Minimum Possible for $j = 1$ to ${}^N C_n$ and $m = 1$ to ${}^N C_n$.

We now repeat this procedure for all the rest of P_i other than P_k and whichever Partition has this Least value, we consider that particular Partition has the *Least Possible Sampling Bias*.

Finding the aforementioned $n!$ Number of Functions

Considering the Set $\overline{A}_{P_i}$, the elements of the Set P_k can be arranged among themselves in $n!$ Number of ways. Now the One-One position wise respective correspondence between the Elements of the Set $\overline{A}_{P_i}$ and the Elements of each of the aforementioned arrangements of the Set P_k gives us the $n!$ Number of Functions.

We can also repeat the same Procedure using the Expected Value in place of the Mean $\overline{A}_{P_i}$.

References

http://www.philica.com/advancedsearch.php?author=12897

http://vixra.org/author/ramesh_chandra_bagadi

Dictionary of Cancer Terms ? selection bias. Retrieved on September 23, 2009.

Medical Dictionary - 'Sampling Bias' Retrieved on September 23, 2009

TheFreeDictionary ? biased sample. Retrieved on 2009-09-23. Site in turn cites: Mosby's Medical Dictionary, 8th edition.

Dictionary of Cancer Terms ? Selection Bias. Retrieved on September 23, 2009.

*Ards, Sheila; Chung, Chanjin; Myers, Samuel L. (1998). "The effects of sample selection bias on racial differences in child abuse reporting". Child Abuse & Neglect. **22** (2): 103–115. doi:10.1016/S0145-2134(97)00131-2. PMID 9504213.*

*Cortes, Corinna; Mohri, Mehryar; Riley, Michael; Rostamizadeh, Afshin (2008). "Sample Selection Bias Correction Theory" (PDF). Algorithmic Learning Theory. **5254**: 38–53. doi:10.1007/978-3-540-87987-9_8.*

*Cortes, Corinna; Mohri, Mehryar (2014). "Domain adaptation and sample bias correction theory and algorithm for regression" (PDF). Theoretical Computer Science. **519**: 103–126. doi:10.1016/j.tcs.2013.09.027.*

Fadem, Barbara (2009). Behavioral Science. Lippincott Williams & Wilkins. p. 262. ISBN 978-0-7817-8257-9.

*Feinstein AR; Horwitz RI (November 1978). "A critique of the statistical evidence associating estrogens with endometrial cancer". Cancer Res. **38** (11 Pt 2): 4001–5. PMID 698947.*

*Tamim H; Monfared AA; LeLorier J (March 2007). "Application of lag-time into exposure definitions to control for protopathic bias". Pharmacoepidemiol Drug Saf. **16** (3): 250–8. doi:10.1002/pds.1360. PMID 17245804.*

Matthew R. Weir (2005). Hypertension (Key Diseases) (Acp Key Diseases Series). Philadelphia, Pa: American College of Physicians. p. 159. ISBN 1-930513-58-5.

*Kruskal, William H. (1960). "Some Remarks on Wild Observations". Technometrics. **2** (1): 1–3. doi:10.1080/00401706.1960.10489875.*

*Jüni, P.; Egger, Matthias (2005). "Empirical evidence of attrition bias in clinical trials". International Journal of Epidemiology. **34** (1): 87–88. doi:10.1093/ije/dyh406.*

Bostrom, Nick (2002). Anthropic Bias: Observation Selection Effects in Science and Philosophy. New York: Routledge. ISBN 0-415-93858-9.

*Ćirković, M. M.; Sandberg, A.; Bostrom, N. (2010). "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks". Risk Analysis. **30** (10): 1495. doi:10.1111/j.1539-6924.2010.01460.x.*

*Tegmark, M.; Bostrom, N. (2005). "Astrophysics: Is a doomsday catastrophe likely?". Nature. **438** (7069): 754. doi:10.1038/438754a. PMID 16341005.*

*Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error". Econometrica. **47**: 153. doi:10.2307/1912352. JSTOR 1912352.*