SAN JOSÉ STATE UNIVERSITY

Know Your Data II

1

SAN JOSÉ STATE UNIVERSITY

Agenda

- Correlation & Covariance Analyses
- Data Transformation

2

2

SAN JOSÉ STATE UNIVERSITY

Correlation & Covariance

3

SAN JOSÉ STATE UNIVERSITY

Correlation Analysis (for Categorical Data)

- X^2 (chi-square) test: correlation between 2 attributes A (a_1, a_2, …) and B (b_1, b_2, …)

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

expected frequency of (a_i, b_j)

- Null hypothesis: The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
  - The larger the X^2 value, the more likely the variables are related
- Note: **Correlation does not imply causality**
  - \# of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

4

4

SJSU SAN JOSÉ STATE UNIVERSITY

Chi-Square Example

Given 1500 people with the following contingency table:

	Male	Female	Sum (row)
Like fiction	250 (e_{11})	200 (e_{12})	450
Like non fiction	50 (e_{21})	1000 (e_{22})	1050
Sum(col.)	300	1200	1500

Are gender correlated to fiction or non-fiction?

- Null hypothesis: The two distributions are independent (no correlation)
- First, compute the expected frequencies.

SJSU SAN JOSÉ STATE UNIVERSITY

Chi-Square Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Degree of freedom (number of values that are free to vary)
  - (#categories in variable A - 1) * (#categories in B -1)

SJSU SAN JOSÉ STATE UNIVERSITY

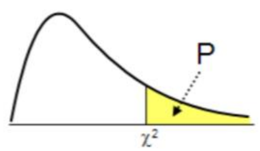
Chi-Square Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

Degree of freedom =?

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Values of the Chi-squared distribution



	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458

7

SJSU SAN JOSÉ STATE UNIVERSITY

Variance for Single Variable (Numerical Data)

- The variance of a random variable X provides a measure of how much the value of X deviates from the mean or expected value of X:

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

where σ^2 is the variance of X, σ is the standard deviation

- μ is the mean, and μ = E[X] is the expected value of X
- That is, variance is the expected value of the square deviation from the mean

- It can also be written as:

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$$

8

8

SJSU SAN JOSÉ STATE UNIVERSITY

Covariance for Two Variables

- Covariance between two variables X_1 and X_2 :

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the mean or expected value of X_1; similarly for μ_2

- Sample covariance between X_1 and X_2 : $\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \widehat{\mu_1})(x_{i2} - \widehat{\mu_2})$
- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \widehat{\mu_1})(x_{i1} - \widehat{\mu_1})$$

9

9

SJSU SAN JOSÉ STATE UNIVERSITY

Covariance for Two Variables

- Positive covariance: $\sigma_{12} > 0$
- Negative covariance: $\sigma_{12} < 0$
- Independence: If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true!!
  - Some pairs of random variables may have a covariance 0 but are not independent
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

10

SJSU SAN JOSÉ STATE UNIVERSITY

Covariance Analysis of Numeric Data

Consider the table of stock prices:

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$E[X_1] = (6 + 5 + 4 + 3 + 2)/\ 5 = 20/5 = 4$

$E[X_2] = (20 + 10 + 14 + 5 + 5)\ /5 = 54/5 = 10.80$

$E[X_1\ X_2] = (6\times20 + 5\times10 + 4\times14 + 3\times5 + 2\times5)/5 = 60.2$

$\sigma_{12} = 60.2 - 4 \times 10.80 = 17$

Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

SJSU SAN JOSÉ STATE UNIVERSITY

Correlation between Two Numerical Variables

- Correlation between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- Sample correlation for two attributes X_1 and X_2:

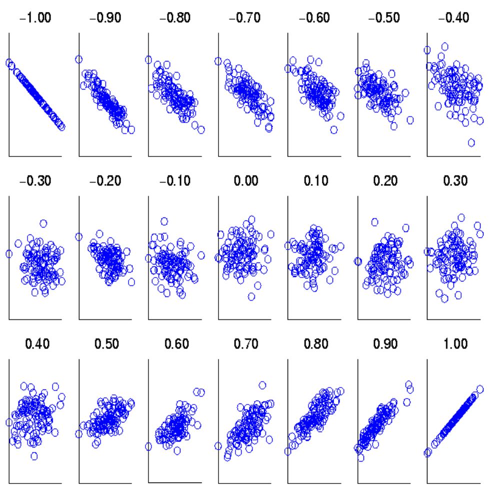
$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)\,(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n}(x_{i2} - \hat{\mu}_2)^2}}$$

  where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2, σ_1 and σ_2 are the respective standard deviation of X_1 and X_2

