

بسمه تعالی

یادگیری ماشین

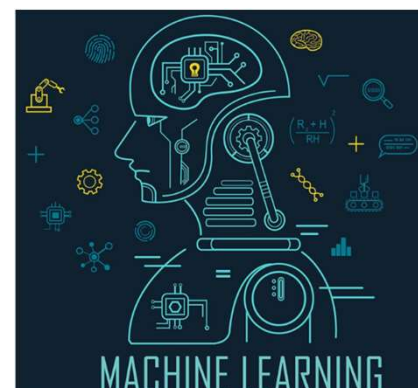
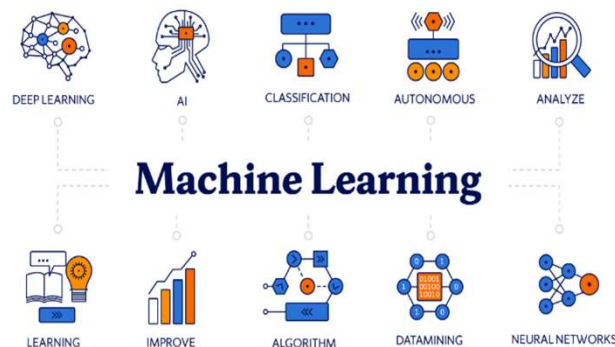
Machin Learning

مدرس

فاطمه دارائی

f_daraei@semnan.ac.ir

<https://fdaraei.profile.semnan.ac.ir>





دانشگاه سمنان

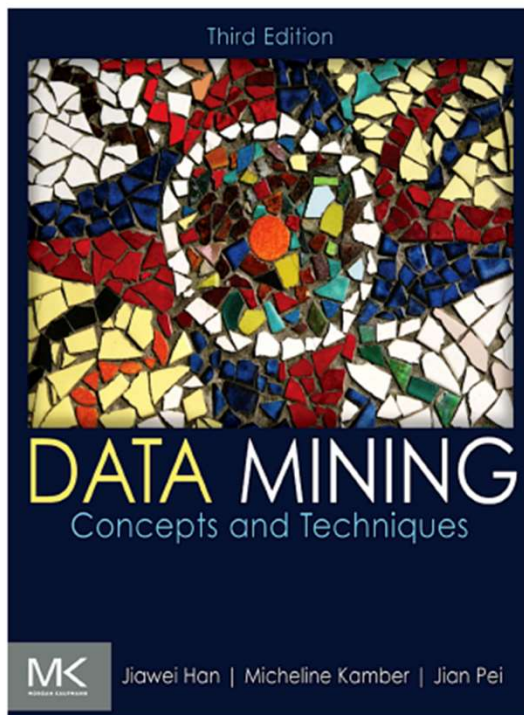
دانشگاه سمنان

Semnan University

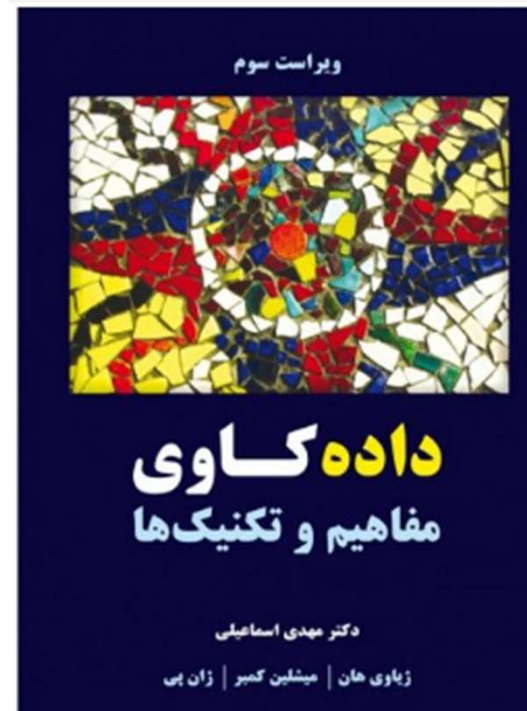
پروفسور فرزانه

مراجع اصلی درس

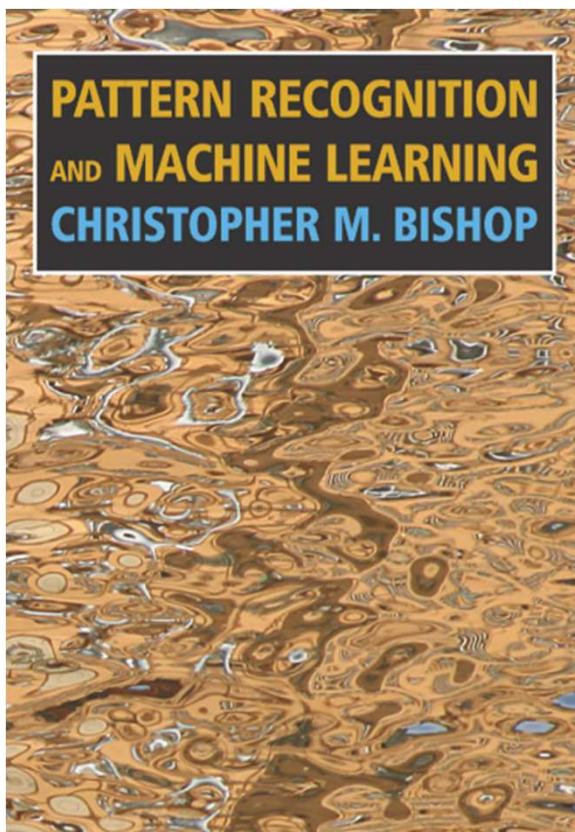
Data Mining: Concepts and Techniques (3rd ed.)
Jiawei Han, Micheline Kamber, and Jian Pei



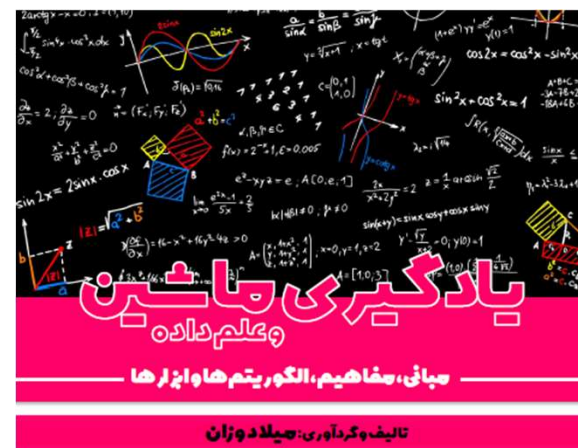
داده کاوی مفاهیم و تکنیک ها
نویسنده ژیاوی هان
مترجم مهدی اسماعیلی



C. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.



کتاب یادگیری ماشین و علم داده
 مبانی، مفاهیم، الگوریتم‌ها و ابزارها
 نویسنده: میلاد وزان



ارزشیابی

۶ نمره

میان ترم



۱۰ نمره

پایان ترم



۲ نمره

تکلیف



۳ نمره

پروژه



مجموع از ۲۱ نمره

سرفصل مطالب

✓ فصل اول: معرفی یادگیری ماشین

شناخت انواع داده ها و ویژگی ها، توصیف آماری داده ها

✓ فصل دوم: پیش پردازش داده ها

تجمیع داده، پاک سازی داده، عملیات پیش پردازش، روش های کاهش داده

✓ فصل سوم: دسته بندی

Classification , Knn

✓ فصل چهارم: درخت تصمیم

Decision Tree

✓ فصل پنجم: رده بندی و خوشه بندی

Kmeans, Clustering

✓ فصل ششم: شبکه عصبی

معرفی یادگیری عمیق، مروری بر شبکه های عصبی کانولوشن، بهینه سازی، Batch normalization



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پروفسور فرزانه گان

علم داده و یادگیری ماشین

مقدار داده‌ها با ظهور رسانه‌های اجتماعی، تلفن‌های هوشمند، اینترنت اشیا و دیگر پیشرفت‌های فناوری با سرعت شگفت‌انگیزی افزایش یافته است. تخمین زده می‌شود که بیش از ۹۰ درصد از کل داده‌های ایجاد شده توسط انسانها در پنج سال گذشته تولید شده‌اند.

این انفجار اطلاعات به عنوان "کلان داده" شناخته می‌شود.

امروزه، داده‌ها به عنوان ابزار و سوختی برای کسب و کارها در جهت بدست آوردن بینش‌های مهم و بهبود عملکرد آنها هستند. علم داده‌ها تقریباً تمام صنایع جهان را تحت سلطه خود درآورده است. امروزه هیچ صنعتی در دنیا وجود ندارد که از داده‌ها استفاده نکند.

اما چه کسی این بینش را بدست خواهد آورد؟

چه کسی تمام داده‌های خام را پردازش می‌کند؟

همه چیز توسط یک تحلیلگر داده یا یک دانشمند داده انجام می‌شود.

