داده‌کاوی

Data Mining

M.M. Pedram
pedram@khu.ac.ir
Kharazmi University
(Fall 2009)

1

اهداف درس

- آموزش مفاهیم یادگیری ماشین و داده کاوی
- ارایه چشم‌اندازی از زمینه های تحقیقاتی فعلی و آینده
- آشنایی با کاربردهای موفق
- آشنایی با اصول انتخاب روش مناسب برای یک مساله خاص

2

ارزیابی

- پایان‌نیم‌سال + امتحان‌های کوچک (کوئیز) ۱۶ نمره
- تکالیف عادی + تکالیف برنامه‌نویسی ۲ نمره
- ۲ مورد تحقیق + گزارش + ارائه ۲ نمره

- زمان ارائه تحقیق اول: ۲۰ آبان
- زمان ارائه تحقیق دوم: ۲۰ آذر

سطوح علم

- معرفت (wisdom)
- دانش (knowledge)
- اطلاعات (information)
- داده (data)
- جهل (ignorance)

"Data is not information; information is not knowledge; knowledge is not wisdom." Gary Flake, Principal Scientist & Head of Yahoo! Research Labs, July 2004.

Data Mining Definition

- **داده کاوی:** استخراج الگوها یا دانش از حجم زیادی داده.

- DATA MINING: exploration & analysis
  - by ***automatic means actionable***
  - of ***large quantities of data***
  - to discover ***actionable*** patterns & rules

- Data mining *a way to utilize massive quantities of data that businesses generate*

Data Mining Definition

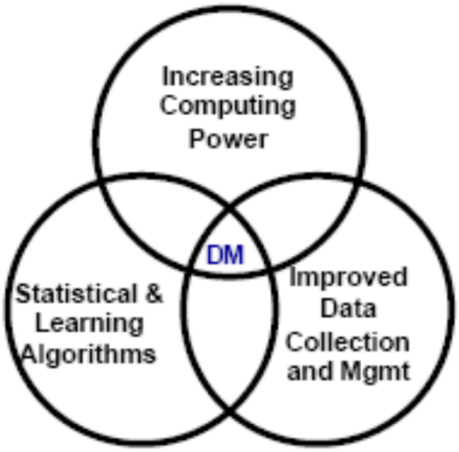
The search for interesting patterns and models,
in large data collections,
using statistical and machine learning methods,
and high-performance computational infrastructure.

Key point: applications are
- data-driven and
- compute-intensive

Characteristics of Data Mining Applications

- *Data*
  - Lots of data, numerous sources
  - Noisy: missing values, outliers, interference
  - Heterogeneous: mixed types, mixed media
  - Complex: scale, resolution, temporal, spatial dimensions

همگرایی سه تکنولوژی



- DBMS
- AI, Machine Learning, Pattern Recognition
- Data Visualization

- اسامی دیگر
  - knowledge discovery (mining) in databases (KDD) (*wrong!*)
  - knowledge extraction
  - data/pattern analysis
  - data archeology
  - data dredging
  - information harvesting
  - business intelligence
  - ...
- آیا هر مطالعه ای داده کاوی است؟
  - query processing.
  - expert systems or small ML/statistical programs

DM Tasks (Goals)

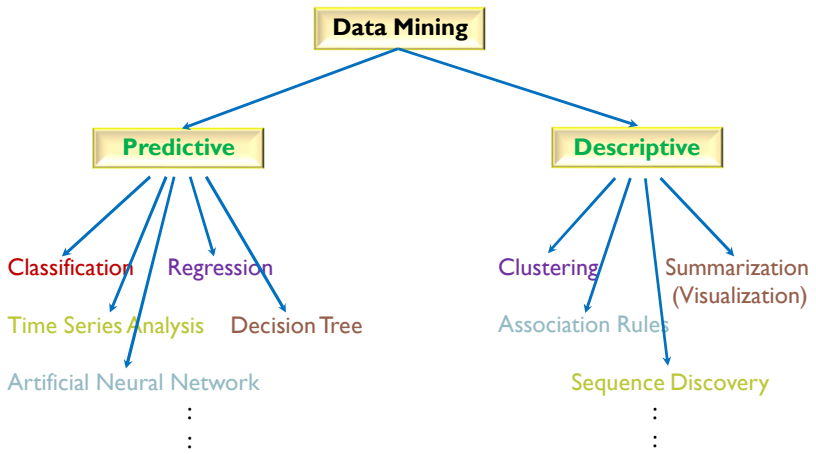
- There are two categories of goals (or high level tasks) in DM:
  - Description: models are constructed to describe particular patterns or relationships in the data
  - Prediction: models are constructed using historical cases to predict outcomes for new cases

DM Tasks (Goals)

Another definition:

- Descriptive data mining: Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.
- Predictive data mining: Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy of those models and procedures when applied to fresh data.