- If $\rho_{12} > 0$: A and B are positively correlated (X_1's values increase as X_2's)
  - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

SJSU SAN JOSÉ STATE UNIVERSITY

Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: [–1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1

13

SJSU SAN JOSÉ STATE UNIVERSITY

Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 X 2 covariance matrix as

$$\Sigma = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = E[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}(X_1-\mu_1 \quad X_2-\mu_2)]$$

$$= \begin{pmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] \end{pmatrix}$$

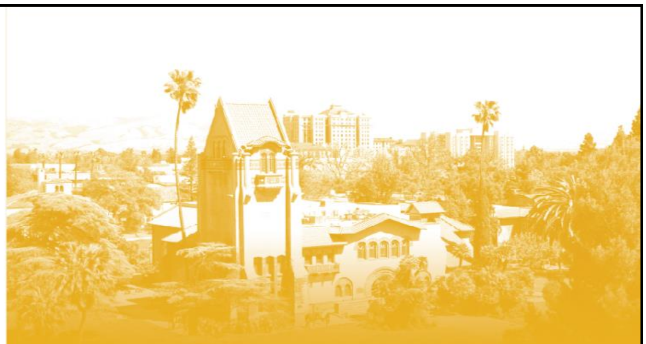
$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Generalizing it to *d* dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix}$$

$$\mathbf{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

14



SJSU SAN JOSÉ STATE UNIVERSITY

Data Transformation

15

SJSU SAN JOSÉ STATE UNIVERSITY

Data Transformation

Maps the entire set of values of a given feature to a new set of replacement values such that each old value can be identified with one of the new values.

- Aggregation
- Sampling
- Discretization & Binarization
- Feature Subset Selection
- Feature Creation
- Dimensionality Reduction

16

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.
    - Days aggregated into weeks, months, or years
- More "stable" data - aggregated data tends to have less variability

17

17

Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling

(a) 8000 points (b) 2000 points (c) 500 points

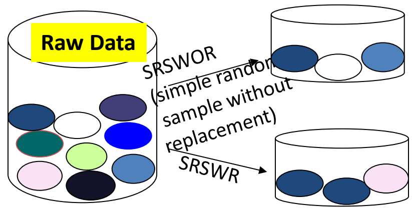
18

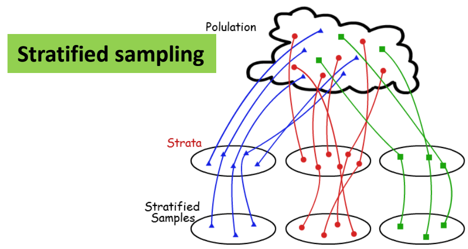
18

SJSU SAN JOSÉ STATE UNIVERSITY

Types of Sampling

- Simple random sampling: equal probability of selecting any particular item
- 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
- Stratified sampling
  - Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)





SJSU SAN JOSÉ STATE UNIVERSITY

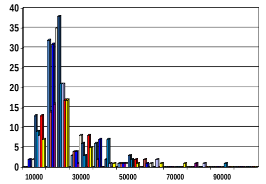
Discretization & Binarization

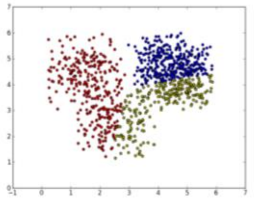
- Discretization converts a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is used in both unsupervised and supervised settings
- Binarization maps a continuous or categorical attribute into one or more binary variables

SJSU SAN JOSÉ STATE UNIVERSITY

Data Discretization Methods

- Binning
  - Top-down split, unsupervised
- Histogram analysis
  - Top-down split, unsupervised
- Decision-tree analysis
  - Supervised, top-down split
- Clustering analysis
  - Unsupervised, top-down split or bottom-up merge
- Correlation (e.g., χ^2) analysis
  - Unsupervised, bottom-up merge





