Data Preprocessing

James Kwok

Why Data Preprocessing?

real-world data is dirty

incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data (i.e., summaries)

- e.g., *occupation* = " "

Example

- "not applicable" data values upon collection (e.g., zip code)
- different considerations between the time when the data was collected and when it is analyzed
- human/hardware/software problems

Why Data Preprocessing?...

noisy: containing errors or outliers

- e.g., $salary =$ "-10"

Example

- faulty data collection instruments
- human or computer error at data entry
- errors in data transmission

inconsistent: containing discrepancies in codes or names

- e.g., $age =$ "42" and $birthday =$ "03/07/1997"

Example

- different data sources

redundant: containing duplicate records or unnecessary attributes

Example

- integration (i.e., merging) of datasets from different sources

Data Preprocessing

No quality data = no quality mining results!

- quality decisions must be based on quality data
- e.g., duplicate or missing data may cause incorrect or even misleading statistics

Why Data Preprocessing?

data size and complexity gravely affect the performance of the data mining tasks

- the larger the number of data objects to be analyzed, the more expensive the mining task
- the larger the number of attributes and the more complicated their value types, the more expensive the mining task

Data Preprocessing

Data preprocessing is a preparation stage before the actual data mining tasks are performed, which attempts to

- remove incomplete, noisy, inconsistent, and redundant data
- reduce the size and complexity of the data in order to refine the quality of the mining results and improve the performance of the mining tasks

Data preprocessing tasks

- data cleaning
- data integration
- data transformation
- data reduction

Types of Attributes

A dataset is a collection of objects

- alternative names for "object" include "record", "tuple", "point", "case", "sample", "entity", and "instance"

An object is characterized by a set of attributes

Example

name, address, eye color, temperature

- alternative names for "attribute" include "variable", "field", "characteristic", and "feature"

Attribute values are numbers or symbols assigned to an attribute

Example

$student\_name =$ 'John'

- alternative names for "attribute value" include "value" and "feature-value"

Types of Attributes...

Categorical

- nominal: provide enough information to distinguish one object from another

Example

zip codes, employee ID numbers, eye color, gender

  - binary attributes: assume only two values (e.g., yes/no, true/false, 0/1)

- ordinal: provide enough information to order objects

Example

grades, $\{good, better, best\}$

Numeric (continuous)

Examples

calendar dates, temperature in Celsius or Fahrenheit

Statistical Descriptions of Data

- gives the overall picture of the data
- involves
  - measuring the central tendency
  - measuring the dispersion
  - graphical display of descriptive summaries

Central Tendency

- the most common measure is the (arithmetic) mean (or average)
- let x_1, x_2, ..., x_N be N observations
- their (sample) mean is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- sometimes each value x_i is associated with a weight that signifies its importance. In this case, the weighted mean is:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

The mean is sensitive to extreme values

Central Tendency...

Example (Scoring for individual diving events)

- panel of seven judges
- the two highest and lowest scores of the panel are thrown out
- the rest of the scores are added together and multiplied by a difficulty rating
- multiplied by 0.6 (for easy comparison with other events)

Trimmed mean: disregards the low and high extremes

Example

- data: 0,4,5,6,7,7,8,10,11,18
- 10% trimmed mean omits 0 and 18
- trimmed mean: $\frac{4+5+6+7+7+8+10+11}{8} = 7.25$

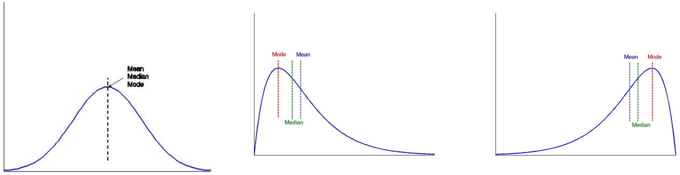
Central Tendency...