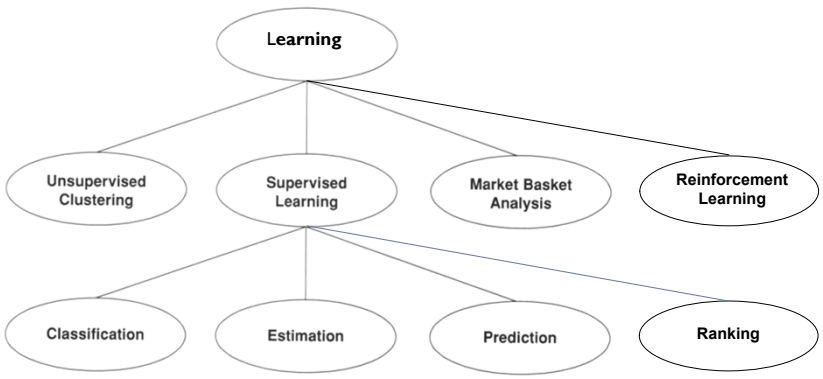
Data Mining Tech

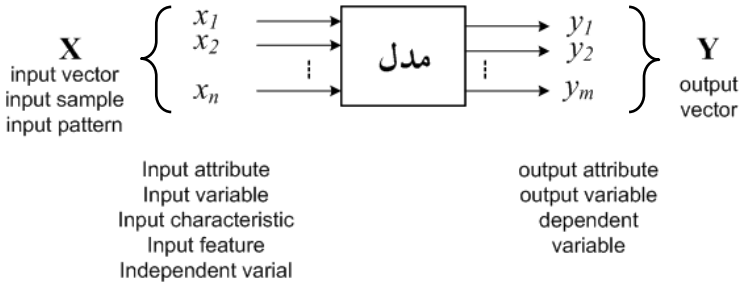


Other names

- Clustering: Segmentation, Database Segmentation
- Link Analysis: Discovering Association Rules, Sequential Patterns, and time Sequences
- Deviation Detection: Visualization, Statistics

Fundamental Types of Learning





انواع متغیر

- عددی(numeric)
- رسته‌ای(categorical)

مراحل کار با مدل

- یادگیری(learning) یا آموزش(teaching)
- فراخوانی(recall)

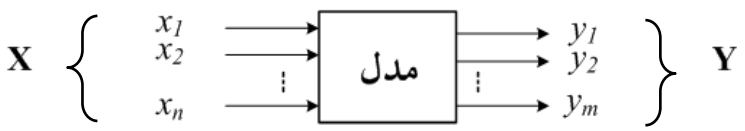
کلاس‌بندی

- متغیرهای خروجی از نوع رسته ای هستند.
- وظیفه مدل، انتساب هر الگوی ورودی به یک کلاس از خروجی است.
- یادگیری مدل از نوع نظارت شده است.
- خروجی ها، وضعیت فعلی(نه در آینده) را بیان می کنند.

- روش‌های کلاس‌بندی
  - درخت تصمیم‌گیری
  - قواعد
  - شبکه‌های عصبی
  - گونه‌های فازی

تخمین

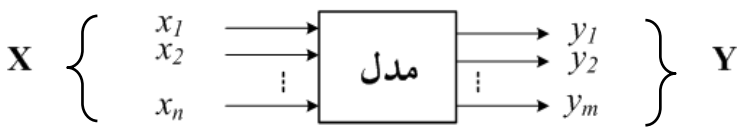
- متغیرهای خروجی از نوع عددی هستند.
- وظیفه مدل، انتساب یک عدد در خروجی به هر الگوی ورودی است.
- یادگیری مدل از نوع نظارت شده است.
- خروجی ها، وضعیت فعلی(نه در آینده) را بیان می کنند.