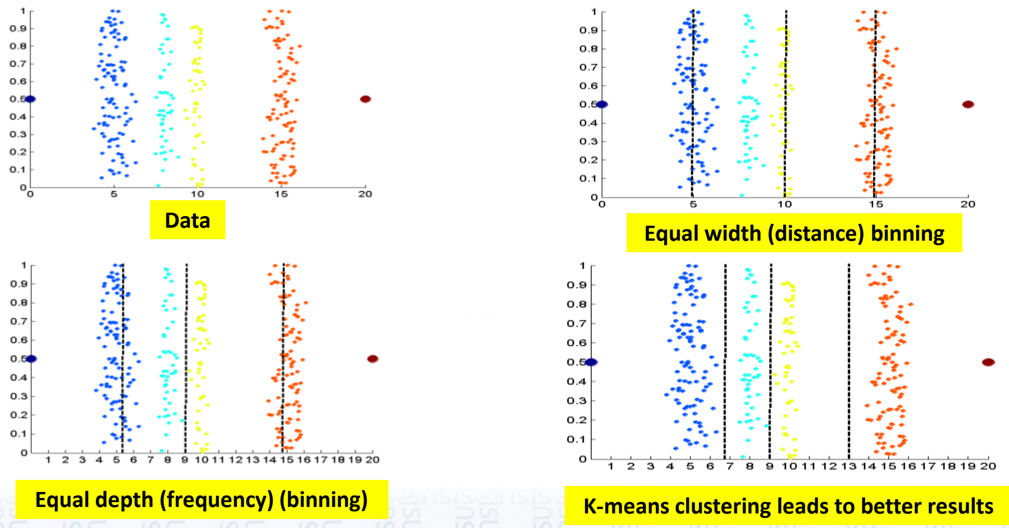
SJSU SAN JOSÉ STATE UNIVERSITY

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be:

    w = (B –A)/N
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Discretization Without Supervision: Binning vs. Clustering



23

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

symmetric binary attributes

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

asymmetric binary attributes

24

24

SJSU SAN JOSÉ STATE UNIVERSITY

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using entropy to determine split point (discretization point)
  - Top-down, recursive split
  - Details to be covered in Chapter “Classification”
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: Find the best neighboring intervals (those having similar distributions of classes, i.e., low $\chi 2$ values) to merge
  - Merge performed recursively, until a predefined stopping condition

25

25

SJSU SAN JOSÉ STATE UNIVERSITY

Dimensionality Reduction

26

SJSU SAN JOSÉ STATE UNIVERSITY

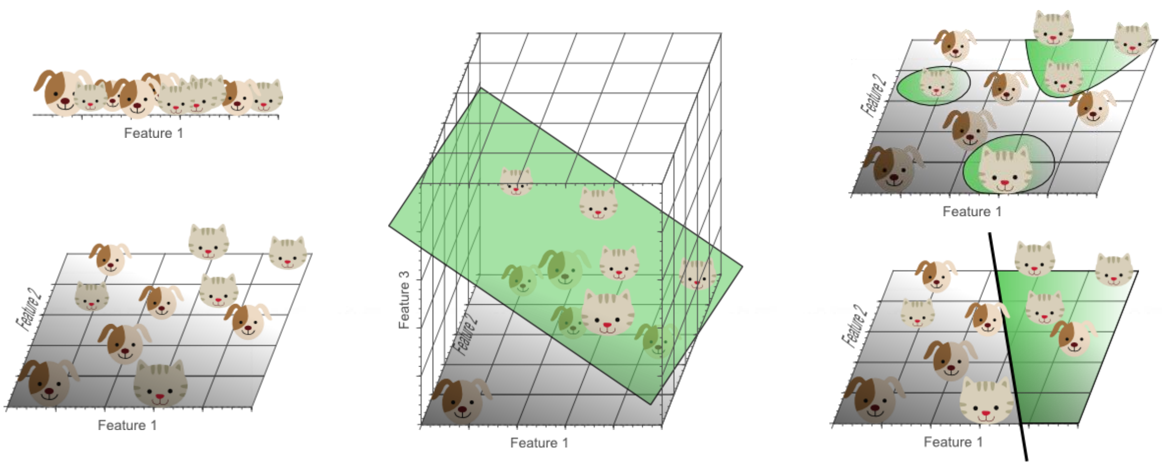
What Is Dimensionality Reduction?

