Tour Version

Data Mining

Tour 2

Data Preprocessing

Data Preprocessing

Terminology

<table>
<tr><td>Data Preprocessing</td><td>

وهى العمليات اللى يتم 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:

accuracy,
completeness,
consistency,
timeliness,
believability
interpretability.
</td></tr>
<tr><td>Reasons of Low Data Quality</td><td>

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)


<table>
<tr><th>Tid</th><th>Refund</th><th>Marital Status</th><th>Taxable Income</th><th>Cheat</th></tr>
<tr><td>1</td><td>Yes</td><td>Single</td><td>125K</td><td>No</td></tr>
<tr><td>2</td><td>No</td><td>Married</td><td>100K</td><td>No</td></tr>
<tr><td>3</td><td>No</td><td>Single</td><td>70K</td><td>No</td></tr>
<tr><td>4</td><td>Yes</td><td>Married</td><td>120K</td><td>No</td></tr>
<tr><td>5</td><td>No</td><td>Divorced</td><td>10000K</td><td>Yes</td></tr>
<tr><td>6</td><td>No</td><td>NULL</td><td>60K</td><td>No</td></tr>
<tr><td>7</td><td>Yes</td><td>Divorced</td><td>220K</td><td>NULL</td></tr>
<tr><td>8</td><td>No</td><td>Single</td><td>85K</td><td>Yes</td></tr>
<tr><td>9</td><td>No</td><td>Married</td><td>90K</td><td>No</td></tr>
<tr><td>9</td><td>No</td><td>Single</td><td>90K</td><td>No</td></tr>
</table>
</td></tr>
<tr><td>Preprocessing Tasks That Improve Data Quality</td><td>
Data cleaning: missing values, noisy data, outliers

Data integration: data from multiple sources

Data reduction: reduced representation of the data set

Data transformation: data scaled to fall within a smaller range like 0.0 to 1.0
</td></tr>
<tr><td>Data Cleaning</td><td>
is 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”)

</td></tr>
</table>

<table>
<tr>
<td></td>
<td>

Intentional manipulating (e.g., by choosing the default value “January 1” displayed for birthday)

</td>
</tr>
<tr>
<td>Missing Values Methods</td>
<td>

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

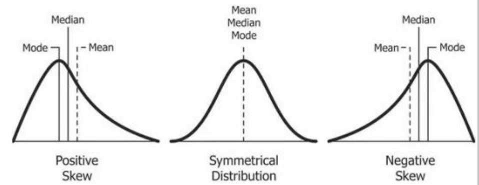
Time Consuming
May not be feasible in a large data set

Using a global constant to fill the missing value

Simple
Not foolproof in datamining tasks => مش مضمون لانه منكم يأثر على عمليات المايننج

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

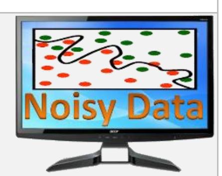


Use the attribute mean or median for all samples belonging to the same class as the given tuple

زى اللى فوقها بالظبط بس على مستوى الداتا اللى من نفس الكلاس

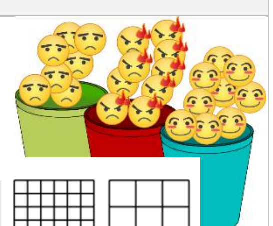
Use the most probable value to fill in the missing value

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”)
</td>
</tr>
<tr>
<td>Noisy Data</td>
<td>


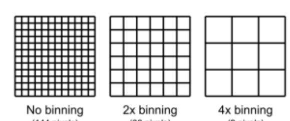
Noise is a random error or variance in a measured variable.
“smooth” out the data to remove the noise.
Smooth Data Techniques:

Binning
Regression
Outlier Analysis

</td>
</tr>
<tr>
<td>Binning</td>
<td>


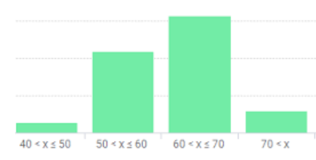
The original data values are divided into “buckets” known as “bins” and then they are replaced by a general value calculated for that bin.
In the context of image processing, binning is the procedure of combining a cluster of pixels into a single pixel.
2 Steps for Binning

Partitioning
Smoothing



“Partition” sorted data by 2 methods:

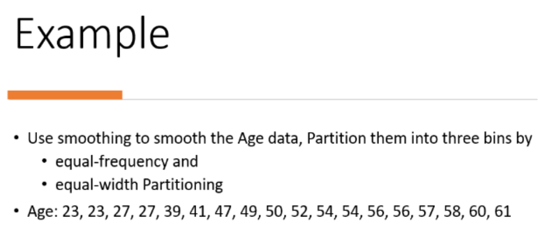
Equal depth (frequency) bins: each bin has same number of values
Equal width bins: interval range of values per bin is equal

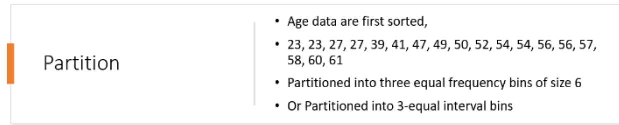


“Smooth” each bin by:

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)

</td>
</tr>
</table>





Partition into (equal-frequency bins:

Bin 1: 23, 23, 27, 27, 39, 41

Bin 2: 47, 49, 50, 52, 54, 54

Bin 3: 56, 56, 57, 58, 60, 61

Partition into (equal-width) bins:

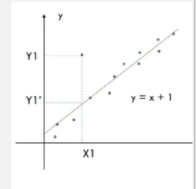
Bin 1: 23, 23, 27, 27

Bin 2: 39, 41, 47, 49, 50

Bin 3: 52, 54, 54, 56, 56, 57, 58, 60, 61

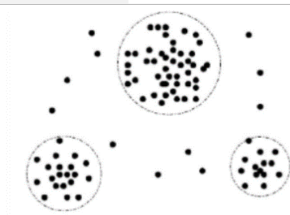
Regression

- <u>**Linear**</u> 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.



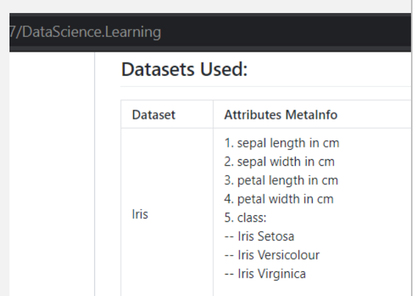
Outlier Analysis

- 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



/DataScience.Learning

Datasets Used:

Dataset	Attributes MetaInfo
Iris	1. sepal length in cm 2. sepal width in cm 3. petal length in cm 4. petal width in cm 5. class: -- Iris Setosa -- Iris Versicolour -- Iris Virginica

- **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 X^2 (chi-square) test

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

- Where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i , B_j) and e_{ij} is the expected frequency of (A_i , B_j) which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

- 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

Example : Determine if is a relationship Between Gender and getting in Trouble

	Get in Trouble	not —	Total
Boys	46	71	117
girls	37	83	120
Total	83	154	237

$$e_{11} = \frac{117 * 83}{237} = 40.97$$

$$e_{12} = \frac{117 * 154}{237} = 76.02$$

$$e_{21} = \frac{120 * 83}{237} = 42.03$$

$$e_{22} = \frac{120 * 154}{237} = 77.97$$

$$X^2 = \frac{(46 - 40.97)^2}{40.97} + \frac{(71 - 76.02)^2}{76.02} + \frac{(37 - 42.03)^2}{42.03} + \frac{(83 - 77.97)^2}{77.97} = 1.87$$

degree of freedom

معناه احنا كده نجي على جدول Critical Statistics ⇐

$DF = (\#\,rows - 1)\ (\#\,col - 1) = (2-1) * (2-1) = 1$

نشوف الجدول اللي تحت وحدد لما يكون DF = 1 والـ alpha (reference) بـ 0.05

هيساوي 3.841 ⇐ ولو قمت قارنت مابينها و Critical

واللي هناخد $3.841 > 1.87$

لو كده بيبقى Null Hyp ⇐ وبالتالي مافيش علاقة مابينهم

- 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

- The correlation coefficient between two attributes, A and B, is

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

  - If $r_{A,B}$ is *greater* than 0, then *A* and *B* are *positively* correlated, The higher the value, the stronger the correlation
  - If $r_{A,B}$ = 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:
  - $Cov(A,B) = E(A.B) - \bar{A}\bar{B}$
  - Measures how two things change together .
  - *Covariance is +ve* → *A* & *B* change together, and if *A* > $\bar{A}$ then *B* > $\bar{B}$
  - *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

٦ أيام

Find mean? Then E(X,Y)
Finally Cov(X,Y)

Temp	No of Cust[illegible]
98	15
87	12
90	10
85	10
95	16
75	7

$$\bar{X} = \frac{98+87+90+85+95+75}{6}$$

$$= 88.33$$

$$\bar{Y} = \frac{15+12+10+10+16+7}{6} = 11,67$$

$$E(X,Y) = \frac{(98\cdot15)+(87\cdot12)+(90\cdot10)+(85\cdot10)+(95\cdot16)+(75\cdot7)}{6}$$