- a measure that is not sensitive to extreme values is the median, which represents the middle value of an ordered set of observations

Example

- median$(1, 5, 2, 8, 7)$= 5
- median$(1, 6, 2, 8, 7, 2) = (2 + 6)/2 = 4$

- mode: the value that occurs most frequently in the set
- midrange: average of the largest and smallest values in the data



Dispersion

- Let x_1, x_2, ..., x_N be a set of (numerical) observations.
- range: difference between the largest and smallest value
- kth percentile: value x_i with the property that k percent of the data are smaller than x_i (what percentile is the median?)
- quartiles: 25th percentile (denoted by Q_1), 50th percentile, and 75th percentile (denoted by Q_3)
- interquartile range:

$$IQR = Q_3 - Q_1$$

- five-number summary: consists of *minimum*, Q_1, *median*, Q_3, *maximum* (in this order)
  - gives a good impression of the center, spread, shape and distribution of the data

Dispersion

- variance $var(X) = E[(X-\mu)^2]$
- given a set of observations x_1, x_2, ..., x_N:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right]$$

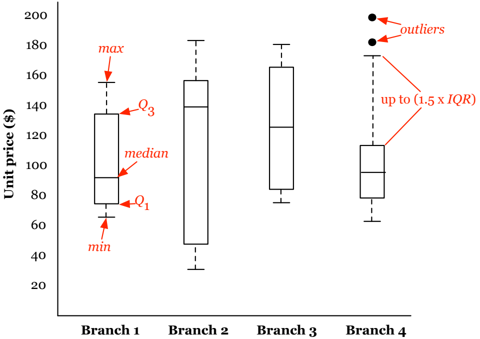
- standard deviation σ: square root of variance σ^2
  - σ indicates the spread of the values around the mean
  - $\sigma = 0$ when there is no spread, i.e., when all observations have the same value. Otherwise, $\sigma > 0$

Graphic Display

- usually useful to provide graphic displays of the data, in order to get some first impression of their characteristics
- examples of graphic displays include
  - boxplots
  - histograms
  - scatter plots

Boxplot

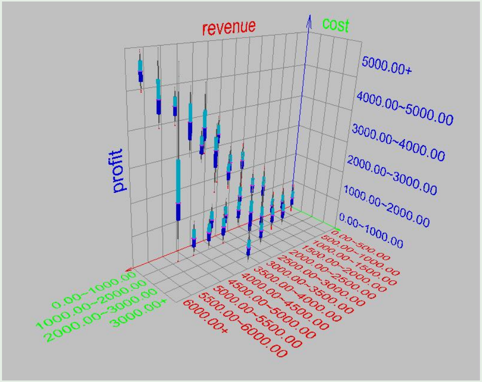
- boxplots incorporate the five-number summary



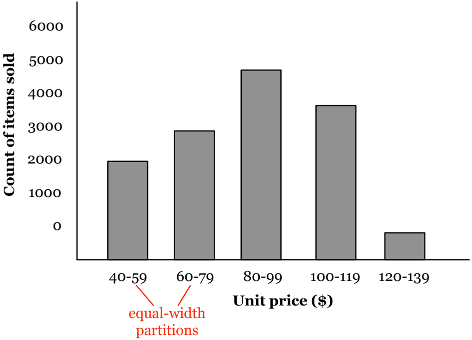
- the ends of the box are at the first and third quartiles
  - height of the box is IQR
- median is marked by a line within the box
- outlier: usually, a value higher/lower than 1.5 x IQR
- whiskers: two lines outside the box extended to minimum and maximum

Boxplot...

Example (3-D boxplot)



Histogram

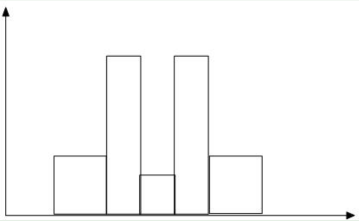


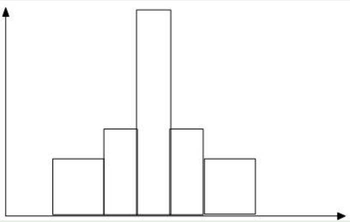
- divide data into buckets and store average (sum) for each bucket

Histogram...

