

Data Preprocessing

# Terminology

|  |  |
| --- | --- |
| Data Preprocessing | * وهى العمليات اللى بتم قبل ما ابدأ فى التحليل لان البيانات بتكون غير مهيئة للتحليل بسبب المشاكل اللى فى البيانات * Datasets are highly susceptible to noisy, missing, and inconsistent data. * Low-quality data will lead to low-quality mining results. * Data quality factors includes:   1. accuracy,   2. completeness,   3. consistency,   4. timeliness,   5. believability   6. interpretability. |
| Reasons of Low Data Quality | * Inaccurate data   + Having incorrect attribute values (e.g., by choosing the default value “January 1” displayed for birthday)   + عدى يا عمى التاريخ بسرعة اختاره اى حاجة ,, المفروض يبقى فيه اسلوب تحقق زى ويب فاليديشن * **Incomplete data**    + **Missing data (may not always be available or of interest)** * **Inconsistent data**    + **different assessments of the quality depend on the intended use of the data** * Timeliness   + (e.g. month-end data are not updated in a timely fashion has a negative impact on the data quality. ) * **Believability**    + **reflects how much the data are trusted by users** * Interpretability   + reflects how easy the data are **understood** (e.g. sales codes) |
| Preprocessing Tasks That Improve Data Quality |  |
| Data Cleaning | Image result for dirty datais about:   * filling in **missing values**   + Missing Values: {Nan, Null, Na, “”} * Smooth out **noise**   + Noisy: Contain **Errors** => Salary = -1000 * Identifying or removing **outliers** * Removing **inconsistencies**   + (e.g. rating was “1, 2, 3”, now rating “A, B, C”) * **Intentional** manipulating (e.g., by choosing the default value “January 1” displayed for birthday) |
| Missing Values Methods | * **Ignore** the Tuple   + **Effective when Class label is missing, and the Required task is Classification**   + Not Effective unless the tuple has varied of missing attributes   + **Ignoring make no use of other attributes of the tuple that can be useful** * **Filling** in missing values manually   + Time Consuming   + May not be feasible in a large data set * **Using** a **global constant** to fill the missing value   + **Simple**   + **Not foolproof in datamining tasks => مش مضمون لانه منكم يأثر على عمليات الماينينج** * **Use** a measure of **central tendency** for the attribute to fill in the missing value   + **Symmetric** data distribution => Use **Mean**   + **Skewed** data distribution => Use **Median** * **Use** the attribute **mean or median** for all samples belonging to the same class as the given tuple   + زى اللى فوقها بالظبط بس على مستوى الداتا اللى من نفس الكلاس * **Use** the **most probable value** to fill in the missing value   + Can be determined using **Regression, Decision Tree Induction, Inference Based tools**   + Uses the most information of present data to predict the missing value   A missing value may not imply an error in the data!   * Forms should allow respondents to specify values such as “not applicable.” Software routines may also be used to uncover other null values (e.g., “don’t know,” “?” or “none”) |
| Noisy Data | * Noisy Data in Data Mining | Soft Computing and Intelligent ...Noise is a **random error** or **variance** in a measured variable. * “**smooth**” out the data to **remove** the noise. * Smooth Data Techniques:   + Binning   + Regression   + Outlier Analysis |
| Binning | * Numerical & Scientific Computing with Python: Binning Data with ...The original data values are divided into “**buckets**” known as “**bins**” and then they are replaced by a **general value** calculated for that bin. * In the context of **image processing**, binning is the procedure of combining a **cluster of pixels into a single pixel**. * Image result for binning2 Steps for Binning   + Partitioning   + Smoothing * “Partition” sorted data by 2 methods:   + **Equal depth (frequency) bins**: each bin has same number of values   + Image result for binning**Equal width bins**: interval range of values per bin is equal * “Smooth” each bin by:   + **bin means**: each bin value is replaced by the bin **mean**   + **bin medians**: each bin value is replaced by the bin **median**   + **bin boundaries**: each bin value is replaced by the **closest boundary value** (min & max in a bin are bin boundaries) |
| Regression | * **Linear** regression involves finding the “best” line to **fit two attributes** (or variables) so that **one attribute can be used to predict the other.** * Replace **noisy or missing values** by **predicted** values. |
| Outlier Analysis | * May be Detected by **Clustering** * The data outside the cluster {circle} may be analyzed as Outlier * باذن الله مشروحة في جزء ال * Clustering |
| Data Integration | * Merging of data from multiple **data stores**. * Be Careful when integration because of Redundency and inconsistencies * فيه مشاكل هتقابلك و انت بتعمل تجميع للبيانات،،، تعالى نبدأ نتعرف عليهم! * **Problems:** * **Entity Identification Problem**   + How can equivalent real-world entities from multiple data sources be matched up?   + For example, how can the data analyst or the computer be sure that customer id in one database and cust\_number in another refer to the same attribute?   + **metadata** can be used to help avoid errors in schema integration.   + You Can See the Metadata about Iris dataset * **Redundency**   + An attribute may be redundant if it can be “***derived***” from another attribute or set of attributes. => Age, Date of Birth, annual revenue, for instance   + So some Redundancies Can be detected using **Correlation analysis**   + **Correlation analysis**, given two attributes, such analysis can measure how strongly one attribute implies the other.   + There are 2 test:     - Chi-Squre for Nominal Data     - Covariance for Numeric   + تعالى نفتح الموضوع ده في سكشن جديد 😀 |
