Introduction to Data Mining

What is data mining?

Probability – reminder !

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

Based on Data Mining – Concept and Techniques,
Jiawei Han, Micheline Kamber & Jian Pei

Motivation by examples

The Challenge

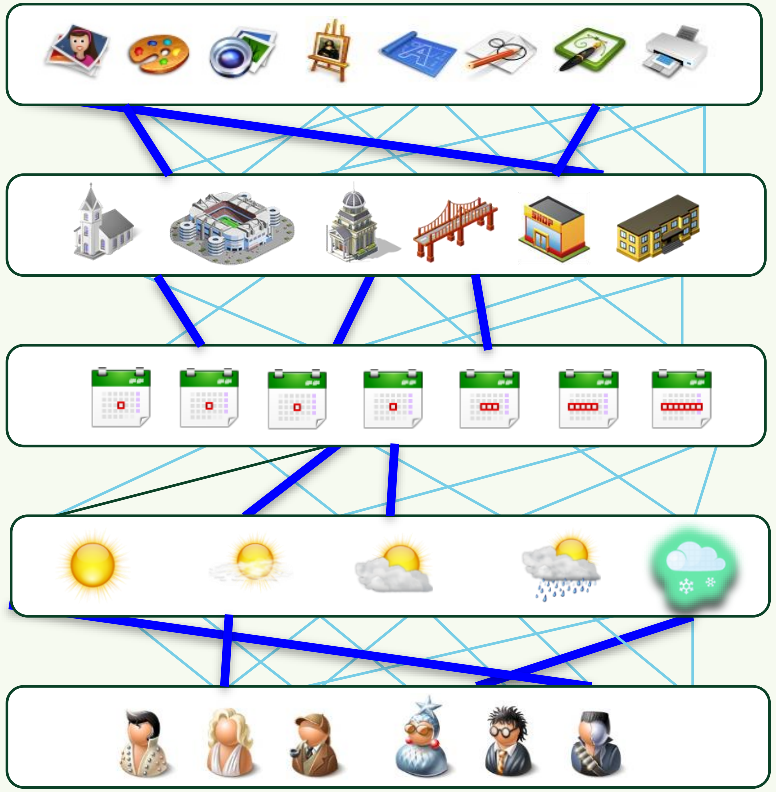
(e)Products (8)

Locations (6)

Week day (7)

Weather (5)

Customers (6)



In this toy example: 4,294,967,296 potential patterns!!!

Sample pattern: "On rainy Tuesdays the customer will buy service A if he is located close to downtown"

The challenge: identifies the significant patterns and maps them into business opportunities.

Motivation by examples

Executions | Scheduling | Management

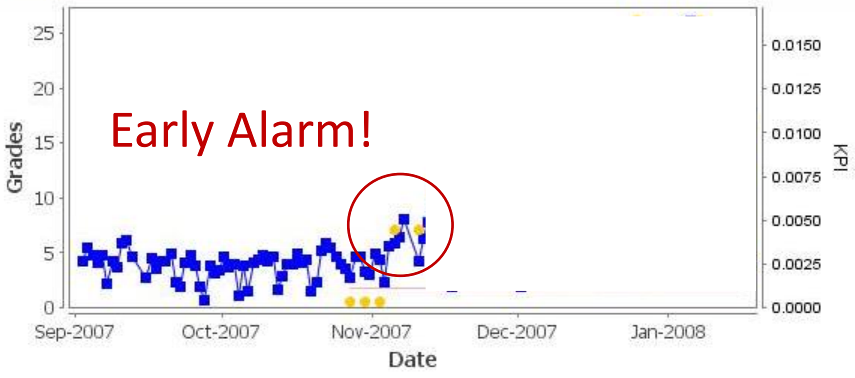
Columns | KPI And Exec | Hierarchy | Sort By | Filter By | Show Charts | Export to Excel

Executions List

		Source Name	KPI Name	UCL	# Grades Above	% Grades Above	% Change in KPI	Anomaly Start	Anomaly End
		Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv South 12982	UL_Bad_NEW %	8.53	25	92.59	123.67	31/10/2007 10:14	02/11/2007 10:14
		City:Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
		City:Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv South 12983	HO Cause Qual UL %	3.58	24	92.59	68.34	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12983	DCR %	1.73	24	88.89	102.14	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12983	UL_Bad_NEW %	7.21	24	88.89	82.92	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12982	DCR %	7.55	24	88.89	116.17	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv	DCR %	6.80	23	85.19	44.02	06/11/2007 10:14	10/11/2007 10:14

« ‹ 1 2 3 4 › »

Source Chart



Info

Source Name: Tel Aviv South 12983
KPI Name: DCR %
UCL: 1.73
 # Grades Above: 24
% Grades Above: 88.89
% Change in KPI: 102.14
Anomaly Start: 03/11/2007 12:14
Anomaly End: 05/11/2007 12:14

Findings | XML Model | Tree Model

Why data mining ?