ادغام هوش مصنوعی و یادگیری ماشین در استخراج اطلاعات از داده‌ها به ما کمک می‌کنند.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

مؤسسه جهانی مک کنزی MGI گزارش می دهد (در سال ۲۰۱۱) که بیشتر شرکت های آمریکایی با بیش از ۱۰۰۰ کارمند به طور متوسط حداقل ۲۰۰ ترابایت داده ذخیره شده داشتند. MGI پیش بینی می کند که میزان داده های تولید شده در سراسر جهان سالانه ۴۰٪ افزایش خواهد یافت.

مقادیر بسیار زیادی از داده های خام!!

رشد انفجاری داده ها: از ترابایت ها به پتابایت ها

در عصر دیجیتال، در هر لحظه ترابایت ها داده تولید می شود. داده ها با سرعت زیاد جمع آوری و ذخیره می شوند (گیگا بایت در ثانیه)

جمع آوری داده و دسترسی به داده

ابزارهای خودکار جمع آوری داده، سیستم های پایگاه داده، وب، جامعه ی کامپیوتری

منابع اصلی داده های فراوان، توسعه نرم افزارهای تجاری آماده بکار

رشد فوق العاده در قدرت محاسباتی و ظرفیت ذخیره سازی

Data,
Data
Everywhere



دانشگاه سمنان

دانشگاه سمنان
Semnan University

پردیس فرزانهگان

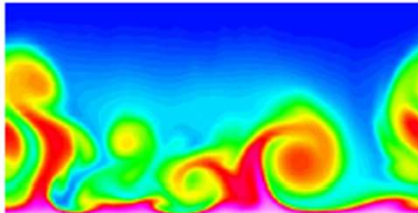
حسگرهای موجود در ماهواره ها، تلسکوپ های فضایی، دستگاه های موبایل، عکس های دیجیتال، اسناد وب.



به روزرسانی های فیسبوک، توییت ها، وبلاگ ها، محتوای تولید شده توسط کاربر. تراکنش ها، داده های سنسورها، داده های نظارتی. پرسش ها، کلیک ها، مرور صفحات.



علم: سنجش از دور، زیست اطلاعاتی، شبیه سازی علمی، ...
جامعه و همه: اخبار، دوربین های دیجیتال، یوتیوب
تجارت: وب، تجارت الکترونیک، تراکنش ها، سهام، ...





حجم زیاد داده می تواند قدرتمندتر از الگوریتم ها و مدل های پیچیده باشد

ما در داده غرق شده ایم، اما از دانش گرسنه ایم!

داده قدرت
است!!

نیاز به تحلیل داده های خام برای استخراج دانش و تحلیل خودکار مجموعه های بزرگ داده

داده ها نیز بسیار پیچیده هستند

انواع مختلفی از داده‌ها:

جداول، سری‌های زمانی، تصاویر، نمودارها و غیره

جنبه‌های مکانی و زمانی

داده‌های **به هم پیوسته** از انواع مختلف:

از تلفن همراه می‌توانیم، مکان کاربر، اطلاعات دوستی، ورود به مکان‌ها، نظرات از طریق توییتر، تصاویر از طریق دوربین، پرسش‌ها به موتورهای جستجو را جمع‌آوری کنیم.

مثال: داده های تراکنش

میلیاردها مشتری واقعی:

WALMART: دویست میلیون تراکنش در روز

AT&T: سیصد میلیون تماس در روز

شرکت های کارت اعتباری: میلیاردها تراکنش در روز

کارت های امتیاز به شرکت ها اجازه می دهند تا اطلاعات مربوط به کاربران خاص را جمع آوری کنند

مثال: داده های سند

وب به عنوان یک مخزن اسناد:

۵۰ میلیارد صفحه وب تخمین زده می شود

ویکی پدیا:

۴ میلیون مقاله (و در حال افزایش)

پورتال های خبری آنلاین:

هر روز یک جریان ثابت از ۱۰۰ مقاله جدید

تویتر:

حدود ۳۰۰ میلیون توییت در روز

مثال: داده های شبکه

وب:

۵۰ میلیارد صفحه که از طریق لینک ها به یکدیگر پیوند داده شده اند

فیس بوک:

۵۰۰ میلیون کاربر

توییتر:

۳۰۰ میلیون کاربر

پیام رسان فوری:

۱ ~ میلیارد کاربر

وبلاگ ها:

۲۵۰ میلیون وبلاگ در سراسر جهان



دانشگاه سمنان

دانشگاه سمنان
Semnan University

پروفسور فرزانه گان

تکامل علوم

قبل از سال ۱۶۰۰: علم تجربی: (Empirical Science)

در این دوره، علم بیشتر بر مشاهده و آزمایش مستقیم استوار بود. دانشمندان با جمع‌آوری داده‌های تجربی از محیط پیرامون خود، به صورت دستی نتایج را تحلیل می‌کردند. این مرحله شامل دوران پیش از توسعه نظریه‌های علمی منسجم است.

۱۶۰۰ تا دهه ۱۹۵۰: علم نظری: (Theoretical Science)

با توسعه نظریه‌های علمی، علوم وارد مرحله نظری شدند. در این دوره، مدل‌های نظری توسعه یافتند که برای توضیح پدیده‌های طبیعی استفاده می‌شدند. این مدل‌ها اغلب به‌عنوان انگیزه‌ای برای انجام آزمایش‌ها و عمومی‌سازی نتایج تجربی به کار می‌رفتند.

دهه ۱۹۵۰ تا دهه ۱۹۹۰: علم محاسباتی: (Computational Science)

با ظهور کامپیوترها و پیشرفت‌های محاسباتی، علوم وارد مرحله محاسباتی شدند. در این مرحله، بیشتر رشته‌های علمی به یک شاخه محاسباتی مجهز شدند که امکان شبیه‌سازی و تحلیل مدل‌های پیچیده ریاضی را فراهم کرد. این تغییر به ویژه در علوم مانند فیزیک، اکولوژی، و زبان‌شناسی مشهود بود.

۱۹۹۰ تاکنون: علم داده: (Data Science)

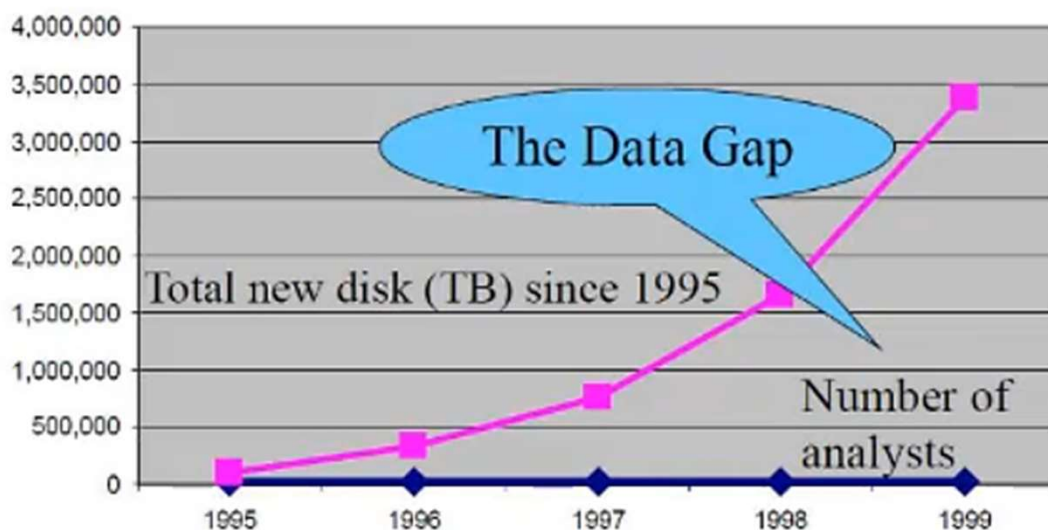
از دهه ۱۹۹۰ به بعد، با افزایش سریع حجم داده‌ها و پیشرفت در فناوری‌های ذخیره‌سازی و پردازش داده‌ها، علوم به سمت داده‌محوری حرکت کردند. در این دوره، داده‌ها از منابع مختلف از جمله ابزارهای علمی جدید و شبیه‌سازی‌ها به سرعت تولید و جمع‌آوری می‌شوند. علوم داده بر تحلیل و استخراج دانش از این داده‌ها تمرکز دارد.

اینترنت و شبکه‌های محاسباتی جهانی باعث دسترسی عمومی به این داده‌ها و تحلیل‌های مرتبط با آن‌ها شدند. داده‌کاوی در این مرحله به عنوان یک چالش و همچنین یک ابزار ضروری برای مدیریت، سازماندهی، و تحلیل داده‌های عظیم مطرح شد.