Two histograms may have the same boxplot representation (i.e., the same values for: min, Q1, median, Q3, max), do they have the same data distributions?

Example

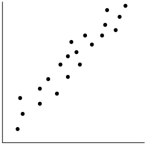




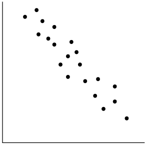
Scatter Plot

- determine whether there appears to be a relationship, pattern, or trend between two numerical attributes

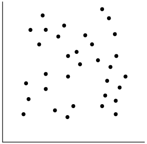
- each pair of values is treated as a pair of coordinates



positive correlation

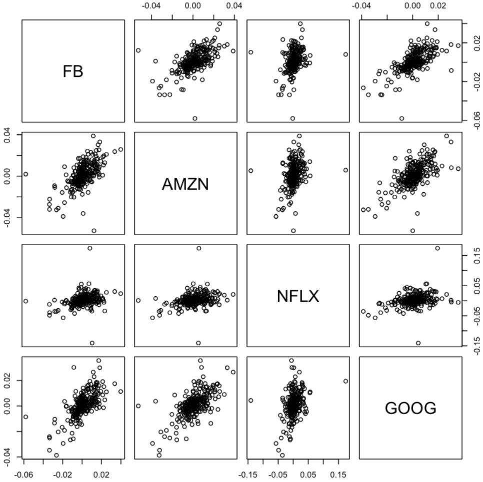


negative correlation



Example: Stocks Returns

- 4 stocks: Facebook, Amazon, Netflix, and Google



Data Cleaning

Data cleaning attempts to

- fill in missing values
  - e.g., Occupation=“”
- smooth out noise, outliers
  - e.g., Salary = “-10”
- correct inconsistencies in the data
  - e.g., Age = “42”, Birthday = “03/07/2010”
  - e.g., discrepancy between duplicate records

Missing Values

What to do with missing values?

- ignore the record
  - usually done when class label is missing (when doing classification)
  - ignoring the tuple → cannot make use of the remaining attributes' values in the tuple
- fill in the missing value manually
  - tedious + infeasible?
- use a global constant to fill in the missing value
  - e.g., "unknown" (a new class?!)
- use the attribute mean to fill in the value
- use the attribute mean of a certain class (in which the record belongs)
- use the most probable value to fill in the missing value
  - extra tools may be needed to compute the most probable value

Missing Values...

Should we fill in all missing values?

- “not applicable” or “don’t know” is implied
- left intentionally blank in order to be provided later

Noise

How to smooth out noise?

- remove outliers found by graphic display (explained in previous slides)
- binning (explained in the next slide)
- regression
  - smooth by fitting the data into regression functions
  - explained in later lectures
- clustering
  - detect and remove outliers
  - explained in later lectures

Binning

- smooths a sorted value by consulting its "neighborhood" (i.e., the values around it)
- e.g., smoothing by bin means/medians

Example

sorted data for *price* in \$: 4, 8, 15, 21, 21, 24, 25, 28, 34

- partition into equal-sized bins
  Bin 1: 4, 8, 15   Bin 2: 21, 21, 24   Bin 3: 25, 28, 34
- smoothing by bin means
  Bin 1: 9, 9, 9   Bin 2: 22, 22, 22   Bin 3: 29, 29, 29

Inconsistencies

How to correct inconsistencies?

- some cases are easy to detect (or even fix), provided that we possess some domain knowledge, also called metadata (i.e., data about the data)

Example

A person's height should not be negative

- other cases are much trickier, in which case we may need to consult an external source of information

Example

check the customer's address in a reimbursement form against the customer database of the insurance company

Data Integration

Data integration combines data from multiple sources into a coherent data store

What should we consider during data integration?

- entity identification problem
- data value conflicts
- data redundancy

Entity Identification Problem

Do two objects from different data sources refer to the same entity?