| Correlation Test for Nominal Data | * a correlation relationship between two attributes, A and B, can be discovered by a (chi-square) test * Where is the observed frequency (i.e., actual count) of the joint event ( , ) and is the expected frequency of ( , ) which can be computed as * tests the hypothesis that A and B are independent, that is, there is no correlation between them.[Null vs Research Hypothesis] * وده اللى كان موضحه دكتور صيام في مادة سنة ثالثة مادة بحث علمى * دلوقتى تعالى نحل المثال ده لاحظ ان الألفا بيكون مرجعى يعنى مش بيتحسب بس احنا بنستخدمه على حسب المجال * ولاحظ ان ال كاى سكوير بتثبت ان لاوجود ترابط و معنى كده ان هي بتثبت ال * Null Hypothesis      * Correlation doesn’t imply causality * يعنى في المثال حتى لو فيه ترابط بين الشغب و النوع ده مش بيثبت ان النوع هو اللى بيسبب الشغب |
| Correlation Coefficient for Numeric Data | * A screenshot of a cell phone    Description automatically generatedThe correlation coefficient between two attributes, A and B, is   + If is *greater* than 0, then *A* and *B* are *positively* correlated, The higher the value, the stronger the correlation   + If = 0, then *A* and *B* are *independent* * Note that Correlation does not imply causality! That is, if A and B are correlated, this does not necessarily imply that A causes B or that B causes A * Covarience:    + Measures how two things change together .   + *Covariance is +ve* → *A* & *B* change together, and if *A* > then *B* >   + *Covariance is -ve* → one is above its mean and one is below   + If *A* and *B* are independent → *Covariance* = 0   + لو كانوا موجب يبقى الأول لو زاد التانى هيزداد و العكس علاقة طردية   + Example |
| Data Reduction | * More is not always better. * Obtain a **reduced representation** of the data set that is much smaller in volume, yet closely maintains the **integrity** of the original data. * Data Reduction Startegies   + Dimensionality Reduction   + Numerosity Reduction   + Data Compression |
| Dimensionality Reduction | * Beginners Guide To Learn Dimensionality Reduction TechniquesReduce the number of attributes under consideration * Methods include:   + wavelet transforms   + principal components analysis (PCA)   + Attribute subset selection |
| Numerosity Reduction Techniques | * Data are replaced or estimated by alternative.   + parametric methods, a model is used to estimate the data (PCA)   + Nonparametric methods histograms, clustering, sampling, and data cube aggregation |
| Data Compression | * Reducing the amount of capacity required to store data.   + **lossless** : No loss of information (e.g. Text )   + **Lossy**: the size of the file is reduced by eliminating data in the file (e.g. Image) |
| Attribute Subset Selection | * How can we find a ‘**good’** subset of the original attributes? * Rmove the redundent or irrelevent attributes * For 𝑛 attributes, there are 2𝑛 possible subsets!!! * Solution: **Heuristic (Greedy) methods**   + while searching for attribute subsets, they always make what looks to be the best choice at the time. * **Heuristic : Stepwise forward selection {empty} => {Reduced set}**   + The **best** of the attributes is determined and added to the reduced set.   + “**best**” is determined by some predetermined criteria * **Heuristic : Stepwise backward selection {Fill} => {Reduced Set}**   + start with the full set of attributes.   + At each step, remove the worst attribute remaining in the set * **Heuristic : Compination of Stepwise Forward and Backward** * **Heuristic: Decision tree induction**   + **و دى هتندرس في شابتر ال classification**   Data Mining: Concepts and Techniques (3rd ed.) — Chapter 3 — - ppt ... |
| Regression | * 𝑦 = 𝑤𝑥 + 𝑏 * y (response variable), can be modeled as a linear function of x (predictor variable) * W (slope) and b (intercept) could be optimized to get the best fitting |
| Histograms (binning) | * The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted:   + 1, 1, 5, 5, 5,5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30. |
| Sampling | * Obtain (smaller) subsets of the dataset called data sample. * Simple random sample without replacement (SRSWOR) of size s: all tuples are equally likely to be sampled. * Simple random sample with replacement (SRSWR) of size s: similar to SRSWOR, but a tuple is drawn recorded then placed back so it may be drawn again      * **Cluster sample** : non overlapping * **Stratified sample** : if the tuples are divided into strata (overlapping) |
| Data Transformation | * Data are transformed into forms **appropriate** for mining * **Transformation** Strategies   + **Smoothing**   + **Attribute Selection**   + **Aggregation** For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.   + **Normalization**: scaling values   + **Discretization**: (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.)   + **Concept Hierarchy**: street can be generalized to higher-level concepts, like city or country |
| Transformation by Normalization | * To help avoid dependence on the choice of measurement units * Normalizing the data attempts to give all attributes an equal weight   + **Min-max normalization**   + **Z-score normalization** |
| Concept Hierarchy Generation | * • It Recursively reduce data by replacing low level concepts (e.g. age values) by higher level concepts (e.g. age groups: youth, adult, or senior). * explicitly specified by domain experts * formed for both **numeric** and **nominal** data |