انگیزه

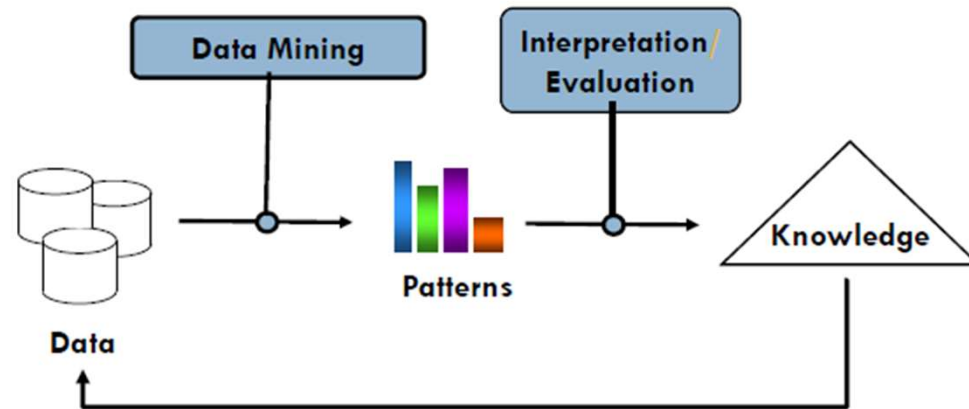
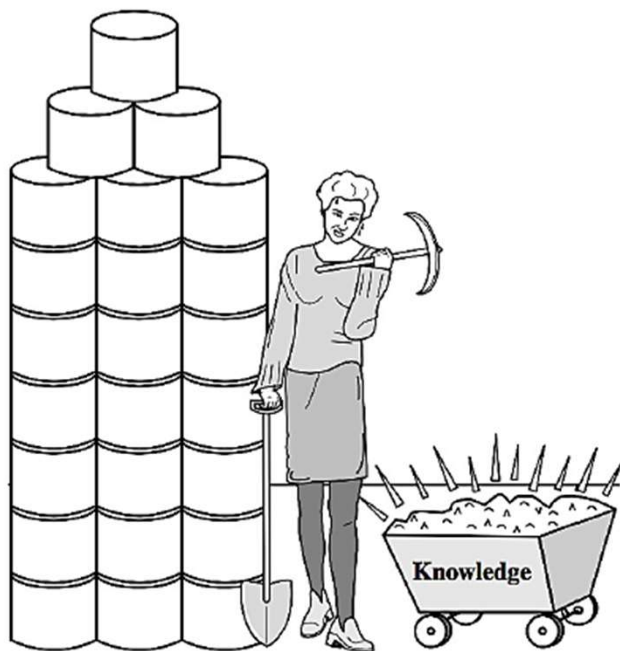
کاوش مجموعه داده های بزرگ
معمولا اطلاعات نهفته ای در داده ها وجود دارند که تاکنون آشکار نشده اند.
برای کشف اطلاعات مفید توسط انسانها هفته ها زمان نیاز است.
خیلی از داده ها هنوز تحلیل نشده اند.





تعریف داده کاوی

- استخراج اطلاعات مفید و ناشناخته از داده ها
- فرآیند استخراج دانش، الگوها و اطلاعات مفید از مجموعه‌های بزرگ داده به وسیله ابزارهای خودکار و نیمه خودکار
- استفاده از تکنیک‌های کارآمد برای تحلیل مجموعه‌های بسیار بزرگ داده و استخراج الگوهای مفید و احتمالاً غیرمنتظره در داده‌ها



تعریف یادگیری ماشین

Traditional Programming



Machine Learning



یادگیری ماشین شاخه‌ای از هوش مصنوعی است که هدف آن توسعه‌ی الگوریتم‌ها و مدل‌هایی است که به رایانه‌ها امکان می‌دهند از داده‌ها یاد بگیرند و بدون برنامه‌ریزی صریح، عملکرد خود را بهبود دهند.

به زبان ساده:

در برنامه‌نویسی سنتی، ما قواعد (قوانین) را به کامپیوتر می‌دهیم و او داده‌ها را پردازش می‌کند تا خروجی بدهد.

در یادگیری ماشین، ما داده‌ها و خروجی مورد انتظار را به کامپیوتر می‌دهیم، و او خودش قانون یا الگو را کشف می‌کند.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

ارتباط یادگیری ماشین با داده کاوی

داده کاوی علمی است که به کشف الگوها، روابط و دانش پنهان در حجم عظیمی از داده ها می پردازد.
شباهت ها:

هر دو از الگوریتم ها و روش های آماری برای تحلیل داده ها استفاده می کنند.
هر دو به دنبال یافتن الگوهای مفید از داده ها هستند.

تفاوت ها:

داده کاوی بیشتر روی **کشف دانش و بینش از داده ها** تمرکز دارد (مثلاً یافتن الگوی خرید مشتریان).
یادگیری ماشین بیشتر روی **ساخت مدل های پیش بینی کننده یا تصمیم گیرنده** تمرکز دارد (مثلاً
پیش بینی اینکه مشتری چه محصولی را در آینده خواهد خرید).

به عبارت ساده:

داده کاوی = کشف دانش از داده ها

یادگیری ماشین = استفاده از داده ها برای یادگیری و پیش بینی آینده

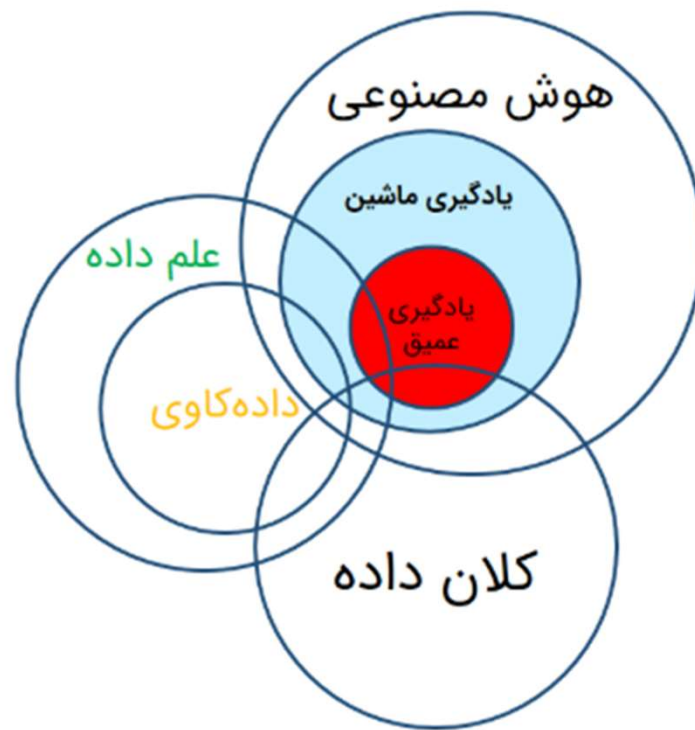
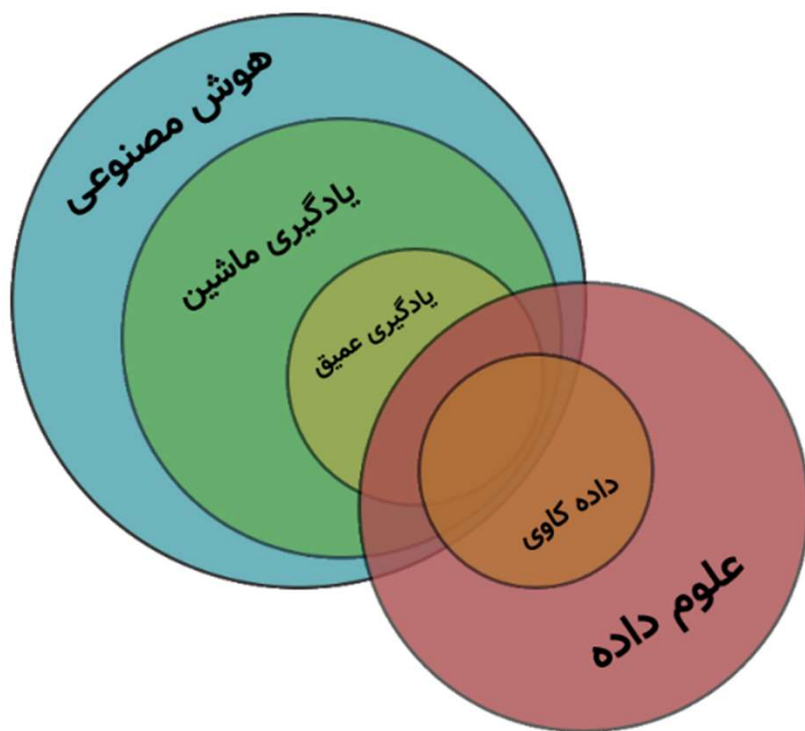


