

Ridandance & Duplicate ۲۵ ماتوچی دادهها را جو در مجموعه داری نشاند این است، آنها تکراری دارند همچنان که پاسخ می‌آورند. مثلاً مکان تقدیم کسی و آن کسی کجا بوده باز اینها هم در این ماتچی است. اینها باید مینیمیز شوند و اینرا دارند.

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
 - Major issue when merging data from heterogeneous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues

92,07,8 Data mining ۱۴
Data Preprocessing = 26,۰۷,۸
جنبش دارد که در زیر آمده است.

Data Preprocessing

- 1- ● Aggregation
- 2- ● Sampling
- 3- ● Dimensionality Reduction
- 4- ● Feature subset selection
- 5- ● Feature creation
- 6- ● Discretization and Binarization
- 7- ● Attribute Transformation

Aggregation چنان که از اسنیت سبقتی است همین ازان Data Reduction است یا تغییر اسکیل (Scale) اینها را برای در داده است. در واقع Aggregation طریق ازان همچویی داده که ازان را بهم جمع می‌نماید: ممکن خواهد بود معمولی را در نزدیکی مطابق باشیست، لیکن این اسنیت اسلامیان را به خاطر اینه که همچویی است؟! فصل برق خانی، فصرف برق، فعایر، فصرف برق، فصنایع، خیر، ازان، فعایر، فصرف برق، راهنمایی و دعوه، آنرا اینها را برق نماید.

بررسی کنیم باید سری جزء معتبر باشیم و سری غافی نباشد که در این صورت همان کلیل است روش های معناداری است و دوستی با توجه به طبقه داداریم و مخصوصاً هم صراحت با رای اسارت همان هم بررسی کامل اینجا نیست بلکه اینجا در مورد اینها تولید شده است اینها را فنون اهم برآورده کنیم همچنان که این صورت ایجاد می شود اینها اینها خوب و بدند و درین میان را با جام می دویم اینجا هم بعد از آن که در داده های اینجا کلیل است بررسی کنیم اینجا مطلع شدند اما اینجا داده های اینجا معتبر نیستند برای این داده های اینجا اینجا معتبر نیستند که اینجا طبقه ایجاد شده است.

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose

- Data reduction
 - ◆ Reduce the number of attributes or objects
 - Change of scale
 - ◆ Cities aggregated into regions, states, countries, etc
 - More “stable” data
 - ◆ Aggregated data tends to have less variability

© Tan, Steinbach, Kumar

Introduction to Data Mining

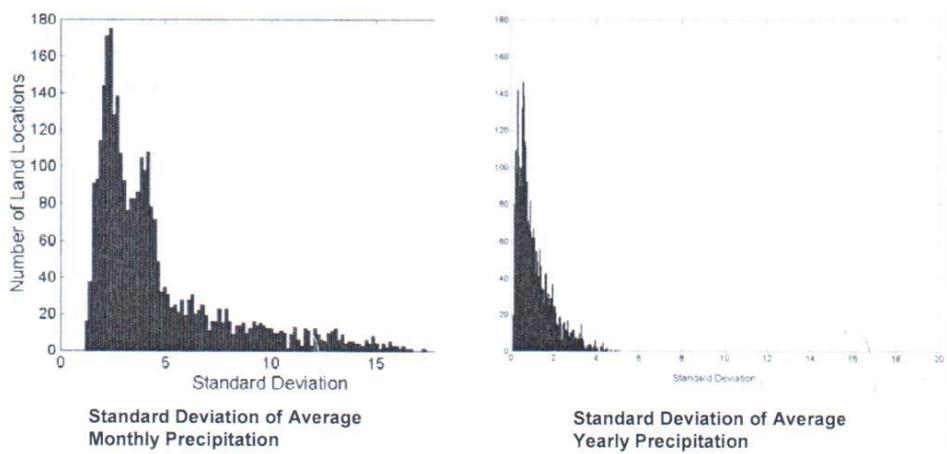
4/18/2004

27

فنا می‌کنند اینجا زد و داده های مال استرداد است، آنچه جو شاید نتواند خود را در بر بگیرد باید باید بگیرد.

Aggregation

Variation of Precipitation in Australia



© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004

28

Sampling نماینده جمع خالی خالی مجهود را مارست، لیکن این است که این داده ها درست جمیع آن ورود مانند - توانند بسیم دلخواه مان بریم، بخطه محسن مارسون مکنند و داده ها را مقدور جمیع آن را کنیم لطفاً (قد طرفان بالاتر بروند) و هم هزینه هاں پائیں ترباید. Sampling در مقابل طریق sensize طریق درود، sensize یعنی سرگواری، لفظی اگر می باشیم مل جانوچه را در نظر بگیریم و مقادیر کان را مطلع کنیم، از این طریق نیز همچنانی **Sampling** می باشد.

Sampling

- Sampling is the main technique employed for data selection.
 - It is often used for both the preliminary investigation of the data and the final data analysis.
 - Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
 - Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

consuming.

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004

29

نهاده ایام اخیر می‌گویند مخبر است قبلاً داریم مادرل عمر لادر را از افراد سریع می‌دانیم و بعد بیشتر نهاده برای راهنمایی در این ایام را در خواسته ایم،
نهاده هم لادر معاشر را بوزانم و خبر احوال آسمان پذیر می‌نمایم و این دلیل در معرفه کارهای عالیاتی نفعمن ترین راهنمایی داریم و همانطوری
نهاده هم می‌خواهد این را در معرفه خطاها یکی باشد برای این ایام ایجاد می‌نمایم که ممکن باشد تبلیغاتی نهاده ایام را در خواسته ایم.

Sampling ...