- Curse of Dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density & distance between points (crucial to clustering etc) becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- Dimensionality Reduction
  - Reducing the number of variables under consideration (with principal variables)
- Advantages of Dimensionality Reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Better visualization

27

SJSU SAN JOSÉ STATE UNIVERSITY

Curse of Dimensionality



https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/#google_vignette

28

28

SJSU SAN JOSÉ STATE UNIVERSITY

Dimensionality Reduction Methods

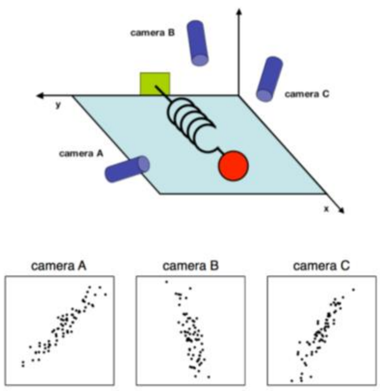
- Dimensionality Reduction Methodologies
  - **Feature selection**: Find a subset of the original variables (or features, attributes)
  - **Feature extraction**: Transform the data in the high-dimensional space to a space of fewer dimensions
- Some typical dimensionality reduction methods
  - Principal Component Analysis
  - Supervised & Nonlinear Techniques

29

SJSU SAN JOSÉ STATE UNIVERSITY

Principal Component Analysis (PCA)

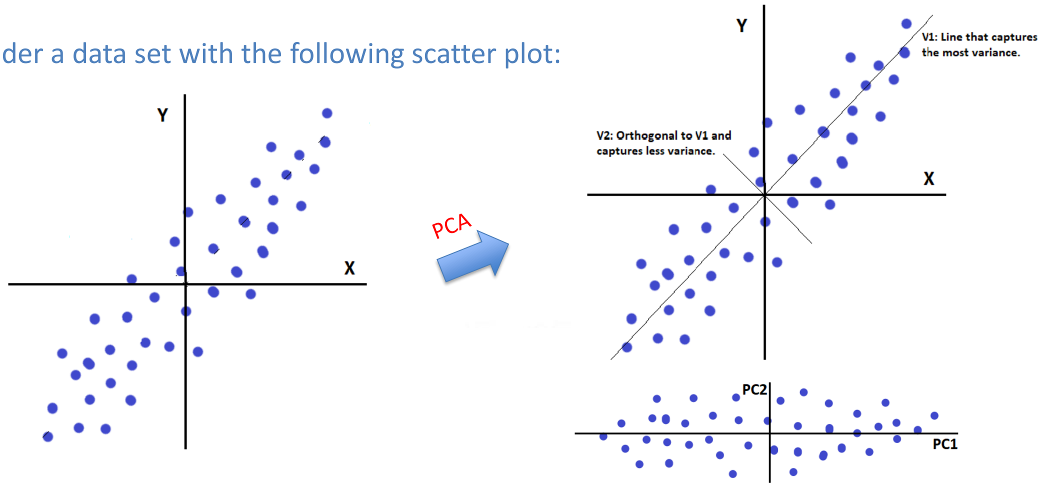
- PCA: A numerical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components
- The original data are projected onto a much smaller space, resulting in dimensionality reduction
- Method: Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



30

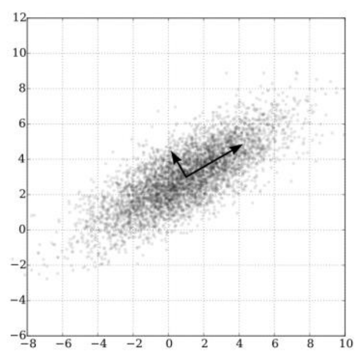
Graphical Explanation of PCA

- Consider a data set with the following scatter plot:



Principal Component Analysis (Method)

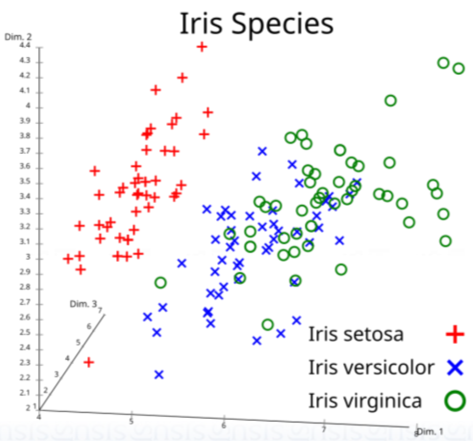
- PCA Steps
  - Standardize the Data (i.e. zero mean & unit standard deviation
  - Compute covariance matrix
  - Compute eigenvalues & eigenvectors
  - Sort eigenvalues
  - Select principal components (top k eigenvectors)
  - Transform the Data
- Works for numeric data only