Example

Is the record that has $customer\_id = 234$ (from one source) equivalent to that where $cust\_num = 234$ (from the other source)?

Metadata can help

- e.g., for each attribute, look at the name, meaning, data type, range of values permitted, etc

Data Value Conflicts

Example

For the same entity, attribute values from different sources may differ

- e.g., *weight* measured in kilograms or pounds

Example

Attribute *total_sales* in one database may refer to the total sales of a company branch, whereas in another it may refer to the total sales for all branches in a specific region

once again, metadata may help

- e.g., What are the acceptable values for each attribute? What is the range of values? What is the standard deviation?

Data Redundancy

record redundancy

- there are two or more identical tuples for a unique entity

attribute redundancy

- one attribute may be "derived" from another attribute, or a set of attributes

Example

annual_income may be derived by *annual_revenue* and *annual_expenses*

- may be able to be detected by measuring how related two attributes are

Numerical Attributes: Correlation Coefficient

Given N tuples, are numerical attributes A and B correlated?

Let

- a_i, b_i: values of attribute A and B for the ith tuple
- $\bar{A}, \bar{B}$: respective means
- σ_A, σ_B: respective standard deviations

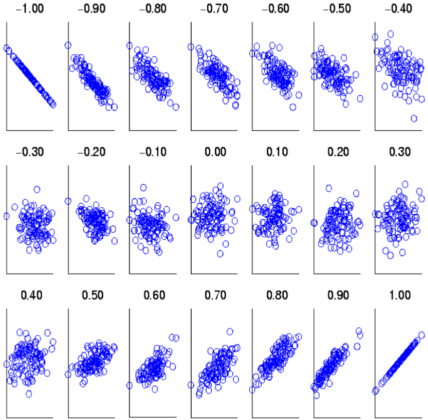
The correlation coefficient is given by

$$r_{A,B} = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{\sigma_A \sigma_B} = \frac{\sum_{i=1}^{N} a_i b_i - N\bar{A}\bar{B}}{N\sigma_A \sigma_B}$$

Correlation Coefficient...

$$-1 \leq r_{A,B} \leq +1$$

- $r_{A,B} > 0$
  - A and B are positively correlated
- $r_{A,B} < 0$
  - A and B are negatively correlated
- $r_{A,B} = 0$
  - A and B are uncorrelated



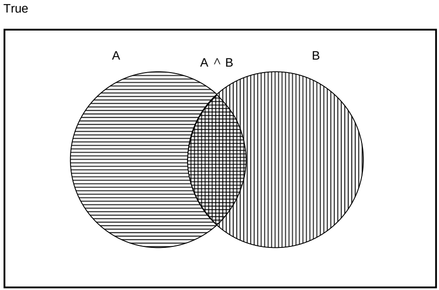
Categorical Attributes: χ^2 test

- A and B be two categorical attributes
- A has c distinct values, B has r distinct values

are A and B independent?

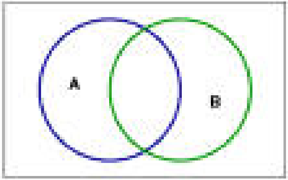
Revision: Axioms for Probability

- All probabilities are between 0 and 1: $0 \leq P(A) \leq 1$
- Necessarily true propositions have probability 1: $P(True) = 1$
- Necessarily false propositions have probability 0: $P(False) = 0$
- The probability of a disjunction: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Conditional Probability

- let A and B be two events such that $P(A) > 0$
- $P(B|A)$: probability of B given that A has occurred



$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \qquad P(A \cap B) = P(A)P(B|A)$$

- probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred

Independence

- two random variables X and Y are independent if

$$P(X|Y) = P(X), \text{ or } P(Y|X) = P(Y)$$

- knowledge about X contains no information about Y
- equivalently, $P(X,Y) = P(X)P(Y)$

Example

- X: result of tossing a fair coin for the first time; Y: result of second tossing of the same coin
- X: result of US election; Y: your grades in this course

Question: Are these independent?