دانشگاه سمنان

دانشگاه سمنان
Semnan University

پردیس فرزندانگان

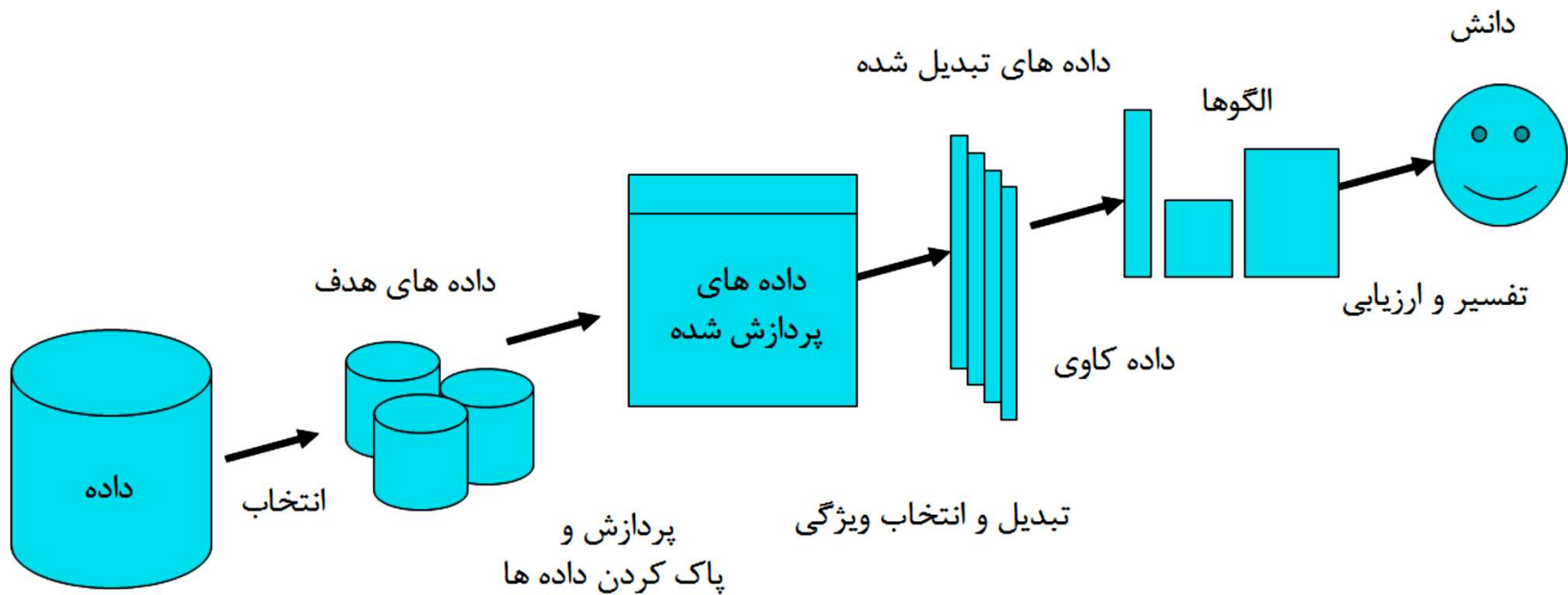
ارتباط یادگیری ماشین با داده کاوی





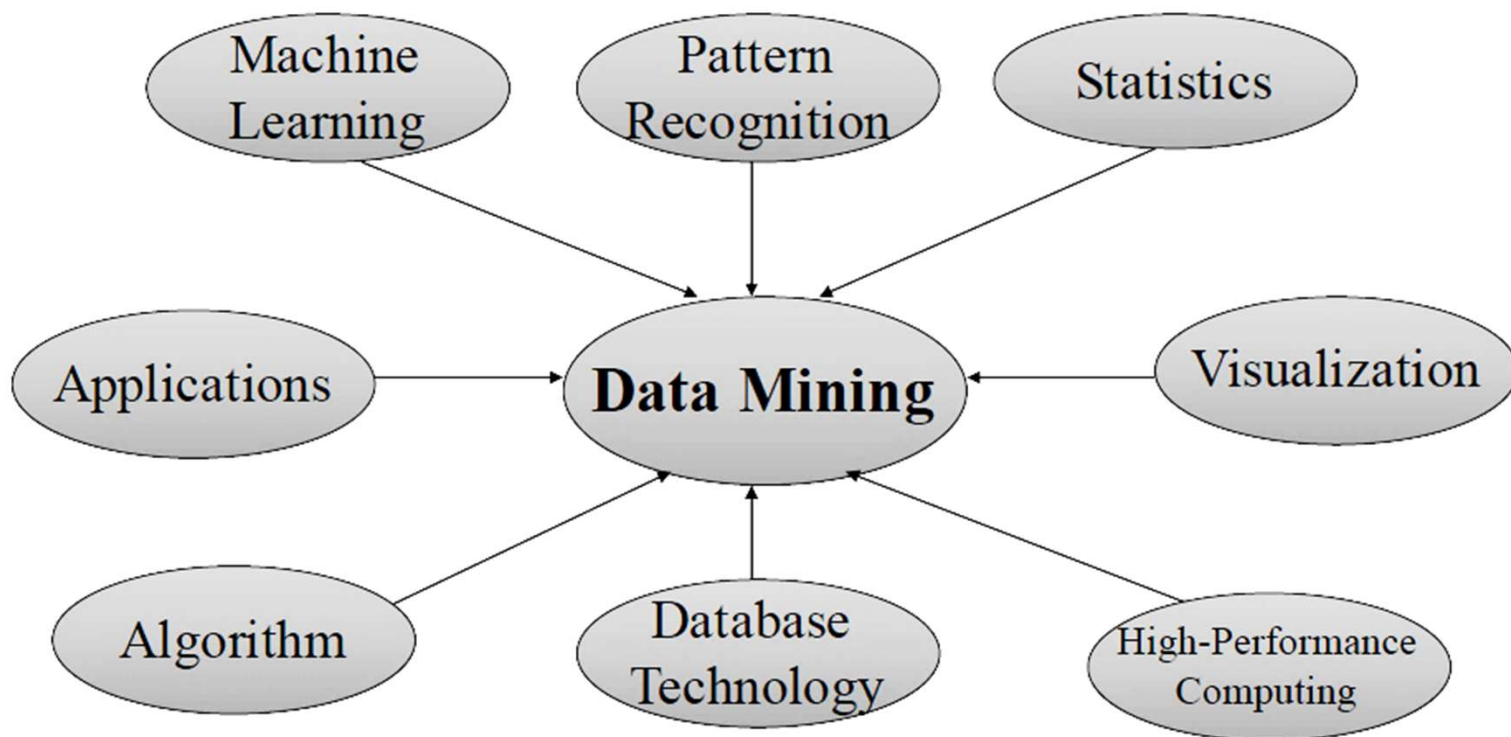
مراحل کشف دانش

داده کاوی نقش اساسی در فرآیند کشف دانش دارد.





داده کاوی: تلاقی چندین رشته



کاربردهای یادگیری ماشین

یادگیری ماشین تقریباً در همه‌ی حوزه‌های فناوری و زندگی روزمره حضور دارد.

۱. پردازش زبان طبیعی NLP و ترجمه ماشینی

- تبدیل گفتار به متن Speech to Text در موبایل‌ها.
- ترجمه متن‌ها با ابزارهایی مانند Google Translate
- چت‌بات‌ها و دستیارهای صوتی

۲. بینایی ماشین و پردازش تصویر

- تشخیص چهره برای باز کردن قفل گوشی.
- سیستم‌های نظارت تصویری هوشمند.
- تشخیص بیماری از روی تصاویر پزشکی (مانند X-Ray یا MRI).
- خودروهای خودران برای شناسایی علائم راهنمایی و موانع



کاربردهای یادگیری ماشین

۳. سیستم‌های توصیه‌گر

- پیشنهاد فیلم در Netflix یا Filimo
- پیشنهاد محصول در دیجی کالا یا آمازون.
- پیشنهاد آهنگ در Spotify یا Radio Javan

۴. مالی و بانکی

- تشخیص تراکنش‌های مشکوک و جلوگیری از کلاهبرداری.
- تحلیل ریسک و اعتبارسنجی مشتریان.
- معاملات الگوریتمی در بازار بورس.

کاربردهای یادگیری ماشین

۵. پزشکی و سلامت

- پیش‌بینی احتمال بیماری با توجه به داده‌های پزشکی.
- تحلیل ژنوم و داده‌های زیستی.
- طراحی دارو با کمک مدل‌های یادگیری ماشین.

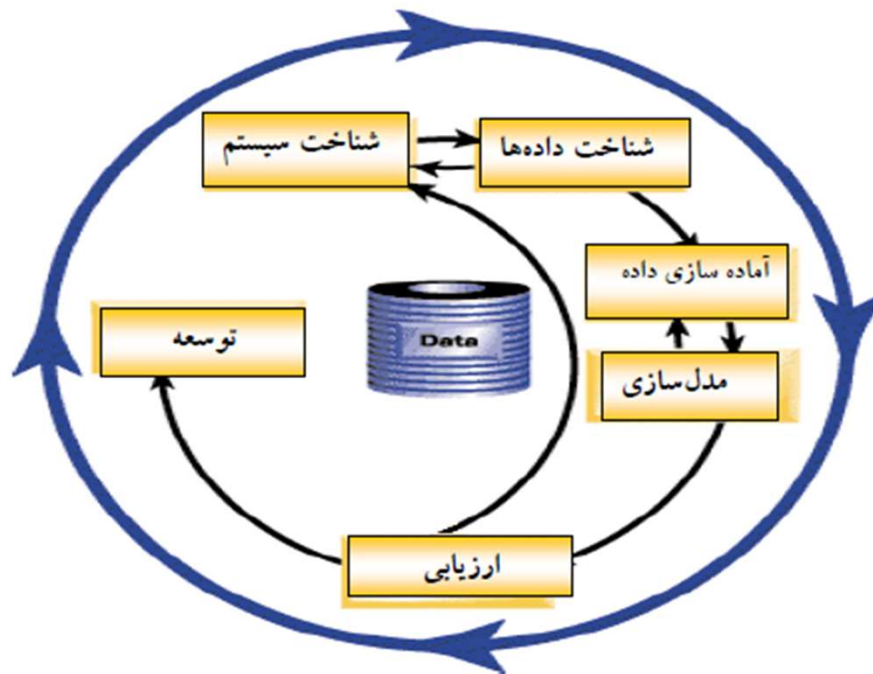
۶. صنعت و اینترنت اشیاء IoT

- نگهداری پیشگیرانه در کارخانه‌ها.
- کنترل هوشمند انرژی در ساختمان‌ها.
- تحلیل داده‌های حسگرها برای پیش‌بینی خطا.