The **Explosive Growth** of Data: from terabytes to petabytes

Data collection and **data availability**

Major sources of **abundant** data

We are drowning in **data**, but starving for **knowledge**!

What is data mining?

Extraction of **interesting** knowledge from **huge** amount of data

non-trivial

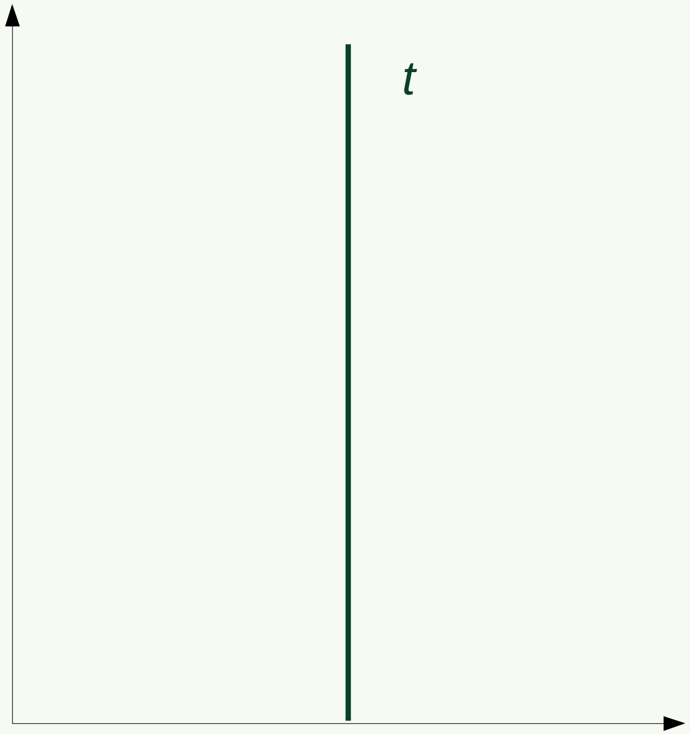
implicit

previously unknown

potentially useful

Example

if income < *t* then person defaulted on loan



X= bad situation

O= good situation

The graph represent historical data.

How to decide when to give a loan?

<u>Definition of terms</u>

- **Data** - is the set of facts (*F*)

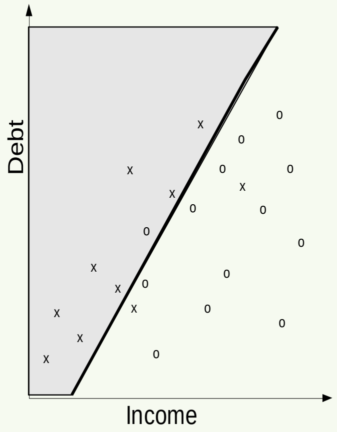
  In the example- collection of 23 cases, each containing three fields: debt income status
- **Pattern** - is an expression *E* describing facts in the subset E_F of *F*.
- **Process** – multi-step process to discover validity, useful and non-trivial but understandable results

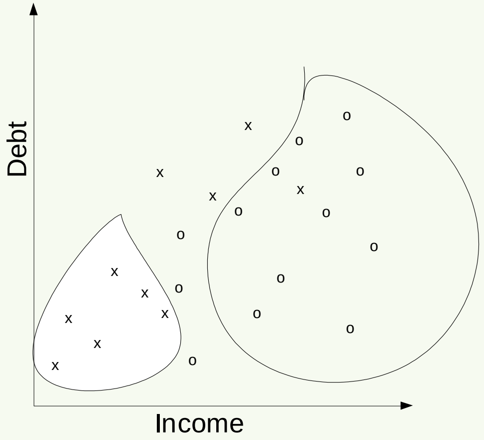
Example – definitions of terms

- **Validity** - The discovered patterns should be valid with some certainty, *C* (E, F).
- **Potentially Useful** - The discovered patterns should do some useful actions that can be measured by some utility function U(E, F).
- **Understandable** - The goal of KDD is to create patterns understandable in order to understand better the data. If this can be measured, it will be measured by s=S(E, F).
- **Interestingness** - The overall measure of pattern value combining all the individual measures: i = I(E, F, C, U, S)
- **Knowledge** - A pattern $E \in L$ is called a knowledge if for some threshold I(E, F, C, U, S)>*v*.

Data mining - tasks

- **Classification** - is the learning of a function that maps a data into one of several predefined classes.
- **Prediction** - is the learning of a function that maps data item into A prediction variable.
- **Clustering** - Identify a set of items with common characteristics
- **And more : reinforcement learning; associating; sequencing**

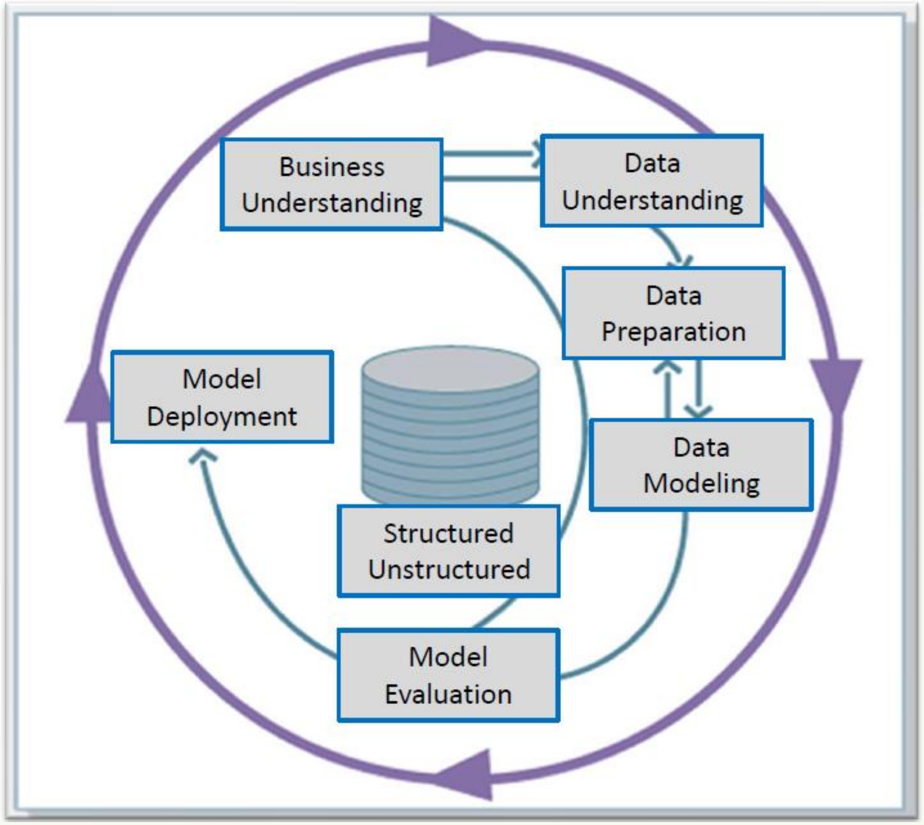




CRISP-DM

The Cross Industry Process for Data Mining –

(www.crisp dm.org) (CRISP DM; Shearer, 2000)



Probability -reminder

- “Introduction to Probability Models”, Sheldon Ross
- “Introduction to Probability and Statistics for Engineers and Scientists”, Sheldon Ross
- “Introduction To Probability”, Dimitri P. Bertsekas, John N. Tsitsiklis

Probability – basics

- Random experiment E, outcome $\omega \in \Omega$, events F, sample space (Ω, F)
- Probability measure $P: F \rightarrow R$
- Axioms of probability, basic laws of probability
- Discrete sample space, discrete probability measure
- Continuous sample space, continuous probability measure
- Conditional probability, multiplicative rule, theorem of total probability, Bayes theorem
- Independence, pair-wise, mutual, conditional independence

Random variables

- $X: \Omega \to R$
- Example:
  - Experiment: Tossing of two coins
  - Random variable: sum of two outcomes
  - $\{X = 2\} \equiv \{\omega: sum\ of\ scores = 2\ \} = \left\{\{1,1\}\right\}$