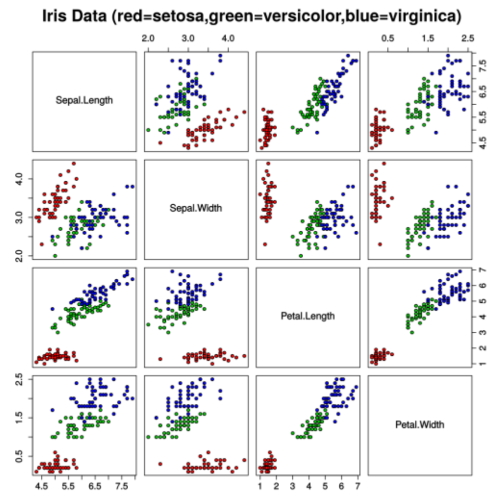


SJSU SAN JOSÉ STATE UNIVERSITY

PCA Example: Iris Dataset

- Iris flower dataset has 4 features, 3 targets:

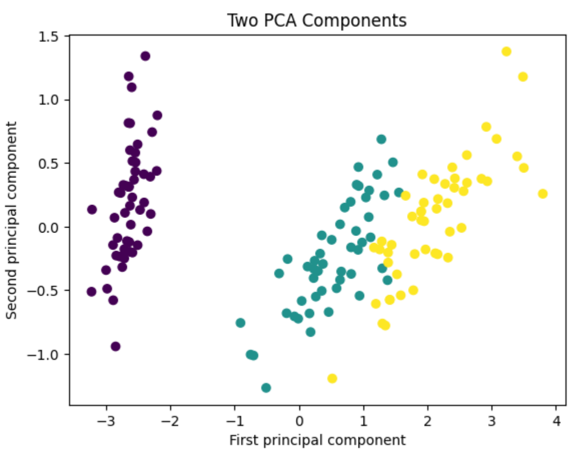


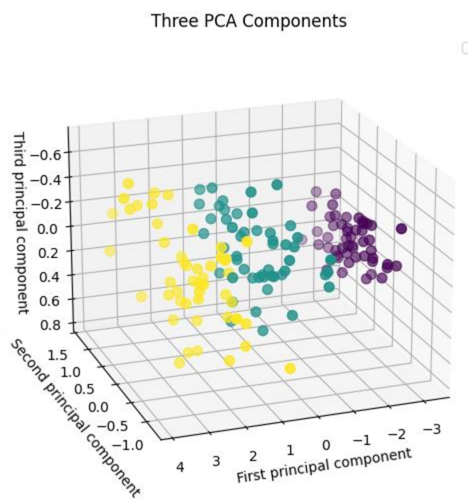


33

SJSU SAN JOSÉ STATE UNIVERSITY

PCA Example: Iris Dataset





34

SJSU SAN JOSÉ STATE UNIVERSITY

Singular Value Decomposition

- A given matrix A can be decomposed as:

$$A_{m\times n} = U_{m\times m} S_{m\times n} V^*_{n\times n}$$

where U and V are unitary (orthogonal), and S is (sorta) diagonal

$$A = USV^* = [u_1\ u_2\ \dots\ u_m] \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix}$$

$$S = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \text{ or } S = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \dots & 0 \\ 0 & 0 & \dots & \sigma_m & 0 & \dots & 0 \end{bmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ real & positive

SJSU SAN JOSÉ STATE UNIVERSITY

Singular Value Decomposition

- The columns of U and V are called the left and right singular vectors

$$U = [u_1\ u_2\ \dots\ u_m]$$
$$V = [v_1\ v_2\ \dots\ v_n].$$

$$A = USV^* \Rightarrow AV = US \Rightarrow Av_k = \sigma_k u_k$$

$$A = USV^* \Rightarrow A^*U = VS^* \Rightarrow A^*u_k = \sigma_k v_k$$

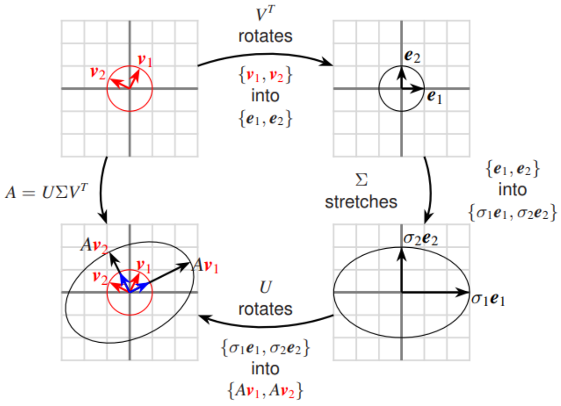
$$\left.\begin{matrix} Av_k = \sigma_k u_k \\ A^*u_k = \sigma_k v_k \end{matrix}\right\} \quad A^*Av_k = \sigma_k^2 v_k$$