متدلوژی CRISP

CRoss Industry Standard Process for Data Mining



فازها:

- شناخت سیستم
- شناخت داده ها
- آماده سازی داده ها
- مدلسازی
- ارزیابی
- توسعه

هر کدام از این فازها به زیربخش هایی تقسیم می شوند:



مراحل انجام پروژه یادگیری ماشین



ابزارها و زبان های برنامه نویسی داده کاوی



مهمترین کتابخانه های پایتون برای یادگیری ماشین





داده چیست؟

مجموعه ای از اشیاء داده‌ای و ویژگی‌های آنها

صفت یک ویژگی یا خصیصه یک شی است.

مثال: رنگ چشم یک فرد، درجه حرارت و غیره.

ویژگی به عنوان متغیر **variable**، میدان **field**، مشخصه **characteristic**

یا ویژگی **feature** نیز شناخته می‌شود.

مجموعه ای از ویژگی‌ها یک شی را توصیف می‌کند.

شی نیز به عنوان رکورد **record**، نقطه **point**، مورد **case**، نمونه **sample**،

موجودیت **entity** یا نمونه موردی **instance** شناخته می‌شود.

ویژگی **Attributes**

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

شیء **Objects**



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزنانگان

انواع صفات و ویژگی‌ها

انواع مختلفی از ویژگی‌ها وجود دارند:

✓ ویژگی‌های دسته‌ای **Categorical** (این ویژگی‌ها کیفی هستند)

اسمی **Nominal**: چنانچه دسته‌هایی که متغیر در آنها قرار می‌گیرد دارای هیچ گونه ترتیب طبیعی نباشد، هر یک از آن متغیرها را متغیر صوری یا اسمی کیفی می‌نامند. مثال: شماره شناسه، رنگ چشم، رنگ مو، وضعیت تاهل، کدهای پستی.

ترتیبی **Ordinal**: در یک جهت معنادار ارزش دارند (رتبه بندی) اما مقدار بین مقادیر متوالی شناخته شده نیست. بین مثال: رتبه‌بندی‌ها (مانند خوب، متوسط، بد یا مثلاً طعم چیپس سیب‌زمینی در مقیاس از ۱-۱۰)، قد در دسته‌بندی‌های {بلند، متوسط، کوتاه}

✓ ویژگی‌های عددی **Numeric** (این ویژگی‌ها کمیتی هستند)

فاصله زمانی **Interval**: اندازه‌گیری در یک مقیاس از واحد‌های هم اندازه، مثال‌ها: تاریخ‌های تقویم، درجه حرارت بر حسب سانتی‌گراد یا فارنهایت.

نسبت **Ratio** مثال‌ها: دما بر حسب کلوین، طول، زمان، شمارش

✓ مورد خاص: ویژگی‌های **Binary** (بله/خیر، وجود دارد/وجود ندارد)

مقادیر صفات و ویژگی‌ها

مقادیر صفات، اعداد یا نمادهایی هستند که به یک ویژگی اختصاص داده شده‌اند.

تمایز بین صفات و مقادیر ویژگی

- همان ویژگی را می‌توان به مقادیر مشخصه‌های مختلف نگاشت کرد.

مثال: ارتفاع را می‌توان با فوت یا متر اندازه‌گیری کرد

- ویژگی‌های مختلف را می‌توان به یک مجموعه از مقادیر نگاشت کرد.

مثال: مقادیر مشخصه برای ID و age اعداد صحیح هستند. اما خصوصیات مقادیر مشخصه می‌تواند متفاوت باشد. شناسه محدودیت ندارد اما سن دارای حداکثر و حداقل مقدار است

خصوصیات مقادیر صفات

نوع یک صفت، بستگی به خواص زیر دارد:

تمايز Distinctness: $\neq$

سفارش Order: $< >$

اضافه Addition: $+$

ضرب Multiplication: $*$ /

- صفت اسمی: تمايز

- صفت ترتیبی: تمايز و نظم

- ویژگی فاصله: تمايز، ترتیب و اضافه

- ویژگی نسبت: هر ۴ ویژگی

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=$, $\neq$)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. ($<$, $>$)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ($+$, $-$)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	For ratio variables, both differences and ratios are meaningful. ($*$, $/$)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation



دانشگاه سمنان

دانشگاه سمنان
Semnan University

پرورش فرزندان

صفات گسسته و پیوسته

صفت گسسته

- فقط مجموعه ای از مقادیر متناهی یا نامحدود دارد. مثال‌ها: کدهای پستی، شمارش یا مجموعه‌ای از کلمات در مجموعه‌ای از اسناد
- اغلب به عنوان متغیرهای عدد صحیح نشان داده می‌شود.
- توجه: ویژگی‌های باینری یک مورد خاص از ویژگی‌های گسسته هستند.

صفت پیوسته

- دارای اعداد واقعی به عنوان مقادیر ویژگی. مثال‌ها: دما، قد یا وزن.
- در عمل، مقادیر واقعی را می‌توان با استفاده از تعداد محدودی از ارقام اندازه‌گیری و نمایش داد.
- ویژگی‌های پیوسته معمولاً به عنوان متغیرهای ممیز شناور نشان داده می‌شوند.



انواع مجموعه‌های داده datasets

داده‌های مرتب (Ordered)

- داده‌های ویدئویی: دنباله‌ای از تصاویر
- داده‌های وابسته به زمان Temporal: سری زمانی
- داده‌های متوالی Sequential: توالی تراکنش
- داده‌های توالی ژنتیکی
- داده‌های فضایی (Spatial)، تصویری و چند رسانه‌ای:
- داده‌های مکانی: نقشه‌ها
- داده‌های تصویری
- داده‌های ویدیویی

رکوردها Record:

- رکوردهای رابطه‌ای
- ماتریسی از داده‌ها مانند ماتریس عددی
- اسناد داده‌ای مانند متن‌ها
- داده تراکنش Transaction
- گراف یا شبکه Graph:
- شبکه جهانی وب
- شبکه‌های اجتماعی یا اطلاعاتی
- ساختارهای مولکولی

داده رکورد

داده‌ای که شامل مجموعه‌ای از رکوردها است، که هر کدام از آن‌ها شامل یک مجموعه ثابت از ویژگی‌ها می‌باشد.

<i>Tid</i>	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



ماتریس داده

اگر اشیاء داده‌ای دارای مجموعه ثابتی از ویژگی‌های عددی باشند، آنگاه می‌توان این اشیاء داده‌ای را به عنوان نقاطی در یک فضای چند بعدی در نظر گرفت که هر بعد نشان‌دهنده یک ویژگی متمایز است.

چنین مجموعه داده‌ای می‌تواند توسط یک ماتریس m در n نمایش داده شود، جایی که m تعداد ردیف‌ها (هر ردیف نشان‌دهنده یک شیء) و n تعداد ستون‌ها (هر ستون نشان‌دهنده یک ویژگی) است.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

داده سند

هر سند به یک بردار ترم term تبدیل می شود، هر عبارت جزء (ویژگی) بردار است، مقدار هر جزء تعداد دفعاتی است که عبارت مربوطه در سند رخ می دهد.

نمایش Bag of word

	season	timeout	lost	win	game	score	ball	play	coach	team
Document 1	2	0	2	0	6	2	0	5	0	3
Document 2	0	0	3	0	0	1	2	0	7	0
Document 3	0	3	0	2	2	1	0	0	1	0



دانشگاه سمنان

دانشگاه سمنان
Semnan University

پردیس فرزندانگان

داده تراکنش

نوع خاصی از داده های رکورد، که در آن:

- هر رکورد (معامله) شامل مجموعه ای از آیتم است. مجموعه ای از آیتم ها را می توان به عنوان یک بردار باینری نیز نمایش داد که در آن هر ویژگی یک آیتم است.
- به عنوان مثال، یک فروشگاه مواد غذایی را در نظر بگیرید. مجموعه ای از محصولات خریداری شده توسط یک مشتری در طی یک سفر خرید، یک معامله را تشکیل می دهد، در حالی که تک تک محصولات خریداری شده، اقلام هستند.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

داده مرتب

سری زمانی

دنباله ای از مقادیر عددی مرتب شده (در "زمان").



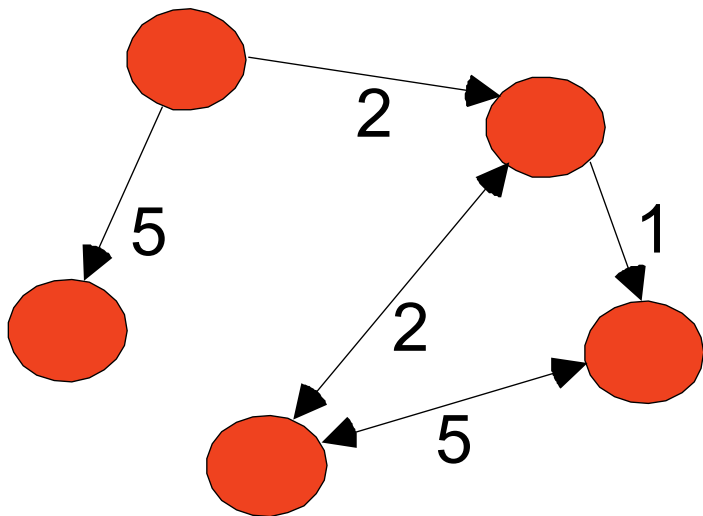
داده مرتب

- داده های توالی ژنومی
- داده یک رشته مرتب شده طولانی است

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

داده گراف یا شبکه

مثال: نمودار وب و پیوندهای HTML

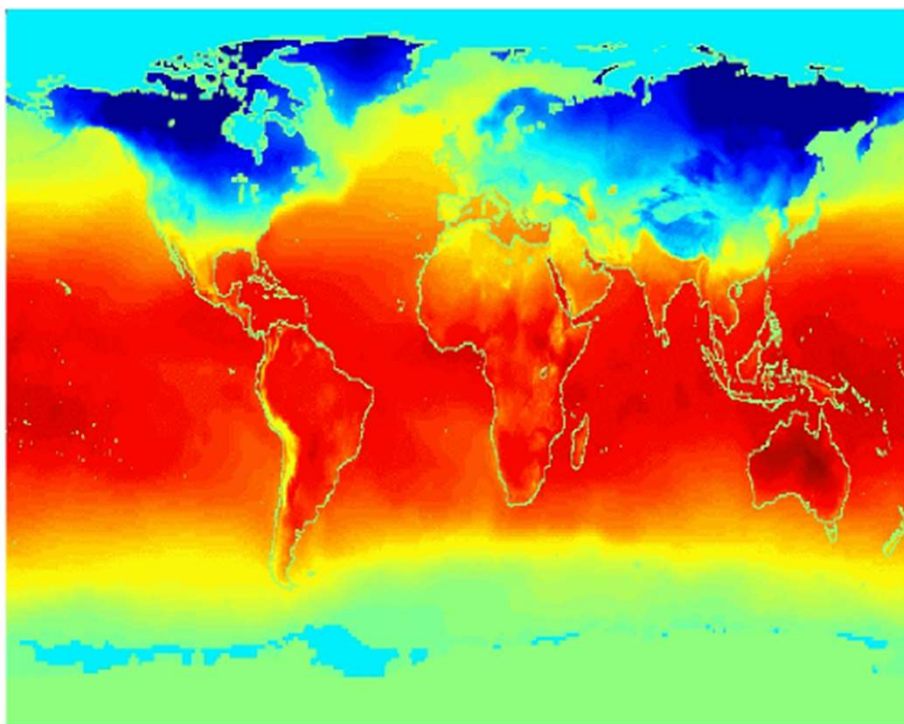


```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

داده‌های فضایی و چند رسانه‌ای

میانگین دمای ماهانه خشکی و اقیانوس

Jan



توصیف آماری داده‌ها

برای خلاصه سازی داده ها یا بدست آوردن یک الگوی خاص از روابط آماری استفاده می شود.
دو نوع شاخص آماری وجود دارد:

- شاخص مرکزی

میانگین ، میانه ، مد

- شاخص پراکندگی

دامنه تغییرات و میانگین انحرافات ، انحراف چارکی ، واریانس و انحراف استاندارد

میانگین

یکی از معروف ترین شاخص های مرکزی میانگین است، که انواع آن عبارت است از: میانگین حسابی ، میانگین وزن دار ، میانگین هندسی و میانگین هارمونیک.

$$\mu_A = \frac{x_1 + x_2 + \dots + x_n}{n} = 1/n \sum_{i=1}^n x_i$$

میانگین حسابی

چنانچه داده ها دارای وزن یکسانی (تأثیر یکسانی) نباشند، برای محاسبه ی میانگین حسابی آن باید هر یک از آنها را در وزن خود ضرب و حاصل جمع نهایی را تقسیم بر مجموع وزن ها کنیم. این میانگین به میانگین وزنی نیز شناخته می شود.

$$\mu_W = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n W_i}$$

میانگین وزن دار

میانگین

$$\mu_G = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n} = \left(\prod_{i=1}^n x_i \right)^{1/n}$$

میانگین هندسی

از میانگین هندسی موقعی

استفاده می‌شود که صحبت از نرخ رشد یک ویژگی مطرح باشد.

برای مثال جهت محاسبه‌ی نرخ رشد جمعیت، نرخ رشد سود و نرخ رشد تولید به این نوع میانگین رجوع می‌شود.

میانگین هارمونیک عبارت است از تقسیم تعداد داده‌ها بر مجموع معکوس مقادیر موجود در داده‌ها

$$\mu_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n 1/x_i}$$

بطور معمول هنگامی که واحد اندازه‌گیری داده‌ها ترکیبی باشد، از این نوع میانگین استفاده می‌شود. برای مثال اگر یک وسیله‌ی نقلیه فاصله‌ی میان دو شهر را در مرحله‌ی رفت با یک سرعت و در برگشت با سرعت دیگری طی کند، برای محاسبه‌ی میانگین سرعت آن از میانگین هارمونیک استفاده می‌کنیم.

$$\mu_H \leq \mu_G \leq \mu_A$$



میانه

میانه عددی است که توزیع داده ها را به دو قسمت مساوی تقسیم می کند، به نحوی که نیمی از داده ها بزرگتر و نیم دیگر کوچکتر از آن هستند.

در مجموعه اعداد مرتب شده داده ی میانی به عنوان میانه لحاظ می گردد. در صورتی که تعداد داده ها زوج باشد، نصف مجموع دو داده ای که در وسط قرار دارند، میانه محسوب می شود.

میانه برای داده های پیوسته ی دسته بندی شده از فرمول زیر محاسبه می شود :

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$



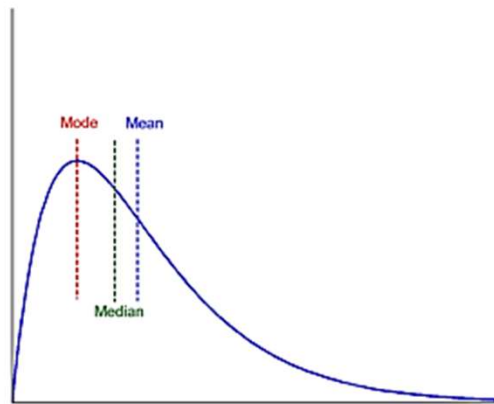
دانشگاه سمنان

دانشگاه سمنان
Semnan University

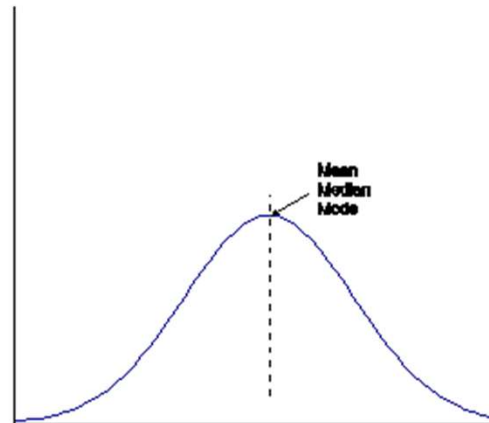
پروفسور فرزانه گان

مد

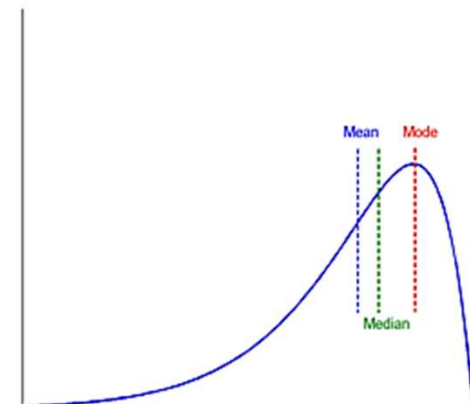
داده ای که فراوانی آن از سایر داده ها بیشتر باشد را نما یا مد می نامیم. اگر دو داده ی مجاور دارای بیشترین فراوانی باشند، نصف مجموع آن ها به عنوان مد انتخاب می شود، در غیر این صورت اگر دو داده مجاور نباشند هر دو به عنوان مد انتخاب می شوند.



کجی مثبت



توزیع متقارن

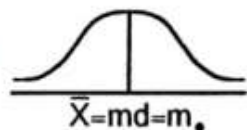


کجی منفی



۱- اگر میانگین و میانه و نما با هم برابر باشد، توزیع متقارن است و کجی آن صفر است.

$$\bar{X} = md = m_o$$

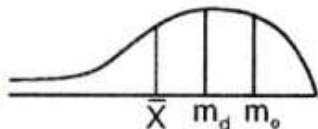


۲- اگر میانگین بزرگتر از میانه و میانه بزرگتر از نما باشد کجی مثبت است، یعنی اکثر نمرات پایین بوده و امتحان سخت است.

$$\bar{X} > md > m_o$$

۳- اگر میانگین کوچکتر از میانه و میانه کوچکتر از نما باشد، کجی منفی است و اکثر نمرات بالاست و امتحان آسان بوده است.

$$\bar{X} < md < m_o$$



شاخص های پراکندگی

دامنه تغییرات : تفاوت میان بزرگترین و کوچکترین داده را به عنوان دامنه ی تغییرات داده ها می شناسیم.

$$R = \text{Max}(x_1, x_2, \dots, x_n) - \text{Min}(x_1, x_2, \dots, x_n)$$

میانگین انحرافات : به منظور دخالت تاثیر تمام داده ها می توان فاصله ی کلیه داده ها با میانگین را محاسبه نمود، که به آن انحراف از میانگین داده ها گوییم.

$$D = \frac{1}{n} \sum_{i=1}^n |x_i - \mu_A|$$

شاخص های پراکندگی (ادامه)

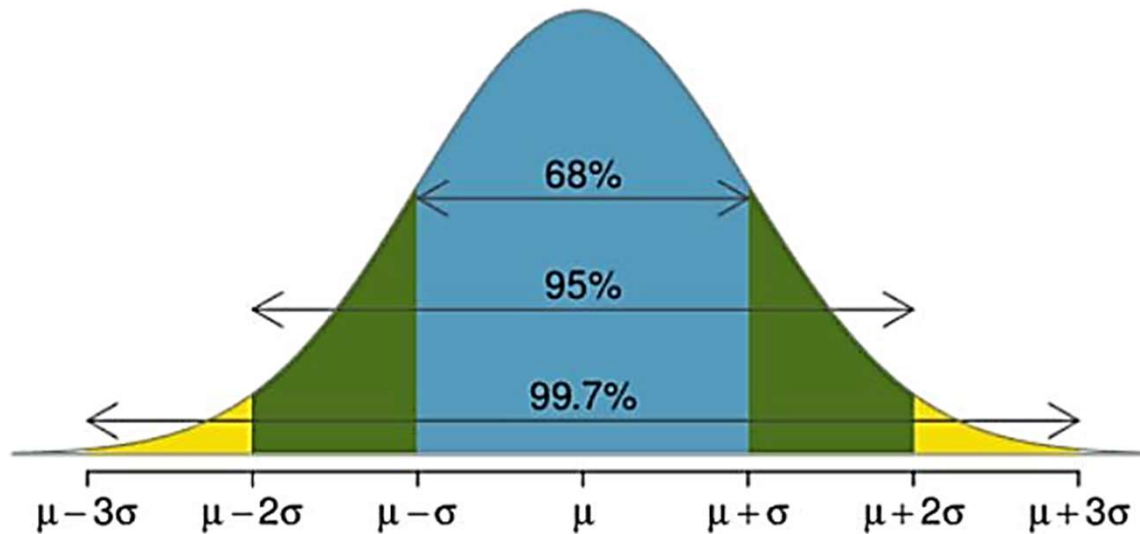
واریانس : میانگین مجذور انحرافات را واریانس گویند.

$$var(x) = \frac{1}{m} \sum_{i=1}^m (x - \bar{x})^2$$

انحراف استاندارد : با جذر گرفتن از مقدار واریانس انحراف استاندارد بدست می آید.

$$\sigma(x) = \sqrt{var(x)}$$

توزیع نرمال



$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

منحنی نرمال (توزیع). (μ : میانگین، σ : انحراف معیار)
 از $\mu - \sigma$ تا $\mu + \sigma$: شامل حدود ۶۸ درصد از اندازه گیری‌ها
 از $\mu - 2\sigma$ تا $\mu + 2\sigma$: شامل حدود ۹۵ درصد از آن است.
 از $\mu - 3\sigma$ تا $\mu + 3\sigma$: حدود ۹۹.۷ درصد آن را شامل می‌شود.



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزندان

کواریانس

اندازه ی تغییرات هماهنگ دو متغیر تصادفی را کواریانس گویند. اگر دو متغیر یکی باشند، کواریانس برابر با واریانس خواهد بود. برای مقادیر ویژگی های x و y مقدار کواریانس برابر است با :

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

اگر این مقدار صفر باشد، میان دو ویژگی همبستگی وجود ندارد و دو ویژگی دارای رابطه خطی نیستند. مقدار مثبت آن نشان می دهد با افزایش یکی، دیگری نیز افزایش می یابد و مقدار منفی دلالت بر این دارد که افزایش یکی باعث کاهش دیگری می شود.

برای داده ها با d بعد، یک ماتریس $d \times d$ به نام ماتریس کواریانس ویژگی داریم که:

$$S_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

همبستگی

همبستگی به رابطه بین دو متغیر اشاره می کند، که می توان آن را با کمک نمودار پراکندگی نشان داد. ارزش مقداری آن را ضریب همبستگی می نامیم.

$$cor(x, y) = \frac{cov(x, y)}{\sqrt{cov(x, x)cov(y, y)}} = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

ضریب همبستگی دارای دامنه ای بین ۱ و -۱ است.

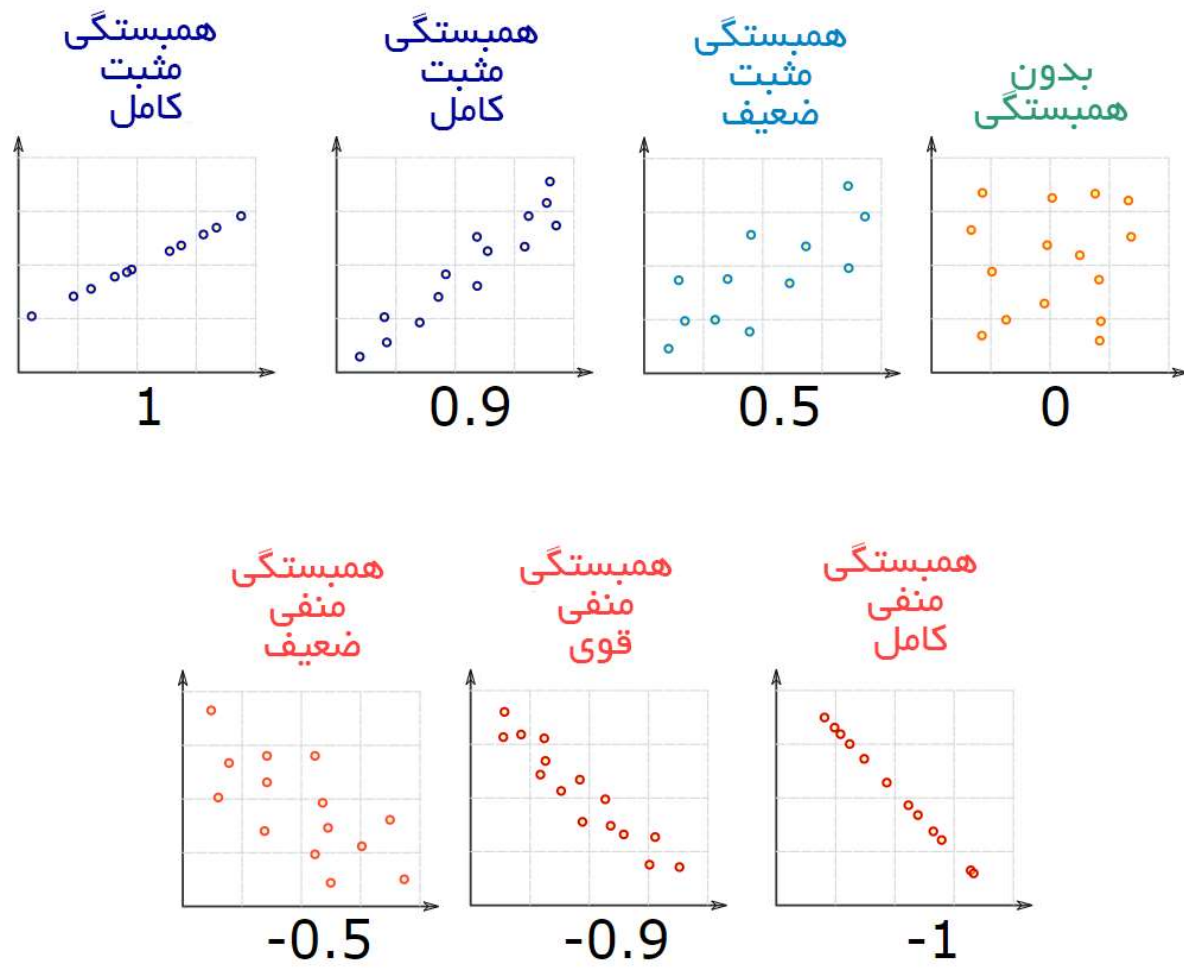
مقدار صفر عدم همبستگی را نشان می دهد و مقادیر ۱ و -۱ به ترتیب دلالت بر همبستگی کامل مثبت و همبستگی کامل منفی برای دو ویژگی فوق را دارد.



دانشگاه سمنان

دانشگاه سمنان
Semnan University

پروفسور فرزانه گان





دانشگاه سمنان

دانشگاه سمنان

Semnan University

پروفسور فرزانه گان

شباهت و عدم شباهت

برای بسیاری از مسائل مختلف، نیاز داریم که میزان نزدیکی دو شیء را اندازه گیری کنیم.
مثال‌ها:

برای یک آیتم خریداری شده توسط مشتری، یافتن آیتم‌های مشابه دیگر.
گروه‌بندی مشتریان یک سایت به‌طوری که مشتریان مشابه، تبلیغات یکسانی را مشاهده کنند.
گروه‌بندی اسناد وب به‌طوری که بتوانید اسنادی که درباره سیاست صحبت می‌کنند را از آن‌هایی که درباره ورزش صحبت می‌کنند، جدا کنید.
یافتن تمام اسناد وب نزدیک به هم که تقریباً تکراری هستند.
برای حل این مسائل، نیاز به یک تعریف از شباهت یا فاصله داریم. این تعریف به نوع داده‌هایی که داریم بستگی دارد.

شباهت

اندازه‌گیری عددی برای بیان میزان مشابهت دو شیء داده‌ای.
مقدار بالاتر نشان‌دهنده شباهت بیشتر بین اشیاء است.
اغلب در محدوده $[0, 1]$ قرار دارد.

عدم شباهت (برای مثال، فاصله)

اندازه‌گیری عددی برای بیان میزان تفاوت دو شیء داده‌ای.
مقدار کمتر نشان‌دهنده شباهت بیشتر بین اشیاء است.
حداقل عدم شباهت معمولاً ۰ است.



ماتریس داده و ماتریس عدم شباهت

Proximity (مجاورت): به شباهت یا عدم شباهت اشاره دارد.

اندازه‌گیری‌های شباهت اغلب می‌توانند به‌عنوان تابعی از اندازه‌گیری‌های عدم شباهت بیان شوند. $sim(i, j) = 1 - d(i, j)$ برای مثال، برای داده‌های اسمی:

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

ماتریس داده:

n نقطه داده با ابعاد p

دارای دو حالت است: سطرها برای اشیا و ستون‌ها برای ویژگی‌ها

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

ماتریس عدم شباهت:

n نقطه داده اما تنها فاصله درج می‌شود.

یک ماتریس مثلثی است.



معیار شباهت برای ویژگی های اسمی

روش تطبیق ساده: عدم شباهت بین دو شی i و j را می توان بر اساس نسبت عدم تطابق محاسبه کرد.

$$d(i, j) = \frac{p - m}{p}$$

m : تعداد موارد منطبق، p : تعداد کل متغیرها

به همین ترتیب، شباهت می تواند به صورت زیر محاسبه شود:

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}$$

مثال:

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

فرض کنید که داده های نمونه جدول مقابل را داریم، با این تفاوت که فقط

شناسه شی و ویژگی test-1 موجود است و test-1 اسمی است.