- روش‌های تخمین
  - قواعد
  - شبکه‌های عصبی
  - رگرسیون آماری
  - گونه‌های فازی
  - سری‌های زمانی

پیش‌بینی

- متغیرهای خروجی از نوع عددی ویا رسته‌ای هستند.
- وظیفه مدل، انتساب یک عدد یا کلاس در هرخروجی به هر الگوی ورودی است.
- یادگیری مدل از نوع نظارت شده است.
- خروجی ها، وضعیت در آینده را بیان می کنند.



- روش‌های پیش‌بینی
  - درخت تصمیم‌گیری
  - قواعد
  - شبکه‌های عصبی
  - رگرسیون آماری
  - گونه‌های فازی
  - سری‌های زمانی

خوشه‌بندی

- متغیرهای خروجی از نوع رسته‌ای هستند.
- وظیفه مدل، کشف واستخراج ساختارهایی در داده های ورودی.
- یادگیری مدل از نوع بدون نظارت است.

کاربردها

- تعیین روابط معنادار در داده‌ها.
- ارزیابی عملکرد یک مدل که یادگیری نظارت شده، دارد.
- تعیین بهترین مجموعه از متغیرهای ورودی برای یادگیری با نظارت.
- آشکارسازی برون‌هشته‌ها(outliers).
  - برون‌هشته: داده‌ای که به طور عادی با سایر نمونه‌ها، گروه بندی نمی شود.

- روش‌های خوشه‌بندی
  - آماری (k-means و c-means)
  - شبکه‌های عصبی خودسازمان‌ده
  - گونه‌های فازی

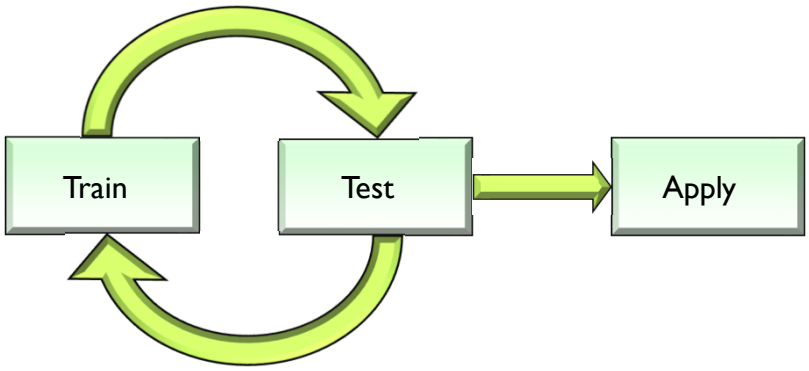
تحلیل سبد بازار(Market Basket Analysis)

یافتن ارتباطات جالب بین محصولات خرده‌فروشی.

- روش‌های تحلیل سبد بازار
  - قواعد همباشی (association rules)

Model Development

Training and Testing(Model validation)



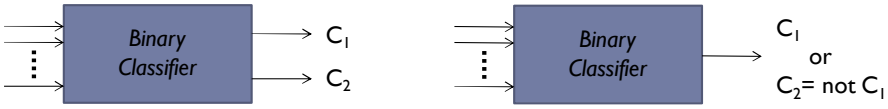
Evaluation of classifiers

- For any data set that is used to test a classifier (model), a *confusion matrix* can be built.

- Classification
  - Binary class
  - Multi-class

Evaluation of classifiers

- Binary classification



		Classifier Output	
		Positive	Negative
Actual	Positive	124	15
	Negative	8	84

agreement

- Error = "proportion of incorrect classification"
  =(8+15)/(124+84+8+15) = 23/231
- Accuracy = 1-Error = "proportion of correct predictions" =208/231
- Precision = "proportion of predicted positive cases that were correct"
  = 124 / (124 + 8)

Evaluation of classifiers

- Entries in the confusion matrix for binary classification have names:

		Classifier Output	
		Positive	Negative
Actual	Positive	124 → TP	15 → FN
	Negative	8 → FP	84 → TN