- The key principle for effective sampling is the following:
 - using a sample will work almost as well as using the entire data sets, if the sample is representative
 - A sample is representative if it has approximately the same property (of interest) as the original set of data

© Tan Steinbach, Kumar

Introduction to Data Mining

4/18/2024

88

بـطـرـاـسـ مـقـدـمـهـ وـبـعـلـسـ (شـمـلـهـ)ـ دـقـتـ. بـنـاطـلـهـ بـيـعـهـ دـادـهـ هـامـهـ نـجـاحـهـ دـفـقـهـ جـطـورـهـ يـعـتـصـمـ مـاـجـنـدـهـ دـلـفـوـتـهـ بـهـ رـاـيـانـ حـمـكـهـ: (ـآـلـهـ ـ3ـ)

<h2 style="text-align: center;">Types of Sampling</h2>	
<p>1- ● Simple Random Sampling (SRS)</p> <ul style="list-style-type: none"> - There is an equal probability of selecting any particular item 	<p>آمار داشت اموزش را درج کرد. با بر اساس محتوای این پیشواز</p> <p>نحوه بگیرید همه چهاره را خود سان را برای کنمند</p>
<p>2- ● Sampling without replacement (SWR)</p> <ul style="list-style-type: none"> - As each item is selected, it is removed from the population 	<p>محی آن از تعداد مدارس که داشتند</p>
<p>3- ● Sampling with replacement</p> <ul style="list-style-type: none"> - Objects are not removed from the population as they are selected for the sample. ◆ In sampling with replacement, the same object can be picked up more than once 	<p>آمدز انتخاب مکنند پر از سکاره داشت اکثر حسنه</p> <p>خوبی همچنان بر اساس چهاره حقیقت بجا هاست</p>
<p>4- ● Stratified sampling</p> <ul style="list-style-type: none"> - Split the data into several partitions; then draw random samples from each partition 	<p>SRS از زیرین آخوند استفاده می کنند لفظی بر جای این داشتند اکثر مدارس را به تصادف انتخاب کردند میانند درجه نسبتی مدارس مختلف اند و بنابراین هم شرکت در مدارس را بر اساس درجه نسبتی مدارس</p> <p>را بسته انتخاب می کنند درجه اس سریکاره ای اثام و هر دو در این فعالیت همچنان</p>

Sample Size

give without \rightarrow feel with
replacement

با جمله‌ای نزدیک است نمایند و جای بین این دو
نهایت مزبور معاصر که در کتاب آثار وجود دارد
مغقول معاشر عوضی مسفعه



بـ ٦٠٠٠ غـ نـ قـ حـ (ـ قـ تـ) (ـ اـ لـ رـ) (ـ دـ اـ رـ)

~~3000 points~~

2000 Points

500 Points

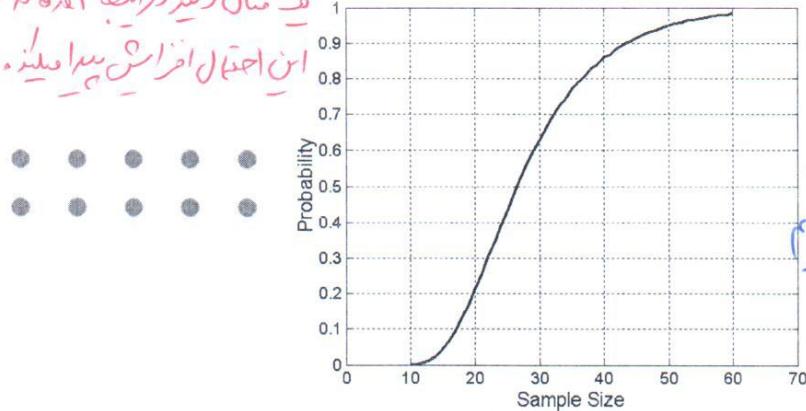
در نهضت پیری اید خانی ابعاد داشته باشند، خوب و قائم باشند تا راهنمایی را بفراراهم در مردم می‌نمایند و محتوا نیز نایاب باشند تا در آنها نظر نمایند.

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004

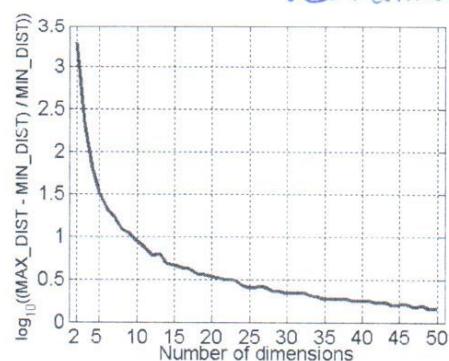
32



(مسئلہ 34): وہی دادھا کہ بندجا اور زرد، خلیٰ از قبیلہ ہارے ماں صفر ہو سکتے، لیکن بندجا مادرس خلیٰ بزرگ دارم، فتنہ ہے۔
دیکھیں اسی حالت میں چند ملینوں اضافہ درویشیں اسے، یعنی این حکیومت کے ^{aliquity} Curse of Dimensionality میں دارے گا، مادرس کے
ہاتھ میں سردار کارڈنل، ۲۵، ۳۰ اور دوستی کی جا ہائے کسی معادلے دارم، پسچا جا صفر است۔ Curse of Dimensionality

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
 - Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
 - Compute difference between max and min distance between any pair of points

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/200

34

دیمیشنالیتی کے نتائج میں ایک ایسا نتیجہ ہے کہ دادا میان از سیدارا ای رائٹر میا دریں ھیا یا چنیں ھما کھڑا پسہ۔ عملہ این مسئلہ
کو، ایسا کہ ای رائٹر نکلا سے کابل جمل نسے۔