X: midterm exam grade; Y: final exam grade

χ^2 Test

- Let the distinct values of A be $\{a_1, a_2\} = \{male, female\}$
- Let the distinct values of B be $\{b_1, b_2\} = \{fiction, non\_fiction\}$
- Create a 2x2 contingency table, putting the values of A as column labels, and those of B as row labels

	male	***female***	***Total***
fiction	250	200	450
non_fiction	50	1000	1050
Total	300	1200	1500

- In every cell (a_i, b_j): observed frequency o_{ij}
  - the actual count of records that have A$=a_i$ and B$=b_j$

are A and B independent?

χ^2 Test...

- For every cell (a_i, b_j) compute the expected frequency e_{ij} of the event that A=a_i and B=b_j, assuming A and B are independent:
  - recall that you have N tuples

$$
\begin{aligned}
e_{ij} &= N \times P(A = a_i \wedge B = b_j) \\
&= N \times P(A = a_i) \times P(B = b_j) \\
&= \frac{1}{N}(count(A = a_i) \times count(B = b_j))
\end{aligned}
$$

	male	***female***	***Total***
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

- observed $\simeq$ expected $\rightarrow$ A and B are independent
- observed $\not\simeq$ expected $\rightarrow$ A and B are dependent

χ^2 Test...

- compute the χ^2 value using the following formula:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  - large $\chi^2 \rightarrow$ more likely A and B are related
- hypothesis testing
  - hypothesis: A and B are independent
- compute the degrees of freedom as $(r-1) \times (c-1)$

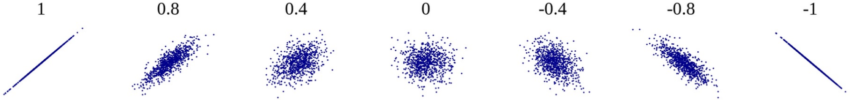
Degrees of freedom	*x^2 value*										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
Level of significance	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Non-significant								*Significant*		

Significance →

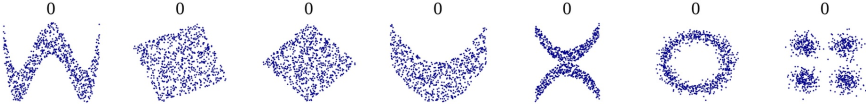
- reject the hypothesis (typically) at significance level 0.001, if χ^2 is larger than the corresponding value at the table
  - In our example, $\chi^2 = 507.93 > 10.83$ and, thus, A and B are not independent (the hypothesis is rejected)

Correlation and Dependence

Zero correlation coefficient = Independent?



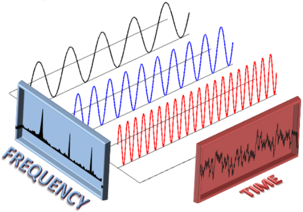
- independence ⇒ uncorrelated?yes
- uncorrelated ⇒ independence? no



Data Transformation

Goal: modify the data in order to improve data mining performance

- attribute/feature construction
  - create new attributes (features) that can capture the important information in a data set more effectively than the original ones
  - e.g., Fourier transform, wavelet transform



- normalization: scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
- discretization

Why Data Normalization?

Example

one variable is 100 times larger than another (on average)

- your model may be better behaved if you normalize (standardize) the two variables to be approximately equivalent

Min-Max Normalization

- min_A, max_A: minimum and maximum values of attribute A
- maps a value v of A to a new range $[new\_min_A, new\_max_A]$ as

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Example

- let income range \$12,000 to \$98,000 normalized to $[0.0, 1.0]$
- then \$73,000 is mapped to $\frac{73600-12000}{98000-12000} = 0.716$

- preserves the relationships among the original values
- typically, we map values to range $[0.0, 1.0]$ or $[-1.0, 1.0]$

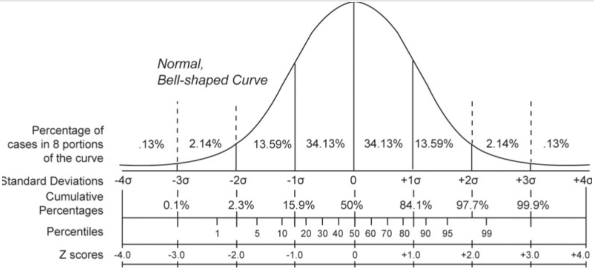
Z-Score Normalization

how many standard deviations from the average that your data lies?

