German University in Cairo

Spring 2019

CSEN1083: Data Mining

Data (2)

Seif Eldawlatly

Data Quality

- Data mining applications are often applied to data that was collected for another purpose, or for future, but unspecified applications

- For that reason, data mining cannot usually take advantage of the significant benefits of "addressing quality issues at the source."

- Data mining focuses on

  (1) the detection and correction of data quality problems

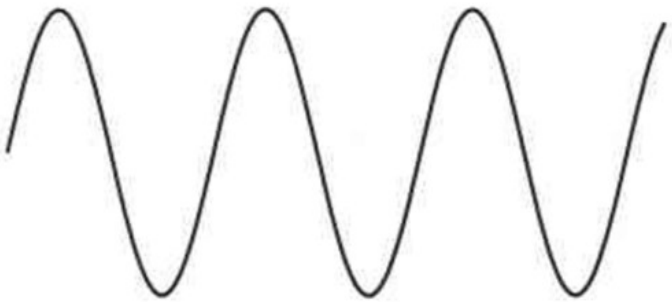
  (2) the use of algorithms that can tolerate poor data quality

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

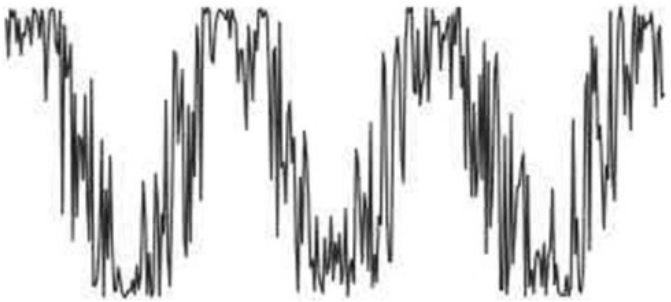
Data Quality: Noise and Artifacts

- There may be problems due to:
  - Human error
  - Limitations of measuring devices
  - Flaws in the data collection process
  - Values or even entire data objects may be missing

- Noise is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects



Signal



Signal + Noise

- Deterministic distortions of the data are often referred to as artifacts

Data Quality: Noise and Artifacts

- Precision: The closeness of repeated measurements of the same quantity to one another. Measured by the standard deviation of a set of value

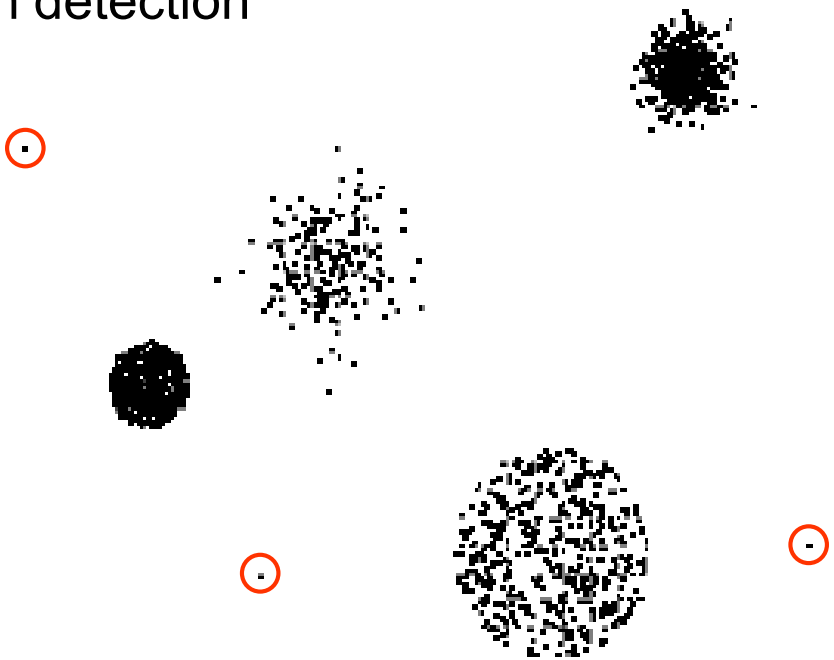
- Bias: A systematic variation of measurements from the quantity being measured. Measured by taking the difference between the mean of the set of values and the known value of the quantity being measured

- Accuracy: The closeness of measurements to the true value of the quantity being measured