eigenvalues & eigenvectors of A*A

SJSU SAN JOSÉ STATE UNIVERSITY

Singular Value Decomposition Illustration

- SVD = rotation + scaling + rotation



SJSU SAN JOSÉ STATE UNIVERSITY

SVD Truncation

- Recall

$$A = USV^* = [u_1 \ u_2 \ \dots \ u_m] \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix} \Rightarrow A = [\sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \dots + \sigma_n u_n v_n^*]$$

- We can form a "truncated" version of A with fewer # of terms:

$$A_k = \sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \dots + \sigma_k u_k v_k^*$$

- Error of truncation: $A_n - A_k = \sigma_{k+1} u_{k+1} v_{k+1}^* + \dots + \sigma_n u_n v_n^*$

SVD Example

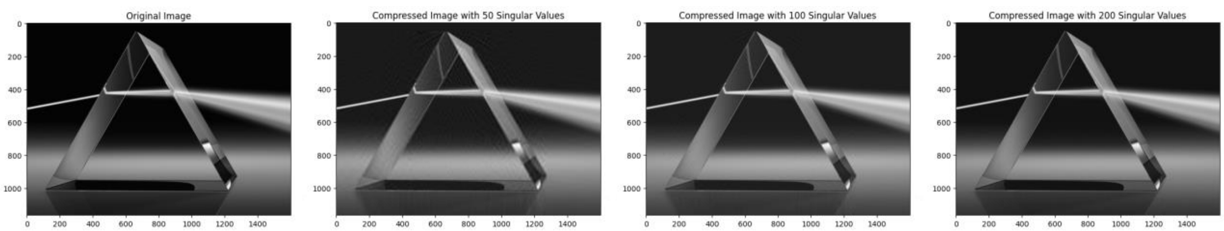
- Consider the 4×5 matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 & 0 & 0 & \mathit{0} \\ 0 & \sqrt{5} & 0 & 0 & \mathit{0} \\ 0 & 0 & 2 & 0 & \mathit{0} \\ 0 & 0 & 0 & 0 & \mathit{0} \end{bmatrix} \quad \mathbf{V}^* = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ -\sqrt{0.2} & 0 & 0 & 0 & -\sqrt{0.8} \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

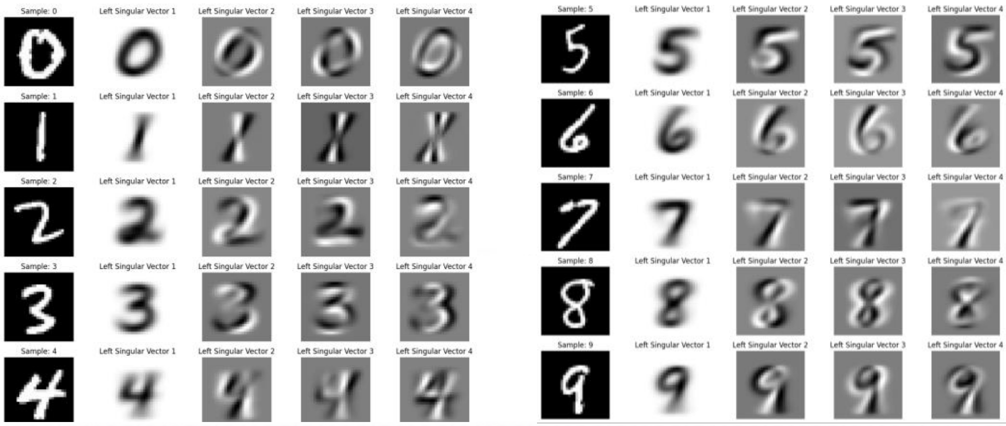
SVD Example: Image Processing

- SVD can be applied to image compression:



SJSU SAN JOSÉ STATE UNIVERSITY

MNIST Dataset with SVD



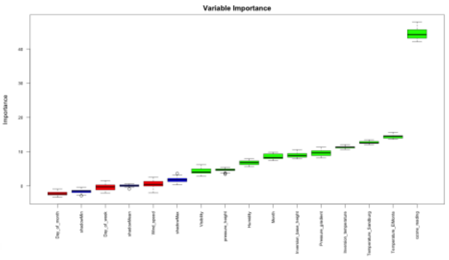
SJSU SAN JOSÉ STATE UNIVERSITY

Feature Selection & Creation

SJSU SAN JOSÉ STATE UNIVERSITY

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 attributes
    - e.g. Purchase price of a product & the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
    - e.g. Student's ID is often irrelevant to the task of predicting his/her GPA



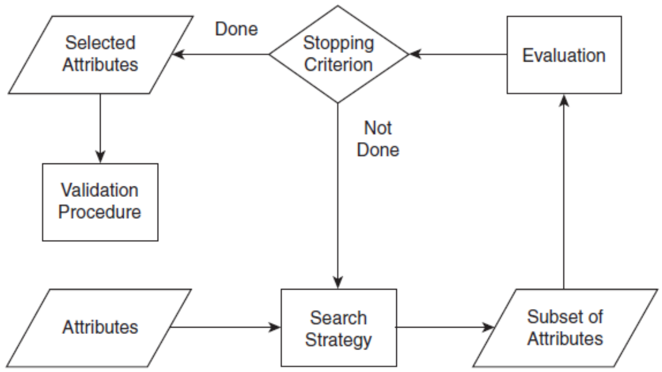
SJSU SAN JOSÉ STATE UNIVERSITY

Feature Selection Techniques

- Brute-force approach
  - Try all possible feature subsets as an input to the algorithm
- Embedded approach
  - Feature selection occurs naturally as a part of data mining algorithm
- Filter approach
  - Features are selected before the data mining algorithm is run
- Wrapper approach
  - Use the target data mining algorithm as a black box to find the best subset of attributes

SJSU SAN JOSÉ STATE UNIVERSITY

Feature Selection Process Flowchart



45

45

SJSU SAN JOSÉ STATE UNIVERSITY

Filter Approach

Set of all Features → **Selecting the Best Subset** → **Learning Algorithm** → **Performance**

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

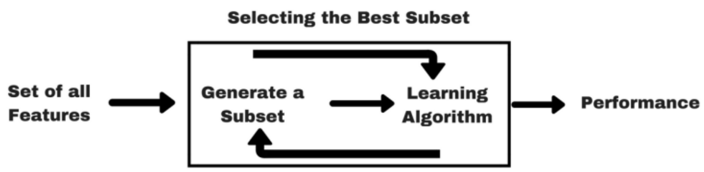
46

46

SJSU SAN JOSÉ STATE UNIVERSITY

Wrapper Approach

- Forward Selection: An iterative method that keeps adding the feature in each iteration which best improves our model till an addition of a new variable does not improve the performance of the model.
- Backward Elimination: Removes the least significant feature at each iteration which improves the performance of the model. Repeat this until no improvement is observed on removal of features.
- Recursive Feature elimination:. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.



47

47

SJSU SAN JOSÉ STATE UNIVERSITY

Feature Creation or Generation

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Feature Extraction (domain specific)

    e.g. extracting edges from images, total price from the sale tax
  - Feature Construction
    - Combining features (see discriminative frequent patterns)

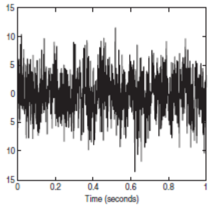
    e.g. Dividing mass by volume to get density
  - Mapping data to new space (see data reduction)

    e.g. Fourier transformation, wavelet transformation, manifold approaches

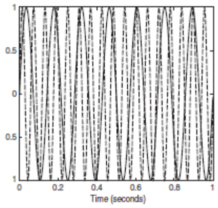
48

SJSU SAN JOSÉ STATE UNIVERSITY

Mapping Data to New Space

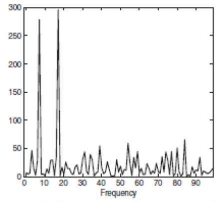


=



\+ Noise

Fourier Transform



SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- Correlation & Covariance Analyses for Numerical and Categorical Data
- Different data transformation techniques
- PCA & SVD