- the new value v' is calculated as $v' = \frac{v - \bar{A}}{\sigma_A}$

Example

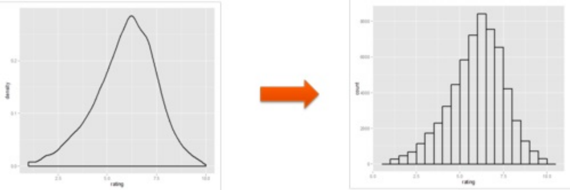
- let $\mu = 54,000, \sigma = 16,000$. Then $\frac{73000-54000}{16000} = 1.225$



- useful when we do not know the minimum and maximum of an attribute, or when we have outliers

Discretization

- divides the range of a continuous attribute (e.g., *age*) into intervals
    - assign these intervals labels such as $0-10, 11-20, \ldots,$ or *youth*, *adult*, *senior*, etc
- reduces data size
- example methods
    - histograms (discussed in the previous lecture)
        - the new dataset consists of the bucket labels



    - cluster analysis (will be discussed in a later lecture)
    - decision-tree analysis (will be discussed in a later lecture)

Data Reduction

Data reduction obtains a reduced representation of the dataset, while allowing (almost) the same analytical results to be produced

Why data reduction?

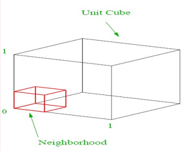
- a database/data warehouse may store terabytes of data
- complex data analysis may take a very long time to run on the complete data set

Data reduction strategies

- dimensionality reduction: reduce the number of attributes
  - principal components analysis (PCA)
  - feature subset selection
- use a smaller form of data representation
  - regression, histograms, clustering, sampling, data cube aggregation

Motivation

Suppose that points are uniformly distributed in a d-dimensional unit hypercube. If we want to construct a hypercube neighborhood to capture a fraction r of the observations, what is the edge length ℓ of this cube?



- volume of cube: $\ell^d = r$; we have $\ell = r^{1/d}$

$d = 1$

- if $r = 0.01$ then $\ell = 0.01$; if $r = 0.1$ then $\ell = 0.1$

$d = 10$

- if $r = 0.01$ then $\ell = 0.63$; if $r = 0.1$, then $\ell = 0.80$
- in order to capture 1% (or 10%) of the data, we must cover 63% (80%) of the range of each input

Motivation...

If $n = 100$ represents a dense sample for one single input, how large should n be in order to have the same sampling density with $d = 10$?

- $n = 100^{10}$
- the number of required points increases exponentially to maintain the same sampling density

Curse of dimensionality

- when dimensionality increases, data becomes increasingly sparse
- density and distance between points $\rightarrow$ less meaningful

Dimensionality Reduction

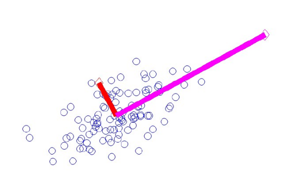
- avoid the curse of dimensionality
- help eliminate irrelevant features and reduce noise
- reduce time and space required in data mining
- allow easier visualization

Dimensionality reduction techniques

- principal component analysis
- feature selection

Principal Component Analysis (PCA)

Find a projection that captures the largest amount of variation in data



- the original data are projected onto a lower-dimensional space, resulting in dimensionality reduction

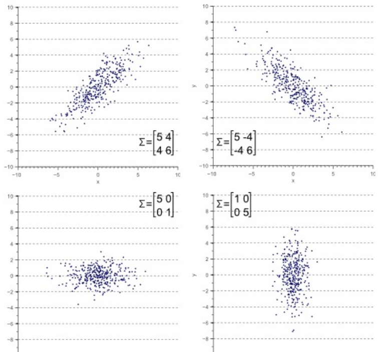
How to Measure Variation in Data?