از آنجایی که در اینجا یک ویژگی اسمی داریم، test-1، $p=1$ را تنظیم می کنیم.

$$\begin{bmatrix} 0 & & & \\ d(2, 1) & 0 & & \\ d(3, 1) & d(3, 2) & 0 & \\ d(4, 1) & d(4, 2) & d(4, 3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

از ماتریس مقابل می بینیم که همه اشیاء غیر مشابه هستند

به جز اشیاء ۱ و ۴



دانشگاه سمنان

دانشگاه سمنان

Semnan University

پردیس فرزانهگان

معیار شباهت برای ویژگی های باینری

یک ویژگی باینری تنها یکی از دو حالت را دارد: ۰ و ۱، که در آن ۰ به معنای وجود ندارد و ۱ به معنای وجود آن است.

اگر تصور شود که همه صفات دودویی دارای وزن یکسانی هستند، جدول احتمالی را خواهیم داشت که در آن:

جدول احتمال برای ویژگی های باینری

		Object j		
		1	0	sum
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
	sum	$q + s$	$r + t$	p

q تعداد صفاتی است که برای هر دو شی i و j برابر با ۱ است.

r تعداد صفاتی است که برای شی i برابر با ۱ و برای شی j برابر با ۰ است.

s تعداد صفاتی است که برای شی i برابر با ۰ و برای شی j برابر با ۱ است.

t تعداد صفاتی است که برای هر دو شی i و j برابر با ۰ است.

تعداد کل صفات p است که $p = q + r + s + t$

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

اندازه گیری فاصله برای متغیرهای دودویی متقارن

برای ویژگی های باینری نامتقارن، این دو حالت به یک اندازه مهم نیستند، مانند نتایج مثبت (۱) و منفی (۰) یک آزمایش بیماری.

$$d(i, j) = \frac{r + s}{q + r + s}$$

اندازه گیری فاصله برای متغیرهای دودویی نامتقارن

زمانی که هر دو یک باشند مهمتر است از زمانی که هر دو صفر باشند.

بنابراین t بی اهمیت در نظر گرفته می شود.



معیار شباهت برای ویژگی های عددی

اندازه گیری های فاصله که معمولاً برای محاسبه معیار شباهت اشیاء توصیف شده توسط ویژگی های عددی استفاده می شوند، شامل فواصل اقلیدسی، منهتن و مینکوفسکی می باشند.

محبوب ترین اندازه گیری فاصله، فاصله اقلیدسی است. فرض کنید $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ و $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ دو شیء باشند که با صفات عددی p توصیف می شوند. فاصله اقلیدسی بین اجسام i و j به صورت تعریف می شود:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}.$$

یکی دیگر از معیارهای معروف فاصله، منهتن Manhattan است که به صورت زیر تعریف می شود:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|.$$

فاصله مینکوفسکی تعمیم فاصله اقلیدسی و منهتن است:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

به این فاصله نرم $L-h$ نیز گفته می شود. اگر $h=2$ باشد فاصله اقلیدسی یا نرم $L-2$ و اگر $h=1$ باشد فاصله منهتن یا نرم $L-1$ خواهد بود.

اگر $h \rightarrow \infty$ به آن فاصله supremum گفته می شود.

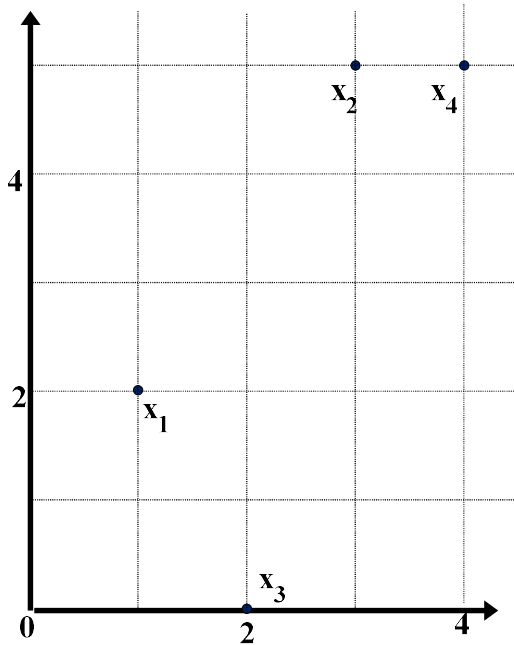
$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

برای محاسبه کافی است که بزرگترین اختلاف در ویژگی ها را در نظر بگیریم.



مثال

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

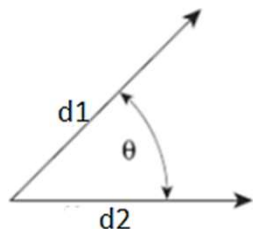
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0



فاصله کسینوسی

اگر d_1 و d_2 دو بردار باشند، فاصله کسینوسی آنها مانند زیر بدست می آید:

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / ||d_1|| ||d_2||$$

که در آن $\cdot$ ضرب نقطه ای بین دو بردار و $||d||$ طول بردار را نشان می دهد.

شباهت کسینوسی معمولا برای مقایسه اسناد که بردارها بر اساس طول سند نرمال سازی شده اند، استفاده می شود.

مثال:

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \cdot d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$||d_1|| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$



فاصله همینگ

فاصله همینگ تعداد موقعیت‌هایی است که بردارهای بیتی در آن‌ها تفاوت دارند.
مثال:

$$p1 = 10101$$

$$p2 = 10011$$

$d(p1, p2) = 2$ به دلیل اینکه بردارهای بیتی در موقعیت‌های ۳ و ۴ با هم متفاوت‌اند.

نرم L1 برای بردارهای باینری

فاصله همینگ بین دو بردار از ویژگی‌های دسته‌بندی شده، تعداد موقعیت‌هایی است که در آن‌ها تفاوت دارند.
مثال:

$$\begin{aligned} x &= (\text{married}, \text{low income}, \text{cheat}), \\ y &= (\text{single}, \text{low income}, \text{not cheat}) \\ d(x, y) &= 2 \end{aligned}$$