Some discrete distributions

- Bernoulli: $X \sim Ber(p)$, $\boxed{P_X(k) = p^k(1-p)^{1-k}; k = 0,1}$
- Binomial: $X \sim Bin(n,p)$, $\boxed{P_X(k) = P\{X = k\} = \binom{n}{k} p^k q^{n-k} \quad ; \quad k = 0,1,\ldots,n}$
- Poisson: $X \sim Poisson(\lambda)$, $\boxed{P_X(k) = P\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!} \quad ; \quad k = 0,1,2,\ldots}$
- Geometric: $X \sim Geo(p)$, $\boxed{P_X(k) = P\{X = k\} = p \cdot (1\text{-}p)^{k-1} \quad ; \quad k = 1,2,\ldots}$

Some density functions

Uniform: $x \sim U(a,b) \equiv f(x) = \frac{1}{b-a}$

Exponential: $x \sim Exp(\lambda) \equiv f(x) = \lambda e^{-\lambda x}$

Standard Normal: $x \sim N(0,1) \equiv f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Gaussian: $x \sim N(\mu,\sigma) \equiv f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$

Much more, Gamma, Beta et. al.

Moments

The *r*-th moment

$$m_r = \sum_i x_i^r \cdot p(x_i)$$

Mean (the first raw moment)

$$(\mu =)\ m_1 = E(x) = \sum_i x_i p(x_i)\ \left(= \int xp(x)dx\right)$$

$a.\ E(aX + bY) = aE(X) + bE(y)$

Variance (the second central moment)

$$Var[X] = E[(X - E[X])^2] = E(x^2) - \mu^2$$

a. $Var(a) = 0$

b. $Var(ax) = a^2 Var(x)$

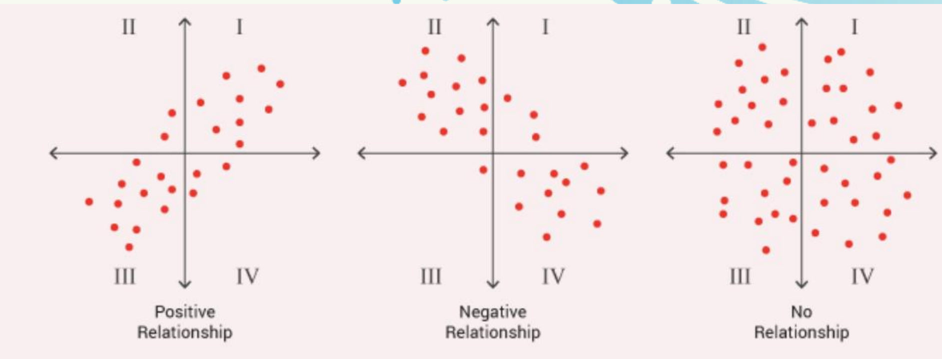
c. $Var(a + x) = Var(x)$

Covariance

Covariance :

$$Cov(x,y) = E\{(x-\mu_x)(y-\mu_y)\} = E(x\cdot y) - \mu_x\mu_y$$

a. if x and y are independent then : $E(x\cdot y) = E(x)E(y) = \mu_x\mu_y$ → Cov(x,y) = 0



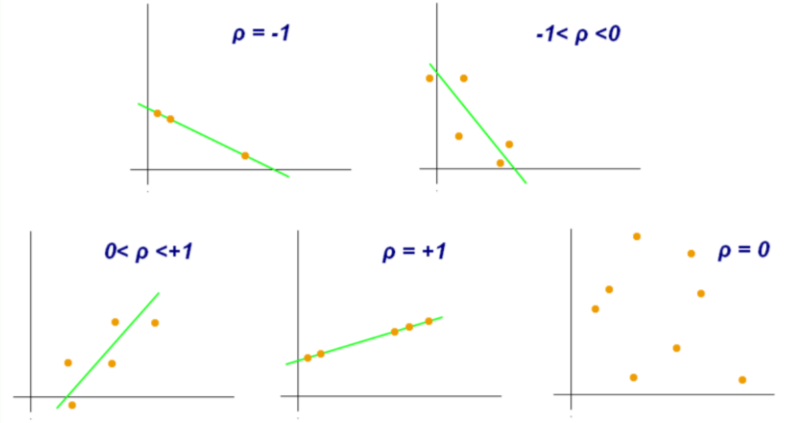
Correlation co-efficient

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x\sigma_y}$$

a. if x and y are independent then : $\rho(x,y) = 0$

b. $|\rho(x,y)| \leq 1$

c. $|\rho(x,y)| = 1 \Leftrightarrow y = ax + b$



Central Limit Theorem

- N i.i.d. random variables X_i with mean μ, variance σ^2
- $S_N = \sum_i X_i$
- $Z_N = \frac{S_N - N\mu}{\sigma\sqrt{N}}$
- As N increases the distribution of Z_N approaches the standard normal distribution $f(Z_N) \sim \mathcal{N}(0,1)$

Conditional probability & Bayes

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad , \quad P(B) > 0$$

נוסחת הכפל:

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$

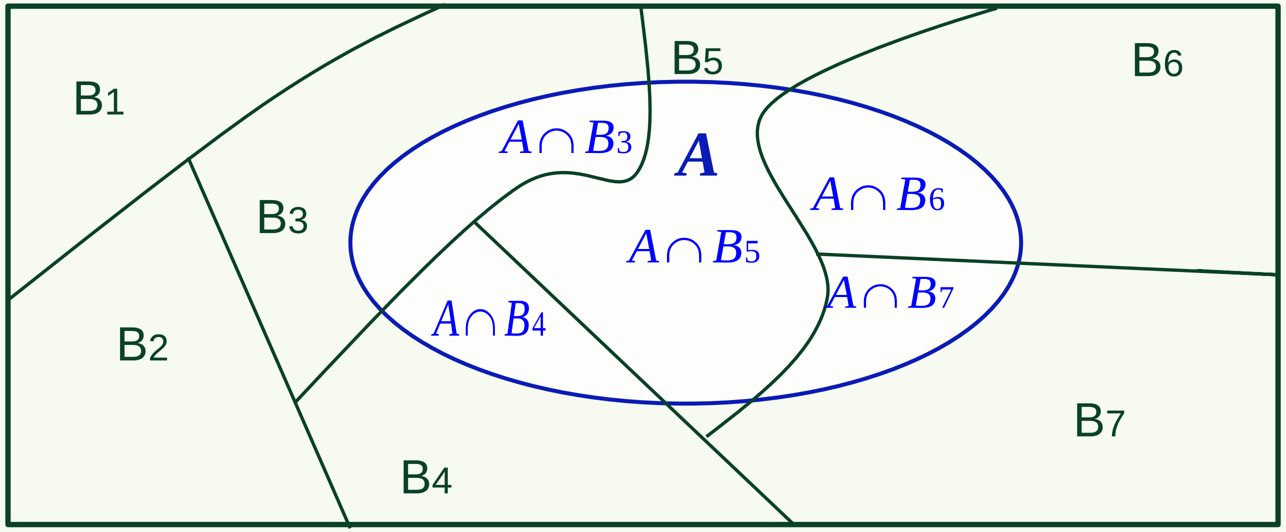
$$P(A \cap B) = P(B) \cdot P(A \mid B)$$

נוסחת בייז:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

Total probability Theorem

{Bi} divide Ω where $B_i \cap B_j = \phi$, $\forall i, j$



$$A = \bigcup_i (A \cap B_i)$$

$$\Rightarrow \qquad P\{A\} = \sum_I P\{A \cap B_i\}$$

$$\Rightarrow \qquad P\{A\} = \sum_I P\{A/B_i\} \cdot P\{B_i\}$$

$$\Rightarrow \qquad P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

Bayes' Theorem: Basics

- Total probability Theorem:

- Bayes' Theorem: $P(H|\mathbf{X}) = \dfrac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$
  - Let **X** be a data sample ("*evidence*"): class label is unknown
  - Let H be a *hypothesis* that X belongs to class C
  - Classification is to determine P(H|**X**), (i.e., *posteriori probability):* the probability that the hypothesis holds given the observed data sample **X**
  - P(H) (*prior probability*): the initial probability
    - E.g., **X** will buy computer, regardless of age, income, …
  - P(**X**): probability that sample data is observed
  - P(**X**|H) (likelihood): the probability of observing the sample **X**, given that the hypothesis holds
    - E.g., Given that **X** will buy computer, the prob. that X is 31..40, medium income

Thank you!

The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University