- 1d: variance
- 2d: variance and covariance

$$\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})$$

  - correlation coefficient?

Given covariance matrix $\begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}$, what will the data look like?



How to Measure Variation in Data?...

- 3d: attributes (x, y, z)

$$\begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

- covariance matrix

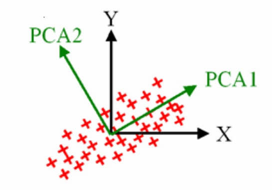
$$\mathbf{C} = \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T$$

projection that captures the largest amount of variation
$=$ projection $\mathbf{w}$ s.t. var$(\mathbf{w}^t\mathbf{x})$ is maximized

PCA

find the eigenvectors of the covariance matrix $\mathbf{C}$

- $\mathbf{Cv} = \lambda \mathbf{v}$
- $\mathbf{v}$: principal components
- λ: eigenvalue
  - measures the variance magnitude in the direction of the eigenvector
  - decreasing eigenvalue $\rightarrow$ decreasing "significance" or strength



Example

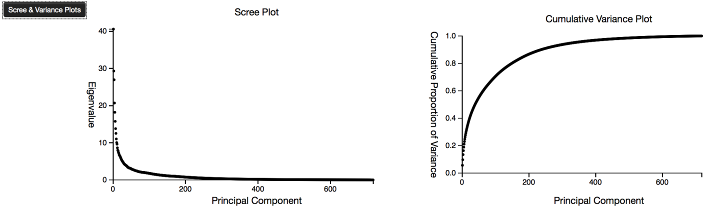
X_1	X_2
19	63
39	74
30	87
30	23
15	35
15	43
15	32
30	73

- $C = \begin{bmatrix} 75 & 106 \\ 106 & 482 \end{bmatrix}$;
- eigenvectors:
  - $e_1 = [-0.98, -0.21], \lambda_1 = 51.8$; $e_2 = [0.21, -0.98], \lambda_2 = 560.2$
  - the second eigenvector is more important! keep!

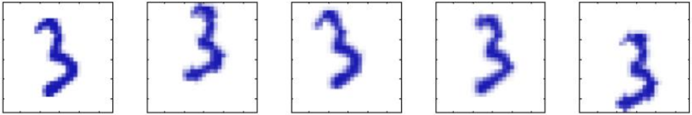
Proportion of Variance Explained

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}$$

- λ_i are sorted in descending order
- decreasing eigenvalue $\rightarrow$ decreasing "significance" or strength
- data size reduced by eliminating components with small eigenvalues

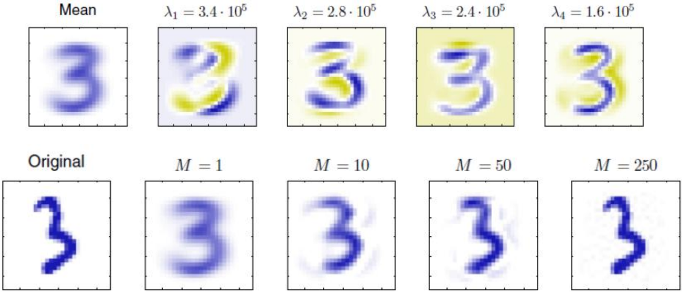


Example



- a collection of 100×100 images created from one image by introducing random displacement and rotation

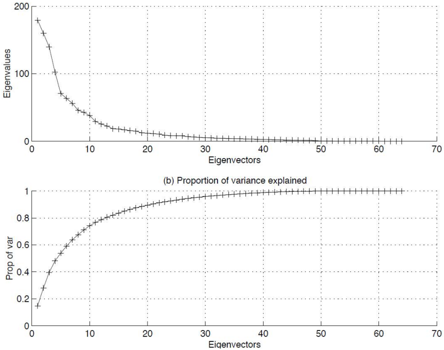
eigenvectors



How to Choose the Number of PCs (k)?

