

Data Preprocessing

Project ENABLE

May 28, 2019



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



Why Do We Need Data Preprocessing?

- Heejun Kim vs. H. Kim

Google Scholar

H. Kim

Articles

About 6,200,000 results (0.06 sec)

My profile

My library

Any time

Since 2019

Since 2018

Since 2015

Custom range...

Sort by relevance

Sort by date

☒ include patents

☒ include citations

☐ Create alert

User profiles for H. Kim

Sung Hyun Kim - Verified email at pucpr.br - Cited by 254064

Jong-Hwan Kim - Verified email at rit.kaist.ac.kr - Cited by 11185

Yong-Hyun Kim - Verified email at kaist.ac.kr - Cited by 6147

Observation of a ${}^6_{\Lambda\Lambda}He$ Double Hypernucleus

..., M Kawasaki, CO Kim, JY Kim, SJ Kim, SH Kim... - Physical review ..., 2001 - APS

A double-hyperfragment event has been found in a hybrid-emulsion experiment. It is identified uniquely as the sequential decay of $\Lambda\Lambda^6He$ emitted from a Ξ^- hyperon nuclear capture at rest. The mass of $\Lambda\Lambda^6He$ and the $\Lambda\bar{\Lambda}$ interaction energy $\Delta B\Lambda\Lambda$ have been ...

Cited by 608 Related articles All 10 versions Web of Science: 376

H^∞ state feedback control for generalized continuous/discrete time-delay system

JH Kim, HB Park - Automatica, 1999 - Elsevier

In this paper, we consider the problem of designing H^∞ state feedback controller for the generalized time-delay systems with delayed states and control inputs in continuous and discrete time cases, respectively. The generalized time-delay system problems are solved ...

Cited by 245 Related articles All 7 versions Web of Science: 117

Quantum-inspired evolutionary algorithms with a new termination criterion, H/sub/spl epsi//gate, and two-phase scheme

KH Han, JH Kim - IEEE transactions on evolutionary ..., 2004 - ieeexplore.ieee.org

From recent research on combinatorial optimization of the knapsack problem, quantum-inspired evolutionary algorithm (QEA) was proved to be better than conventional genetic algorithms. To improve the performance of the QEA, this paper proposes research issues on ...

Cited by 583 Related articles All 14 versions Web of Science: 278

[PDF] aps.org

Find article @ UNC

Find article @ UNC

[PDF] ieee.org

Find article @ UNC



Data Quality

- Data analysis can be highly vulnerable to noisy, missing, and inconsistent data.
- Today's real-world data likely origin from multiple, heterogenous sources.
- Many reasons for low-quality data (e.g., human error at data entry) and know them is important.
- Data quality should be properly addressed.



Simple Descriptive Analysis for Data Cleaning

- Sort all values in ascending or descending order
 - All values are identical: No information. Ignore this variable.
 - Some values occur abnormally with very high frequency: “Alabama” in 20% cases for the “state” part of addresses.
 - The mode age of respondents is 69 in a student satisfaction survey: Default values of data of birth are 01 (day), 01 (month), and 1950 (year).



Methods for Processing Missing Values

- Ignore the instance: Simplest, but not very effective. In general, this method is not recommended.
- Use a measure of central tendency of the variable (e.g., the mean or median) to fill in the missing value: Use mean for normal data distribution and use median for skewed data distribution.
 - Good if the data distribution for a variable is centered or skewed.
 - Meh if the data distribution is more evenly distributed.
- Use the most probable value (imputation)
 - Use regression, inference-based tools based on a Bayesian formalism, or decision tree induction.
 - This approach uses the most information from the available data.



Methods for Processing Noisy Values

- Problem with an extreme outlier. For instance, the department at UNC which has the highest average annual income of alumni is “Geography.”
- Binning: sort values and put them into a number of “buckets,” or bins. This approach basically utilize the values of the neighborhood and perform local smoothing.
- Remove outliers
 - Remove instances whose values are more than two standard deviations away from the mean for a given attribute.
 - Create clusters to put similar values into groups, or “clusters.” Remove instances whose values fall outside of the created clusters.



An Example of Binning

- Sorted data for medical expenditure (in dollars):
104429, 105405, 111544, 121635, 132497, 135662, 153429, 160648, 174984
- Partition into bins
 - Bin 1: 104429, 105405, 111544
 - Bin 2: 121635, 132497, 135662
 - Bin 3: 153429, 160648, 174984
- Smoothing by bin means
 - Bin 1: 107126
 - Bin 2: 129931.3
 - Bin 3: 163020.3



Data Integration

- Data integration is processes to combine data from disparate sources into a unified view of data.
- Potential problems
 - Redundancies and inconsistencies in the resulting dataset.
 - Entity Identification Problem: Different names (e.g., patient_id vs. id) used for the same concept in different databases.
 - Data value conflict (e.g., Heejun vs. H.).
 - Instance duplication

Any Questions?

Bivariate Analysis

Next Class



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL