92/6/25 Data mining جلسه ۲

در این فصل TAN آمده فصل های دیگر را مدرنی کرده و خلاصه ای گفته.

جلسه پیش درمورد اینکه چرا ما باید داده ها را ذخیره کنیم صحبت کردیم. انتظاراتمان آن مثال قبلی مان را کامل کرده و از چند دیدگاه این مسئله را بررسی کرده و مسائلی که بیشتر اهمیت داده مسائلی است که ما هم بصورت روزمره با آنها سروکار داریم.

Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining

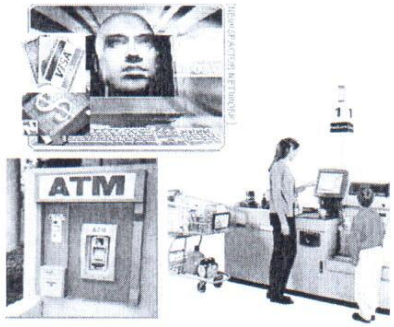
by

Tan, Steinbach, Kumar

دلایلی که از Datamining استفاده میکنیم

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions



خرید از بزرگ فروشگاه خیلی بزرگ

مسائل بانک ها

- Computers have become cheaper and more powerful

ارزان شدن کامپیوترها

- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

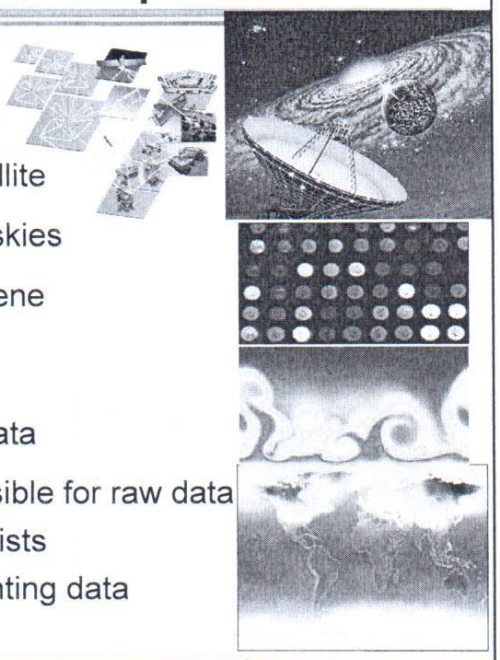
سرعت قدرت کامپیوترها زیاد شد و نرم افزارهای قوی تری برای کامپیوترها بوجود آمد.

پیشرفت در نرم افزار

پیشرفت در سخت افزار

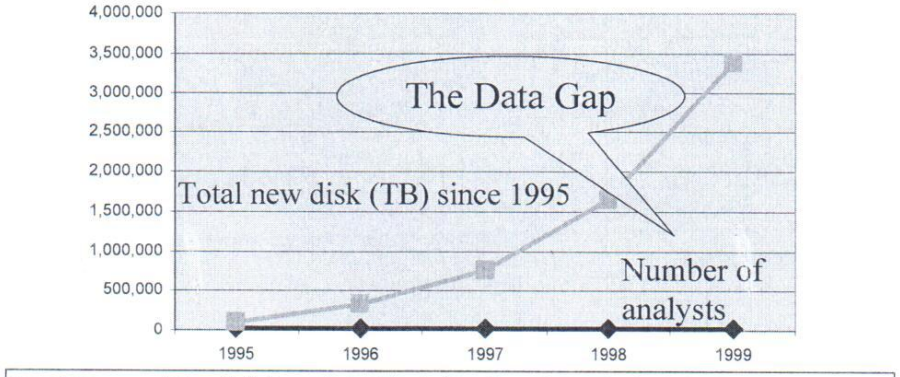
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



Mining Large Data Sets - Motivation

- **There is often information "hidden" in the data that is not readily evident**
- **Human analysts may take weeks to discover useful information**
- **Much of the data is never analyzed at all**



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

مثال های زیادی میتوان در این زمینه زد، اما هدف ما این است که ما میتوانیم برای داده کاوی مشخص کنیم این است که ما هدف اصلی مان، کارکردن با حجم بزرگی از داده هاست و روش های Data Analizez نمیتوانند معمولاً به این سؤالات پاسخ دهند. حالات پیش تفاوت Data Analize و Data mining چیزهایی اش را گفتیم. همین حجم بزرگ داده هایند و مشکلی است که ما نمیتوانیم بر روی Data ها Analize انجام دهیم علاوه بر اینکه در Data Analizez همچنان از قبل مشخص است، یعنی ما به دنبال کشف یک چیز جدید در داده ها نیستیم. یک فرضیه را که قبلاً داشته و ما در مورد آن فرضیه ادعا داریم تطبیق دهیم. تحلیل داده ها به یک تصمیم ختم میشود و این تصمیم را ما با بررسی داده ها بگیریم.

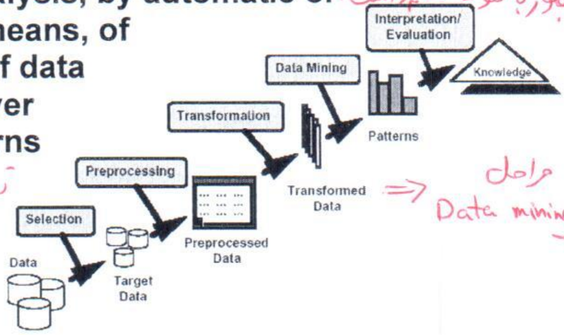
What is Data Mining?

- **Many Definitions**
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

تعریف ۱: استخراج غیر بدیهی، یعنی ما میخواهیم اطلاعاتی را استخراج کنیم از یک سری داده که واقعاً هم نیستند که قبلاً مجهول بوده و بطور بالقوه مفید است.

تعریف ۲: توصیف، تحلیل به دلیل یک روش اتوماتیک یعنی un super vise یعنی برای یک حجم بزرگی از داده ها و کشف الگوهای با معنی.

چندین تعریف داده کاوی وجود دارد هر کس از زاویه دید داده کاوی را تعریف کرده است. بعضی ها داده کاوی را همان تحلیل داده ها تعریف کرده اند اما بعداً پشیمان شده اند.



مراحل Data mining

ولی در Data mining شاید در نهایت منجر به یک تصمیم شود، ولی آن مرحله دوم است، مرحله اول این است که اصلاً سؤال ما چیست یعنی حتی ما در داده کاوی ممکن است سؤالی هم پیدا کنیم و بعد بخواهیم پاسخ دهیم و یا سؤال و پاسخ را با هم دیگر پیدا کنیم.

What is (not) Data Mining?

چه داده کاوی نیست؟ چه داده کاوی است؟

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

وقتی به دفترچه تلفن شماره ای را نگاه میکنیم این واقعاً Data mining نیست به این دلیل که از آن نمیخواهیم اطلاعاتی بگیریم نباید هم بتوان اطلاعاتی به دست آورد اما واقعاً آن طور Data ها Determinstice است، داده های stocastic نیست. یا مثلاً یک query در مورد آمازون یک سایتی که اسمش کارش کتاب فروشی بوده اما ما را هم چی می فرستد بخوانیم.

بگیرم که این هم دوباره در دسته داده کاوی نمی گنجد. ولی اگر بیائیم اسامی آن آدم ها در دفترچه تلفن که در منطقه خاصی از امریکا هستند را بررسی کنیم این ممکن است که داده کاوی باشد.

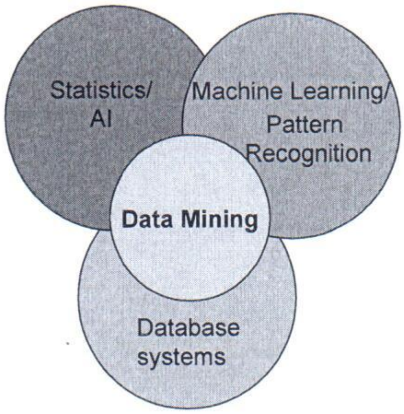
خیلی بعید است که داده کاوی مال نیست؟! مال آمار هست؟ مال مهندسی کامپیوتر هست؟! مال علوم کامپیوتر؟! اگر نگاه کنید متوجه میشوید که داده کاوی مال همه است جز صنایع البته هم میتوانند استفاده کنند. در رابطه داده کاوی از سه بخش عمده درست شده، مهم ترین بخش، قدیمی ترین بخش statistics است و بعد Database system هست و نهایتاً Pattern Recognition یا Mashin learning است و اگر این ۳ را ما با هم ادغام کنیم به Data mining می رسیم.

چرا آماری ها خیلی در داده کاوی موفق نیستند؟! چون Database system بلد نیستند. Mashin learning شان خیلی قوی نیست. چرا کامپیوتری ها خیلی قوی نیستند؟! چون از روش های آماری گریزان هستند، اینها را نمیتوان از هم جدا کرد، ... چون عامل ... (min19)

یک کامپیوتری هم آن آمار بلد نباشد، کار خوبی نمی‌تواند ارائه دهد. پس لازم است این است که ما آمار بدانیم Database
بدانیم، Pattern Recognition و Mashin learning هم بدانیم. تمام درس‌های Pattern Recognition را میتوان دو جور درس
داد: با Base آماری و بدون Base آماری که با Base آماری اش خیلی بهتر است چون برای حل یک مسئله معمولاً ۳ روش
داریم: تحلیلی، عددی و استفاده از شبیه سازی که آماری است. ما معمولاً در روش اول چیزی برای حل کردن به ما نرسیده، تمام شده، همه چی بوده
حل کردند یا آنچه هم مونده خیلی حل کردنش سخته، این روش‌های آماری معمولاً فرض‌های flexiable تری دارند و راحت‌تر میتوانند با مسائل
برخورد کنند و حلش کنند.

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data mining اما باید Data Anazise دو کار انجام می‌دهد: در Data Analiyiz ما آمار را به دو بخش تقسیم می‌کنیم که آمار توصیفی، آمار استنباطی
در داده‌کاوی بجای Infrence از کلمه Prediction استفاده میکنیم، Prediction به معنای پیش بینی است البته پیش بینی آینده را
forcasting می‌گویند با prediction متفاوت است. Prediction به معنی Interpolation است یا آینده نزدیک یا گذشته نزدیک.

Data Mining Tasks

در داده‌کاوی برخلاف آمار که هدف همان Infrence است هدف prediction است. حالا ممکن است جایی هم Infrence انجام دهد یعنی استنباطی انجام دهد.

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

کار را حالا دیگری که در داده‌کاوی انجام میشه توصیف است، یعنی واقعاً یک Base ریاضی نداریم اما با یک شکل کشیدن مسئله را میتوانیم خیلی
اطلاعات بگیریم. یک سایتی هست به نام gapminder که یک سری داده‌ها را نشان می‌دهد. مثلاً خط فقر که می‌آید نقشه GAS
این از جهان به ما نشان می‌دهد، یا یک کشور و قسمت‌ها را در یک نمایش آن را به ما نشان می‌دهد. ما با چند تا کلیک متوجه می‌شویم که

فقیرترین قاره کجاست، قوی‌ترین قاره کجاست و ... یعنی اطلاعاتی که ما باید روزها دنبالش بگردیم. مثلاً برای GPA تعدادهاش را اینجا با
یک نمودار آنی به ما نشان می‌دهد. نقشه GAS این، یعنی یک نقشه را میگیرد و در آن رنگ آمیزی میکند، آن رنگ‌هایی که شبیه یکدیگر است
در آمدهایشان یکی است.

چه مطالبی را میخواهیم در Data mining آموزش دهیم!! classification / clustering / ... در صفحه بعد (اسلاید ۹)

Sequential Pattern Discovery و Association Rule Dis در تا ابزاری هستند که تقریباً همه جا به کار می آید و اطلاعاتی از داده ها استخراج میکنند که خیلی به ما کمک میکند. Regression برای Interpolation استفاده میشود و extrapolation ولی با Regression خیلی از چیزها را میتوان Predict کرد: ما قد و وزن را داریم میخواهیم ببینیم قد و وزن چه رابطه ای با هم دارند. براساس این می آییم یک نمودار رشد برای نوزادان درست می کنیم. این نمودار رشد یک Regression بوده که به ظاهر غیر خطی است اما واقعاً خطی است.

Data Mining Tasks...

1- Classification [Predictive]

2- Clustering [Descriptive]

3- Association Rule Discovery [Descriptive]

4- Sequential Pattern Discovery [Descriptive]

5- Regression [Predictive]

6- Deviation Detection [Predictive]

Deviation Detection مثل Anomaly Detection است، Anomaly یعنی داده ی نابهنجار، میخواهیم ببینیم آیا سرقتی توی بانک صورت گرفته براساس تراکنش هایی که انجام شده، صورت آنها با این مسئله درگیر هستند، آیا میتوانند به موقع تشخیص دهند میتواند خیلی از وقایع جلوی دزدی ها را بگیرند.

Deviation به معنی انحراف است یعنی کشف انحراف، این کشف انحراف خیلی مهم است، کشف Anomaly یا شد یعنی کشف داده های نابهنجار.

این کل مطالبی است که ما میخواهیم این ترم درس دهیم. تفاوت classification و clustering:

classification: supervise است یعنی تعداد کلاس ها را از قبل می دانیم.

clustering: unsupervise است یعنی تعداد کلاس ها را از قبل نمی دانیم.

روش هایشان شبیه هم دیگر است، در classification چون تعداد را می دانیم، همزمان pretictive است، یعنی یک نفر جدید می آید میخواهیم بدانیم در کدام کلاس بگذاریمش ولی در clustering، Descreptive است یعنی توصیف میکنیم یک سری داده داریم و اصلاً نمی دانیم کلاس هایشان چگونه است، اینها را میخواهیم تعیین کنیم.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 9

* در واقع آن هایی که روش شان مشخص است و حدس میزنیم که به چه نتیجه ای میرسیم بحث میکنیم predeetive و آنهایی که واقعاً ما نمی دانیم یعنی ذوق زده میشویم وقتی نتیجه اش را می بینیم اینها Descriptive میشوند.

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

© Tan,Steinbach, Kumar Introduction to Data Mining 4/18/2004 10

در classification ما یک سری Data داریم که این Data موجود است، بحث میکنیم training set (داده های آموزشی)
هر داده یا رکوردها شامل چند تا متغیر است، به آن ها میگوییم Attribute (هر رکوردها شامل چند متغیر است و یک Attribute به عنوان متغیر هدف، یعنی یک برچسبی دارد که این برچسب به ما میگوید این داده یا رکورد متعلق به چه کلاسی است) label پس یک سری عدد یا بردار داریم، بردارها را میخواهیم بگذاریم در چند تا کلاس. ما براساس داده های آموزش مان کلاس ها را مشخص می کنیم، تعداد کلاس ها مشخص

است ، حالا روش ما باید طوری باشد که هر وقت داده جدیدی آمد بتواند این داده را در کلاس هایی که از قبلاً مشخص شده بگذارد.

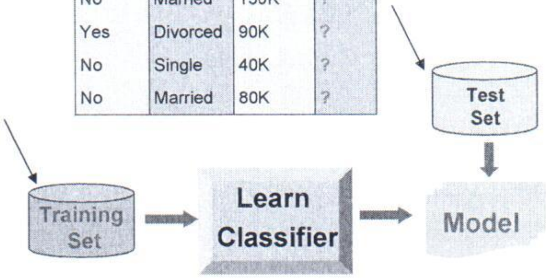
در اسلاید اقبالی از classification را می بینید که هر رکورد ما شامل انواع و اقسام متغیرها است : متغیر عددی ، متغیر اسمی ، متغیر باینری و حالا این ها را ما میخواهیم دسته بندی کنیم ، وقتی اینها را دسته بندی کردیم حالا یک set جدیدی میدهد و سوال اینجاست که اینها چیست ؟! ما براساس اطلاعات قبلی که داریم ، در مورد اینها نظر بدهید که این ها label اشان چی هست . ما این ها را به دو تا cluster تقسیم کردیم (yes , No) حالا این کلاس ها برای داده های جدید مشخص نیست ، روش ما باید طوری باشد این کار را انجام دهد.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



کاربردهایی که classification دارد خیلی خیلی زیاد است ، مثلاً یک مثال که از کشور خودمان بزنیم ، مثال ما میخواهیم استان های کشور را براساس چند شاخص خوشه بندی کنیم ، شاخص هایی که در نظر میگیرند مثل بهداشت ، درآمد ، آموزش ، فرهنگ ، نتیجه این خیلی جالب میشود یک نقشه GAS این میشود که خود مان که رنگ آمیزی شده ، آن رنگ هایی که شبیه هم آمده هستند ، یعنی آنها استان ها یا شهرهایی (کشورهای) هستند که شبیه هم آمده هستند.

Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

اگر ما بخواهیم دسته بندی کنیم خیلی جالب است مثلاً ممکن است به نتایجی برسیم که باورش سخت باشد ، مثلاً ما بگوییم از قبل که ما میخواهیم کشورمان را به ۳ بخش تقسیم کنیم . مثلاً اصفهان و خمینی شهر و تهران در یک کلاس هستند که اولین کلاس ما می شود چون در اقتصاد این سه ناحیه خیلی خیلی تفاوت زیادی دارد نسبت به بقیه.

Classification: Application 2

خودآزمایی

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 4

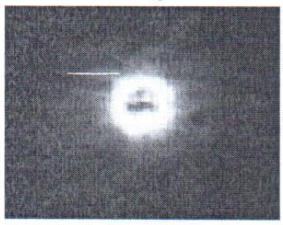
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Courtesy: http://aps.umn.edu

Early

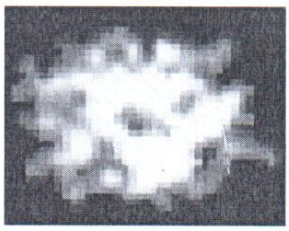


Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate

Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

هدفمان در clustering این است که داده های مشابه را دور هم جمع کنیم و تعداد کلاس ها را هم برآورد کنیم. یعنی یک قدم بیشتر از classification داریم. پس ما سعی داریم که داده هایی که شباهت بیشتری به هم دارند و فاصله کمتری از هم دارند بیاوریم یک جا و فاصله بین کلاس ها را ماکزیمم کنیم. برای همین از یک اندازه ای استفاده میکنیم به نام similarity measure که می تواند براساس داده ها تعریف شود. اگر داده ها خیلی بزرگ نباشند و داده های پرت نداشته باشیم همان فاصله اقلیدسی خوب است ولی اگر داده ی کوچک و بزرگ داشته باشیم و فاصله ی بین داده ها را زیاد کند، معیار اقلیدسی خیلی خوب جواب نمی دهد. به جای توان ۲ از قدر مطلق استفاده میکنیم یا از Lα استفاده میکنیم که با تجربه میتوان این معیار را خوب در نظر گرفت. پس می خواهیم:

within group = min

between group = max

براساس این مقیاس روش های کلاستر کردن به وجود آمده.

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

با وجود اینکه کلاسترینگ یک روش خیلی خیلی قدیمی است ولی روشی است که جدیدا خیلی بیشتر از آن استفاده میشود که شاید همان دلایلی باشد که داده کاوی زیاد از آن استفاده میشود (بالا رفتن سرعت سخت افزاری و امکانات نرم افزاری)

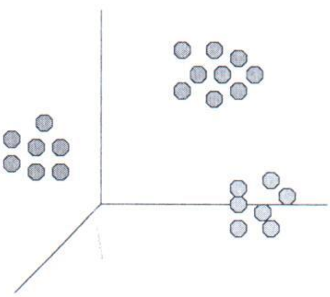
روش های کلاسترینگ خیلی خیلی زمان بر هستند و روش های کلاستری خیلی خوب را نمیتوان خیلی مواقع استفاده کرد به خاطر همین زمان بر بودنشان.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



ما میتوانیم مشتری های یک سوپرمارکت را کلاستر کنیم. چطور؟! با جمع آوری یکسری اطلاعات

ما میتوانیم متن ها را کلاستر کنیم که بحث روز است. ما میتوانیم یکسری واژه ها را در اختیار داریم روی وب یا روی Database. چطور؟ داکیومنت های مشابه را براساس خوشه بندی کنیم؟! با اینکه میخواهیم ببینیم چه واژه ای بیشتر در این داکیومنت ها استفاده میشود. هدف ما پیدا کردن گروه هایی از داکیومنت هاست که به هم شبیه است و روشی که انتخاب می کنیم هم از آن اطلاعات و آن ترم هایی که در هر داکیومنت

صورت مقدار استفاده می‌شود را اول نشان می‌دهیم، براساس این اساس سعی می‌کنیم اول آنها را کلاستر کنیم، بعد در داخل
هر خوشه‌ها یی بندی طوری باید باشد که کلاسترهایی که داریم ارتباط داشته باشند با آن داکیومنت های جدیدمان.

Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Los Angeles یک نمونه از این Document clustering را در این اسلاید 21 آورده که گفته 3204 مقاله از

را در نظر بگیرید، similarity measure ها ان را تعریف می کنیم ① چند تا کلمه در این Document ها وجود دارند

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- ① Similarity Measure: How many words are common in these documents (after some word filtering).

برخی از کلمات را در نظر نمی گیریم.

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

داده های stock را بر اساس مثل ... cluster کنیم

- Observe Stock Movements every day.
- Clustering points: Stock-{UP/DOWN}
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day.
  - We used association rules to quantify a similarity measure.

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

اسلاید 23 : ارزش های Descriptive است نه prective مثل کلاسترینگ. مثالی که جلسه پیش زدم
اینه کسی که پوشک بچه میخرد، چه چیز دیگری میخرد در یک مغازه بزرگ. گفتم نشان دادند حتما شیر هم میخرد.

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

حوزه‌های کاربرد

- Marketing and Sales Promotion:
  - Let the rule discovered be
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

اسلاید 27: خیلی وقتها داده هایی که ما داریم بصورت Sequential تولید میشوند یا نه به صورت دنباله ای نیستند ولی اگر ما آنها داده ها را پشت سر هم قرار دهیم میتوانیم به نتیجه ای برسیم. .... Sequential Pattern در واقع برمیگردد به یک بخشی از Mashin learning که اینجا استفاده میشود. ما میخواهیم ببینیم چه ارتباطی بین متغیرها است. معمولاً متغیرهای مستقل یا پارامترهای مستقل خیلی وجود ندارند، به هم یک جوری وابسته اند. اول این وابستگی ها چگونه است. ممکن است زمان هم نقش داشته باشد.

Sequential Pattern Discovery: Definition

دنباله ای

مشخصه دوم در شکل بالا این زمان است که این کار را انجام داده است.

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.

مثلاً A و B و C داریم میخواهیم بدانیم بعد از این چه اتفاقی می افتد.

(A B) (C) → (D E)

- Rules are formed by first disovering patterns. Event occurrences in the patterns are governed by timing constraints.

(A B) (C) (D E)

<= xg | >ng | <= ws

<= ms

در عین ... sequential گاهی وقتها به Time series میشود. یعنی به داده هایی که باید بعد دیگر هم اضافه شده و آن بعد زمان است. اگر داده ها به زمان وابسته باشند آن را میتوانیم Time series بنامیم. دقت کنید که تحلیل داده های معمولی با تحلیل داده های Time series تفاوت دارد چون غیر قابل تکرار هستند، چون در زمان دارند تولید میشوند. در سری های زمانی ما معمولاً علاقمندیم پیش بینی آینده یا گذشته حتی.

Sequential Pattern Discovery: Examples

چه Pattern ای داده های ما در آینده خواهند داشت. chang point رخ می دهد یعنی یک نقطه تغییر داریم. روند داریم، تغییرات فصلی داریم، سیکل داریم. اینها را ما میتوانیم predict اش کنیم، کار پیچیده ای نیست. هرچند که روش ما Descriptive است. مثال هایی از ... Pattern را می بینید که اگر مثلاً

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) --> (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

در فروشگاه کتاب، کسی این دو تا کتاب را خرید، کتاب بعدی که میخواهد بخرد این است. پس اگر ما در آمازون این دو تا کتاب را خریدیم، خود آمازون چند تا کتاب دیگر ترجیحاً پیشنهاد میکند و بعد اگر ما بخریم تخفیف هم بما می دهد. چون دیده که مثلاً اگر کسی کتاب C و C++ را خرید کتاب برنامه نویسی پرل را میخرد هم. میخرد پس خوب که ما خودمان بهش پیشنهاد دهیم.

Regrission خیلی خیلی قدیمی است، همان نمودار نوزادان در قبل گفتیم، ما هم به تصادف ۱۰۰ تا نوزاد را از جاهای مختلف که یک سن دارند را انتخاب میکنیم، در یک سن مخصوص یا در سن های مختلف و بعد سعی میکنیم به این داده ها یک خط برازش دهیم، خط امکان ندارد، یک منحنی برازش دهیم، روش های ریاضی خیلی خوب نیستند چون این داده هایی که ما داریم هیچ منحنی نمیتوان از آنها رد کرد هر منحنی با درجه خود نشان که فاصله صفر ندارد و خیلی پارامترهای زیادی دارد به خاطر همین مدل آماری میگیریم و یک خطاهایی را قبول می کنیم و بر اساس آن خطاهایی که ما قبول کردیم میتوانیم یک منحنی برازش دهیم. منحنی که برازش دادیم میتوانیم آن نمودار را رسم کنیم.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

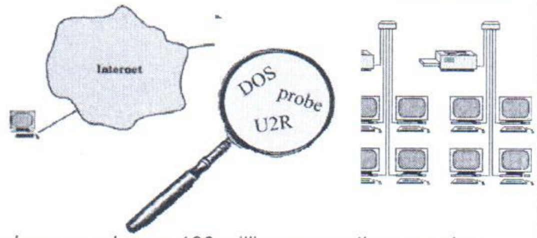
Regression با Time seri تفاوت است، هدف Regression، Enterpolation است، هدف Time seri extrapolation است. هدف Regression، prediction است، هدف Time seri: forcasting است.

Regression یک ابزار خیلی خوبی است که به طور خیلی کار learning را انجام دهد در واقع یک statiscal learning است. اگر فرض های آماری که روی مسئله است برقرار نباشد، ما این مسئله را با neual Network حل می کنیم. در واقع neual Network کاری را انجام میدهد همان کار Regression است اما تفاوت Regression و neual Network و Regression چیه؟ در neual فرض های اولیه زیادی وارد نمی شود در حالی که در Regression فرض های اولیه زیادی داشت. به خاطر همین است که اگر فرض های اولیه آماری برقرار باشند حتماً حتماً جواب بهتری از Neual Network می دهند.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

اسلاید ۳۰ = Anomaly گفتیم یعنی نابهنجار – نابهنجار یعنی خرید نابهنجار، داده نابهنجار، یعنی یک اتفاقی در Database – مثلاً داده که به روال معمول نیست می‌خواهیم اینها را مشخص کنیم، یا اگر اتفاق معمول افتاده باشد، می‌خواهیم ببینیم آیا واقعاً این داده نابهنجار به صورت رندوم قبلاً پیش آمده یا اینکه به صورت systematic پیش آورده. اینها کارهایی است که ما میخواهیم انجامش دهیم و یکی از بحث های روز است Anomaly Detection. اما در داده کاوی ما با چالش های جدی روبه رو می شویم که در داده کاوی

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality ⇒ یک اصل انکارناپذیر است
- Data Ownership and Distribution
- Privacy Preservation } دو مورد آخر به عهده خودتان مطالعه کنید.
- Streaming Data

با داده‌های مواجه می‌شویم که اینها با مقیاس اسمی - فاصله ای و... اندازه‌گیری شده‌اند. لزوماً با مقیاس نسبتی اندازه‌پذیری نشده‌اند.
با فاصله همین Scalable نیستند، نمی‌شه با هم جمع‌شان کرد، ضربشان کرد، قواعد خاصی بر روی آنها حاکم است. موضوع بعدی
ما یکی از موضوعاتی هست که روز داریم برایش راه حل جدی می‌دهیم، Dimensionality یا بُعد داده‌ها است یعنی تعداد متغیرهای
که در داده‌کاوی خیلی وقتها از تعداد داده‌های ما بیشتر است. مثلاً ما می‌خواهیم ۵۰ نفر را به عنوان شاهد نگه‌داریم، ۵۰ نفر را یک
آزمایش روی‌شان انجام دهیم، می‌خواهیم آن‌ها اینها را با هم مقایسه کنیم، چند تا ژن مثلاً ده هزارتا که در اینجا ما ۵۰ تا داده داریم،
۱۰ هزار تا متغیره هیچ روشی آماری وجود ندارد که ما بتوانیم ۱۰ هزارتا متغیر با ۵۰ تا داده بتوانیم ازش نتیجه بگیریم. این ابعاد که بزرگ
شود هیچ روشی نمی‌تواند این داده‌ها را تحلیل کند و باید روش جدیدی ابداع کنیم برای داده‌های Highdenmential یا اینکه بعد داده‌ها را
کم کنیم مثلاً روش‌های PCA، ICA و random projection و... که روش‌هایی هستند که بعد داده‌ها را کم می‌کنند، داده‌ها را می‌برند به
یک فضای دیگر، در آن فضا مسئله را حل می‌کنند و جواب را برمی‌گردانند، هرکدام یکسری محدودیت نسبت به روشی دارند.
ساختار داده‌های ما هم کلاً خیلی پیچیده باشد، پیچیدگی به معنی نوع داده‌ها، توزیع داده‌ها، هم بستگی داده‌ها، اندازه داده‌ها
و غیره‌ی در داده‌کاوی Data quality است، وقتی یک سرشماری انجام می‌شود در یک کشور، هر ده سال یکبار سرشماری انجام
می‌شود، این داده‌هایی که از طریق سرشماری به دست آورده می‌شود، یک سال طول می‌دهد تا Quality‌شان را چک کنند تا ایرادات آن را
برطرف کنند بنابراین همین است که نتیجه سرشماری ابداً در سال اعلام می‌شود، اگر کیفیت داده‌ها خوب نباشد ما هیچ راه‌حلی در داده‌
کاوی نداریم. مالکیت داده‌ها و توزیع داده‌ها خیلی سخت است، ما نمی‌توانیم درک درستی از کل Data بگیریم.