Use the proportion of variance explained:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_d} \quad (\lambda_i \text{ are sorted in descending order})$$



- e.g., stop at proportion of variance > 0.9

Attribute/Feature Subset Selection

- another way to reduce dimensionality of data
- redundant attributes
  - duplicate much or all of the information contained in one or more other attributes

Example

"purchase price of a product" and "the amount of sales tax paid"

- irrelevant attributes
  - contain no information that is useful for the data mining task

Example

students' ID is often irrelevant to the task of predicting students' GPA

Goal

Select the minimum possible subset of attributes, such that the quality of the data mining task is not compromised

Set of all features → Selecting the best subset → Learning Algorithm → Performance

Challenging

- there are 2^d possible attribute combinations of d
- very difficult to test all possible (exponential) combinations of attributes attributes

Typical Attribute Selection Methods

Forward selection	Backward elimination
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$
Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	=> $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$

1. greedy forward selection
   - the best single-attribute is picked first
   - the next best attribute condition to the first, ...

which attribute is the best?

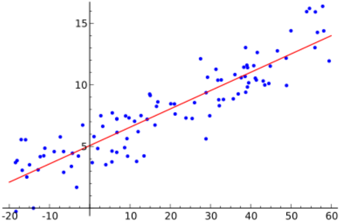
   - e.g., correlation between the attribute and target value
2. greedy backward elimination
   - repeatedly eliminate the worst attribute
3. combined attribute selection and elimination

Numerosity Reduction

- reduces the data volume by choosing smaller forms of data representation

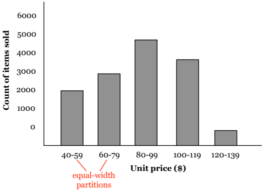
Parametric Methods

- assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- e.g., in linear regression the data can be modeled to fit a straight line (will be studied in a later lecture)

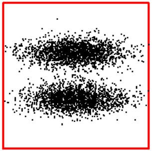


Nonparametric Methods

- do not assume models, e.g.,
- histograms (discussed in the previous lecture)



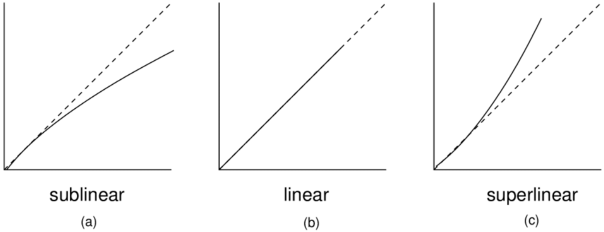
- clustering



  - e.g., a cluster of points can be represented by their centroids
  - clustering will be discussed in a later lecture
- sampling
- data cube aggregation

Sampling

- obtaining a small sample to represent the whole data set
- allow a mining algorithm to run in complexity that is potentially sublinear to the size of the data



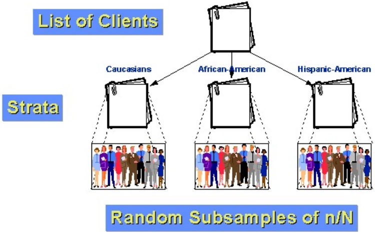
With replacement or not?

- sampling without replacement
  - once an object is selected, it is removed from the population
- sampling with replacement
  - a selected object is not removed from the population

Sampling...

How to choose a representative subset of the data?

- simple random sampling
  - there is an equal probability of selecting any particular item
- stratified sampling
  - partition the data set (e.g., by age group), and draw samples from each partition



  - useful when the data is skewed

Data Cube Aggregation

- use the smallest representation which is enough to solve the task