TP = hit = "number of positive cases correctly identified" = 124

FN = a miss = "number of positive cases incorrectly classified as negative" = 15

FP = false alarm = "number of negative cases incorrectly classified as positive" = 8

TN = correct rejection = "number of negative cases correctly identified" = 84

Evaluation of classifiers

- Entries in the confusion matrix for binary classification have names:

		Classifier Output	
		Positive	Negative
Actual	Positive	124 → TP	15 → FN
	Negative	8 → FP	84 → TN

Recall = **Sensitivity** = ***SENS*** = **TP** rate = "proportion of positive cases correctly identified" = TP/(TP+FN) = 124 / (124 + 15)

FN rate = "proportion of positive cases incorrectly classified as negative" = 15 / (124 + 15)

FP rate = "proportion of negative cases incorrectly classified as positive" = 8 / (84 + 8)

Specificity = ***SPEC*** = **TN** rate = "proportion of negative cases correctly identified" = 84 / (84 + 8)

Evaluation of classifiers

- Entries in the confusion matrix for binary classification have names:

		Classifier Output		
		Positive	Negative	
Actual	Positive	124 → TP	15 → FN	***Recall*** or ***SENS***
	Negative	8 → FP	84 → TN	***SPEC***
		Precision	***NPV***	

Recall = **Sensitivity** = ***SENS*** = **TP** rate = TP/(TP + FN) = 124 / (124 + 15)

Precision = **PPT** = **Positive Predictive value**= "proportion of correctly classified ones as positive case to the all ones classifies as positive" = TP/(TP + FP) = 124 / (124 + 8)

NPV = **Negative Predictive value** = "proportion of correctly classified ones as negative case to the all ones classifies as negative" = 84 / (84 + 15)

Specificity = ***SPEC*** = **TN** rate = TN/(TN + FP) = 84 / (84 + 8)

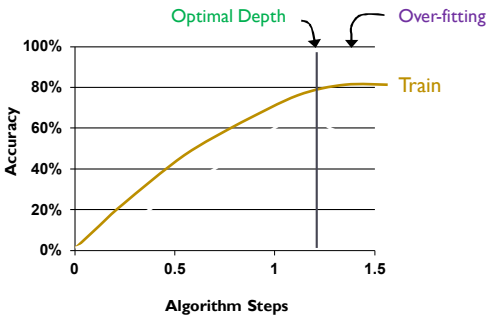
Evaluation of classifiers

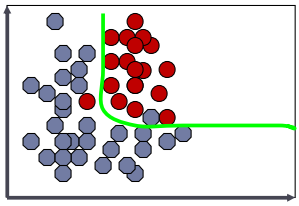
- Multi-Class classification

		Classifier Output		
		C_1	C_2	C_3
Actual	C_1	140	20	22
	C_2	17	54	8
	C_3	12	4	76

agreement

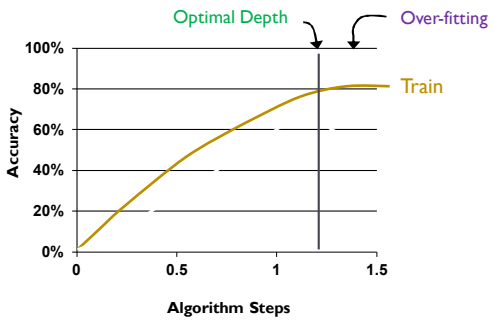
Over-fitting

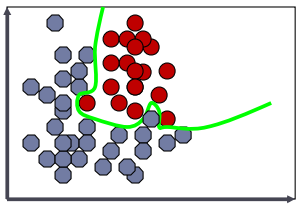




- Running too many epochs can result in over-fitting.

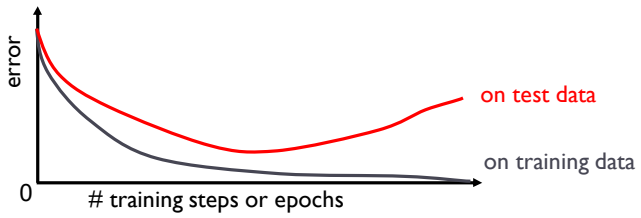
Over-fitting





❖ Running too many epochs can result in over-fitting.

