5/30/2019 Chapter 2

Thursday, May 30, 2019 6:16 PM

Discrete Attributes vs Continuous Attributes

Discrete - integers, jumps in values
Continuous - real/double/float, smooth increase

Asymmetric attributes
- Only presence is regarded as important
    - Presence is a non-zero attr val
        - Words present in docs
        - Items present in cust transactions

Association Analysis

Asymmetric attrs typically arise from objects that are sets

Examples

ID #s
- Nominal

# of cylinders
- Nominal

Biased scale
- Interval or ratio?

The type of operations you choose should be "meaningful" for the type of data you have
- Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data

- The data type you see - often numbers or strings - may not capture all the properties or may suggest properties that are not there
- Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
- Many times, what is meaningful is measured by statistical significance
- But in the end, what is meaningful is measured by the domain

Data Quality

Poor data quality negatively affects many data processing efforts

Examples of data quality problems:
- Noise and outliers
- Missing values
- Duplicate data
- Full Data not available

Outliers
- Data objects with characteristics that are considerably different that most of the other data objects in the data set

Case 1

- Outliers are noise that interferes with data analysis

Case 2

- Outliers are the goal of our analysis

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
- Example:
  - One person with multiple email addresses

Similarity and Dissimilarity Measures

Similarity Measure

- Numerical measure of how alike two objects are, higher when more alike
- Often range of 0 to 1

Dissimilarity Measure

- Minimum dissimilarity is 0
- Upper limit varies

Proximity

- Refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

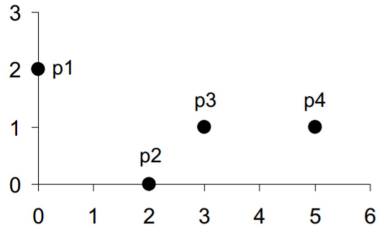
Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

Standardization is necessary, if scales differ

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

01/22/2018 Introduction to Data Mining, 2nd Edition 39

Density

- Measures the degree to which data objects are close to each other in a specified area
- Notion of desnity is closely related to that of proximity
- Concept of density is used for clustering and anomaly detection

  Examples
  - **Euclidean density**
    - Number of points per unit volume
  - **Probability Density**
    - Estimate what the distribution of the data looks like
  - **Graph-based density**
    - Connectivity

Aggregation

- Combining two or more attrs or objs into a single attr or obj
- Purpose is:
  - Data reduction
    - Reduce number of attrs or objs
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More "stable" data
    - Aggregated data tends to have less variability

Sampling

- Sampling is main technique employed for data reduction
  - Often used for both the preliminary investigation of data and final data analysis
- Statisticians often sample because **obtaining** entire set of data of interest is too expensive or time consuming
- Processing entire set of data of interest is too time consuming or expensive

Effective Sampling

- If the sample is **representative** it will work as well as using the entire data set
  - A sample is representative if it has approx the same props of interest as the original set of data

Simple Random Sampling

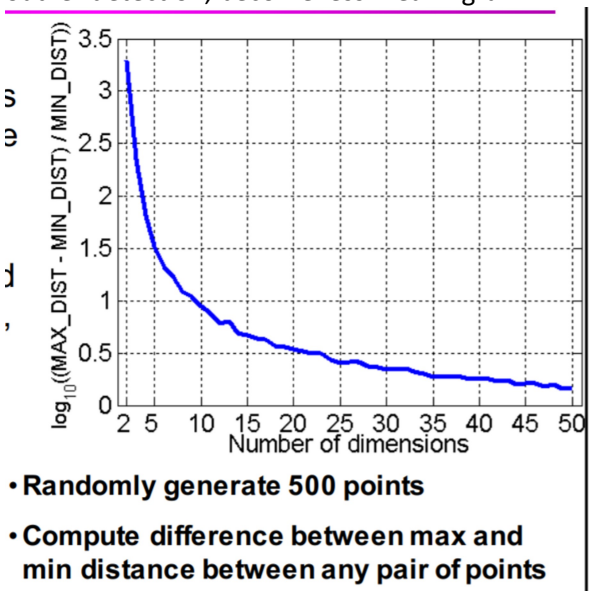
- There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from population
- Sampling with replacement
  - Objs are not removed from the population as they are selected for the sample.
  - In sampling with replacement, same obj can be picked up more than once

Stratified sampling

- Split the data into several partitions; then draw random samples from each partition

Curse of Dimensionality

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



Dimensionality Reduction

Purpose

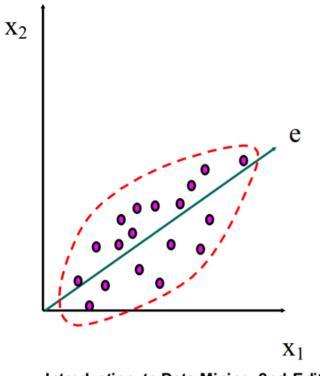
- Avoid curse of dimensionality
- Reduce amount of time and mem required by data mining algos
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

Techniques

- Principal Components Analysis (PCA)
- Singular value decomposition
- Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attrs
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example, students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