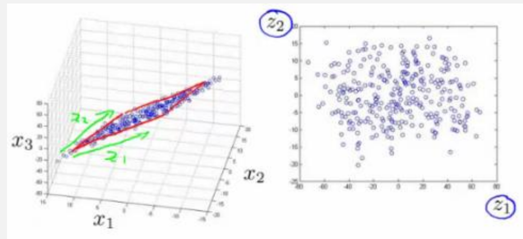
$$= 1051.5$$

$$Cor(X,Y) = E(X\cdot Y) - \bar{X}\bar{Y} = 1051.5 - (88.33\cdot 11,67)$$

$$= 20.94$$

الرقم موجب وبالتالى كل مايزداد درجة الحرارة كل مايجى زبائن أكتر

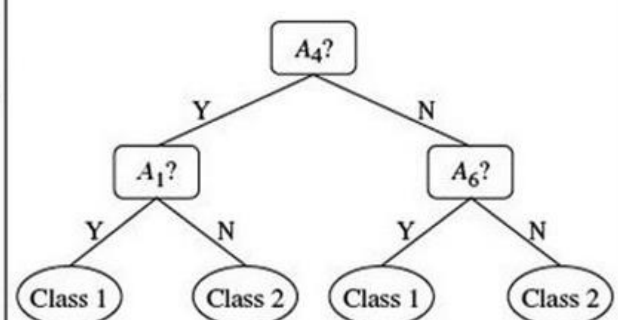
Data Reduction	○ More is not always better.
	○ Obtain a **reduced representation** of the data set that is much smaller in volume, yet closely maintains the **integrity** of the original data.
	○ Data Reduction Startegies
	    ○ Dimensionality Reduction
	    ○ Numerosity Reduction
	    ○ Data Compression

Dimensionality Reduction	▪ Reduce 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)

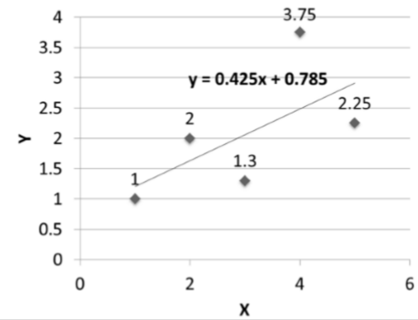
Attribute Subset Selection	▪ How can we find a '**good'** subset of the original attributes?
	▪ Rmove the redundent or irrelevent attributes
	▪ For n attributes, there are 2^n possible subsets!!!
	▪ Solution: **Heuristic (Greedy) methods**
	    ○ while searching for attribute subsets, they always make 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**

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

<table>
<tr><td>Regression</td><td>

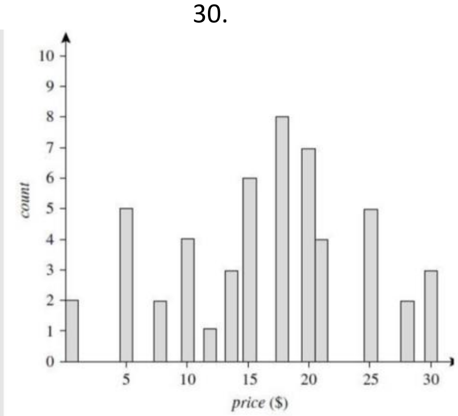
- $y = wx + b$
- y (response variable), can be modeled as a linear function of x (predictor variable)
- W (slope) and b (intercept) could be optimized to get the best fitting

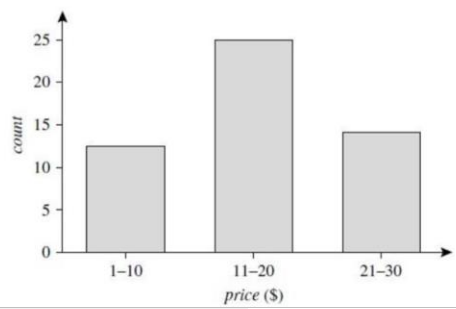
<table>
<tr><th>X</th><th>Y</th></tr>
<tr><td>1.00</td><td>1.00</td></tr>
<tr><td>2.00</td><td>2.00</td></tr>
<tr><td>3.00</td><td>1.30</td></tr>
<tr><td>4.00</td><td>3.75</td></tr>
<tr><td>5.00</td><td>2.25</td></tr>
</table>



</td></tr>
<tr><td>Histograms (binning)</td><td>

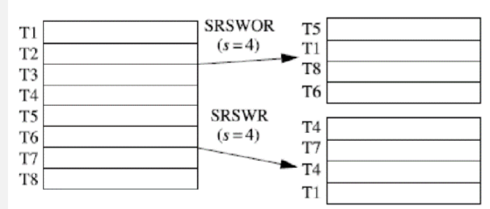
- The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted:
    - 1, 1, 5, 5, 5,5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18,18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30.



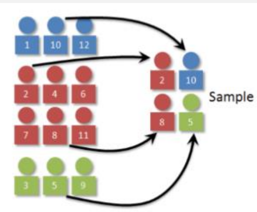


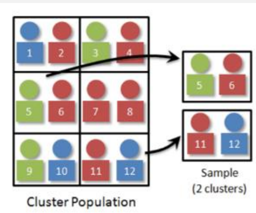
</td></tr>
<tr><td>Sampling</td><td>

- 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)





</td></tr>
<tr><td>Data Transformation</td><td>

- 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

</td></tr>
</table>

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** • $v = \frac{v-min}{max-min}(new_{max} - new_{min}) + new_{min}$ • Suppose that the minimum and maximum values for the attribute age are 13 and 70, respectively. We would like to map age to the range [0.0, 1.0]. • By min-max normalization, a value of 35 for age is transformed to $map(35) = \frac{35-13}{70-13}(1-0) + 0 = 0.39$ ○ **Z-score normalization** • Normalized based on the mean and standard deviation . • $v = \frac{v-mean}{standard\ deviation}$ • Useful when the actual minimum and maximum of attribute A are unknown, or • when there are outliers that dominate the min-max 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