Over-fitting



Scientific vs. Commercial Data Mining

Goals:

- **Science – Theories: Need for insight and theory-based models, interpretable model structures, generate domain rules or causal structures, support for theory development**
- **Commercial – Profits: black boxes OK**

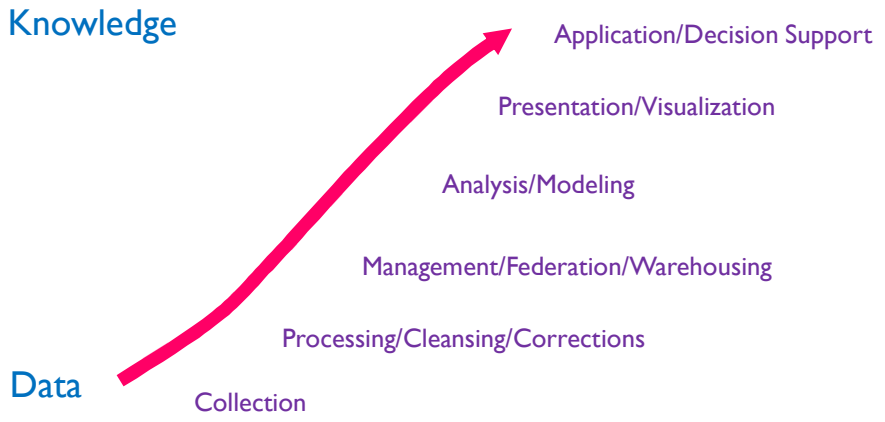
Types of data:

- **Science – Images, sensors, simulations**
- **Commercial - Transaction data**
- **Both - Spatial and temporal dimensions, heterogeneous**

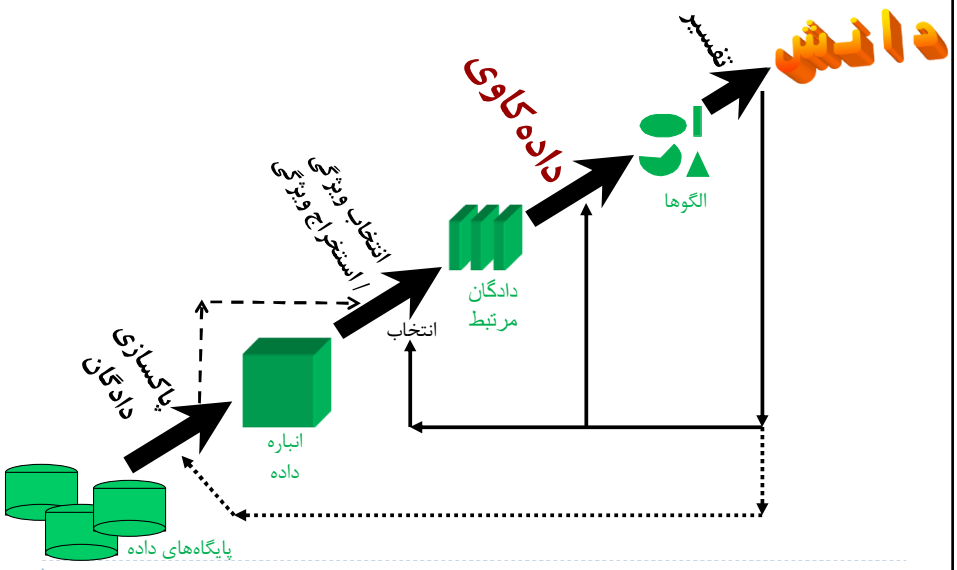
IT (information technology) tools fit both enterprises

- **Database systems (Oracle, DB2, etc),**
- **integration tools (Information Integrator),**
- **web services (Blue Titan, .NET)**

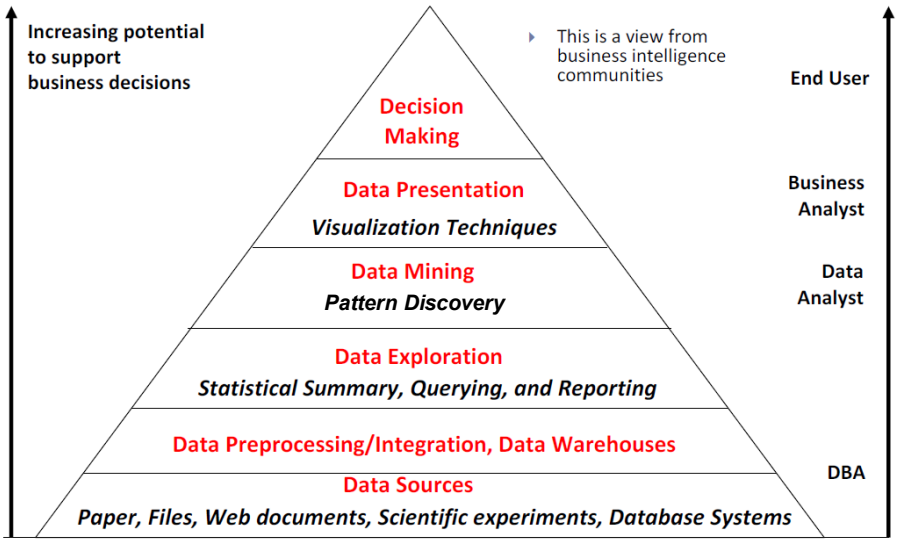
Knowledge Discovery Process



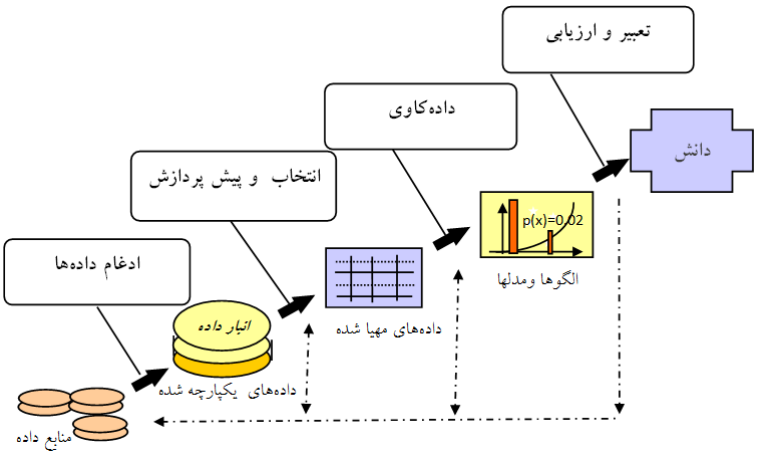
فرایند اکتشاف دانش(KDD)



KDD Process



اکتشاف دانش در پایگاه داده ها (*KDD*)



1. تشکیل انبار داده:
   - با توجه به عنوان، این مرحله براي تشکیل محیطي پیوسته و یکپارچه جهت انجام مراحل بعدي و داده کاوي در آن، انجام مي‌گیرد.
   - انبار داده مجموعه پیوسته و طبقه‌بندي شده است که دائما در حال تغییر بوده و پویا است و براي کاوش آماده مي‌شود.

2. انتخاب داده‌ها و پیش‌پردازش:
   - در این مرحله براي کم کردن هزینه‌هاي عملیات داده‌کاوي، داده‌هایي از پایگاه داده انتخاب مي‌شوند که مورد مطالعه هستند و هدف داده‌کاوي دادن نتایجي در مورد آن‌هاست.
   - براي انجام عملیات داده‌کاوي لزوماً باید تبدیلات خاصي روي داده‌ها انجام گیرد ممکن است این تبدیلات خیلي راحت و مختصر مثل تبدیل *byte* به *integer* باشد یا خیلي پیچیده و زمان بر و با هزینه‌هاي بالا مثل تعریف صفات جدید و یا تبدیل و استخراج داده‌ها از مقادیر رشته‌اي و ... باشد.

3. داده‌کاوی:

- در این مرحله با استفاده از تکنیک‌هاي داده‌کاوي داده‌ها مورد کاوش قرار گرفته، دانش نهفته در آنها استخراج شده و الگوسازي صورت مي‌گيرد.

4. تفسیر نتیجه:

- در این مرحله نتایج و الگوهاي ارائه شده توسط ابزار داده‌کاوي مورد بررسي قرار گرفته و نتایج مفید معین مي‌شود و در صورت عدم دقت، روند تکرار مي‌شود.