Discretization

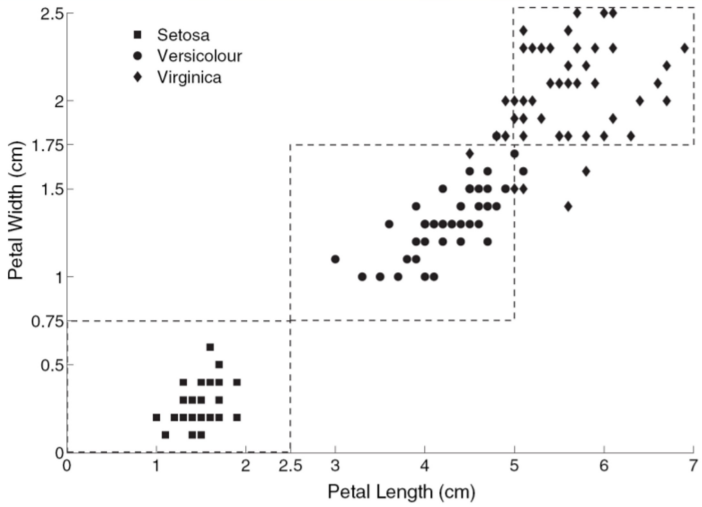
- Process of converting a continuous attr into an ordinal attr
- Potentially infinite number of values are mapped into a small # of categories
- Many classification algos work best if both indep. And dep. Variables have only a few values
- We give an illustration of the usefulness of discretization using the Iris data set

Iris Sample Data Set

- Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

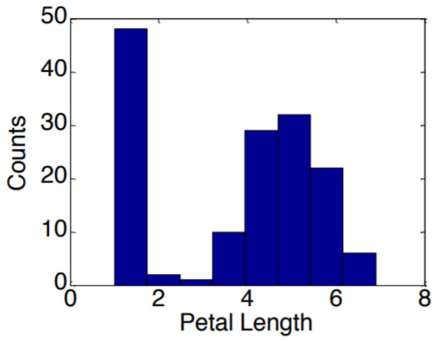
Discretization: Iris Example



Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

Discretization: Iris Example ...

- How can we tell what the best discretization is?
  - Unsupervised discretization: find breaks in the data values
    - Example: Petal Length



  - Supervised discretization: Use class labels to find breaks

Binarization

- Binarization maps a continuous or categorical attr into one or more binary varaibles
- Typically used for association analysis
- Often convert a continuous attr to a categorical attr and then convert a categorical attr to a set of binary attrs
  - Associates analysis needs asymmetric binary attrs
  - Examples: eye color and height measured as {low, medium, high}

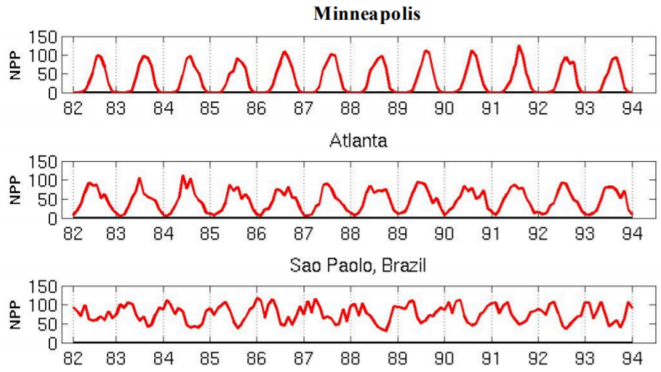
Attribute Transformation

- Function that maps the entire set of values of a given attr to a new set of replacement vals such that each old val can be identified with one of the new values

- Simple functions: $x^k, \log(x), e^x, |x|$
- **Normalization**
  - Refers to various techniques to adjust to differences among attrs in terms of freq of occurrence, mean, variance, range
  - Take out unwanted, common signal, e.g., seasonality
- In statistics, **standardization** refers to subtracting off the means and dividing by the standard deviation

---

Example: Sample Time Series of Plant Growth

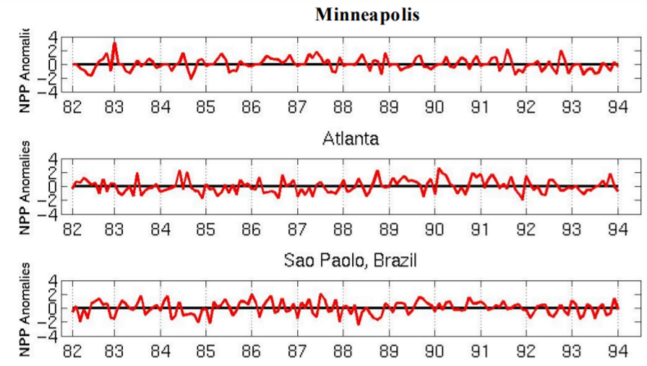


Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Seasonality Accounts for Much Correlation



Normalized using monthly Z Score:

Subtract off monthly mean and divide by monthly standard deviation

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000

Demonstration

Module 4 Data preprocessing file on canvas

Data.replace to replace documented empty with NaN null
Vals

Possibly take median, and set that for null vals

Or fill with default values
.fllna() to fill null vals

Use .dropna() to drop null values

Use .duplicated() to get all duplicated values
.drop_duplicates() to drop duplicates

.reindex(np.random.permutation()) to randomly reshuffle

Drop fields using .drop() to drop attrs or features you don't want

.insert() to add field into data frame

Normalize data using zscore, Example: replacing mpg with z-score

Going over different python functions

Going over aggregation
Going over sampling
Going over discretization
Going over concept of using PCA to reduce features for efficiency

Next Thursday's assignment

Take the data sets such as UCI breast cancer data set in module 4 or use other bad data sets found on google

- Do different types of modifications to data
- Learn different python functions
  - Such as strip off data and create new data set
- Kaggle/UCI data
  - Clean data
- Go on Google if you want data with issues
  - Message TA, Siddharth if you want data set with issues "dirty data"
    - Optional
- Don't use same dataset for weekly projects throughout the semester
  - Purpose is to learn how to solve different problems with data set