- Example: Suppose that we have a standard laboratory weight with a mass of 1g. We weigh the mass five times, and obtain the following five values: {1.015, 0.990, 1.013, 1.001, 0.986}
  - The mean of these values is 1.001, and hence, the bias is 0.001
  - The precision, as measured by the standard deviation, is 0.013

Data Quality: Outliers

- Outliers are data objects with characteristics that are considerably different than 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
    - Credit card fraud
    - Intrusion detection



Data Quality: Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Example:

Case	Attributes			Decision
	Temperature	Headache	Nausea	Flu
1	high	?	no	yes
2	very_high	yes	yes	yes
3	?	no	no	no
4	high	yes	yes	yes
5	high	?	yes	no
6	normal	yes	no	no
7	normal	no	yes	no
8	?	yes	?	yes

Data Quality: Missing Values

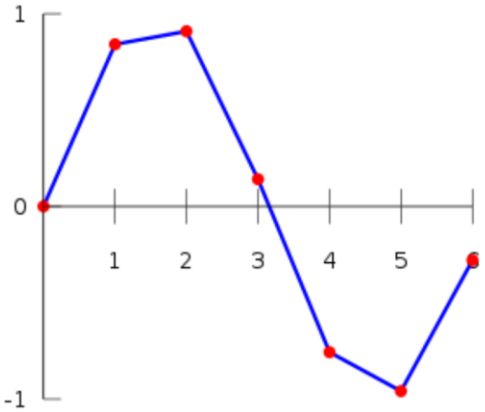
- Handling Missing Values:

  (1) Eliminate Data Objects or Attributes: A simple and effective strategy is to eliminate objects or attributes with missing values

  – This should be done with caution since the eliminated attributes may be the ones that are critical to the analysis

  (2) Estimate Missing Values: Sometimes missing data can be reliably estimated

  – Example: Consider a time series that changes in a smooth fashion, but has a few, widely scattered missing values. The missing values can be interpolated



Data Quality: Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
- Examples: Same person with multiple email addresses
- Data Cleaning: Process of dealing with duplicate data issues

Data Preprocessing

- We address the issue of which preprocessing steps should be applied to make the data more suitable for data mining

- We will discuss the following topics:
  - Aggregation
  - Sampling
  - Dimensionality reduction
  - Feature subset selection
  - Feature creation
  - Discretization and binarization
  - Variable transformation

Data Preprocessing: Aggregation

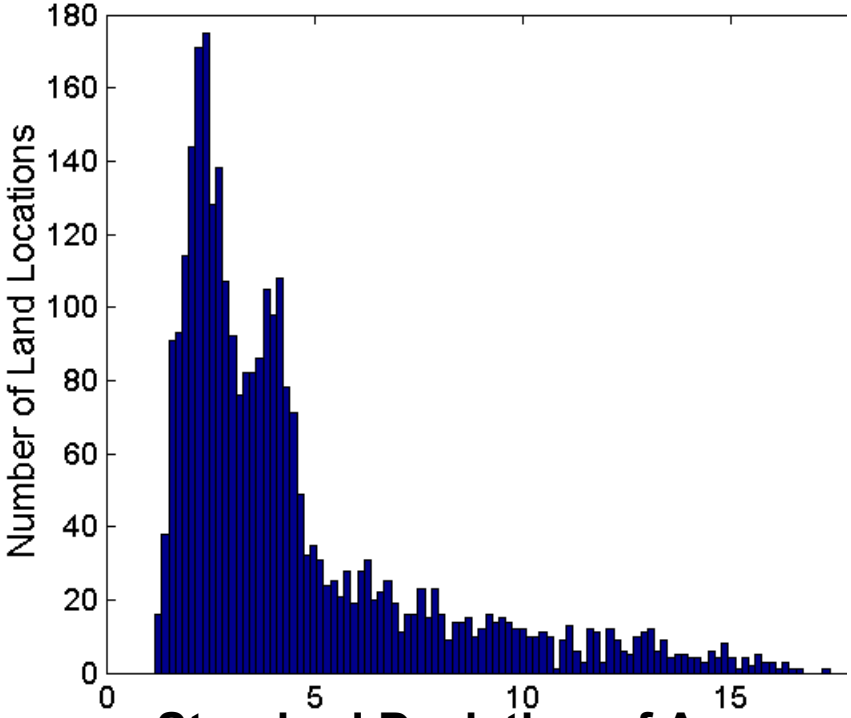
- The combining of two or more objects into a single object

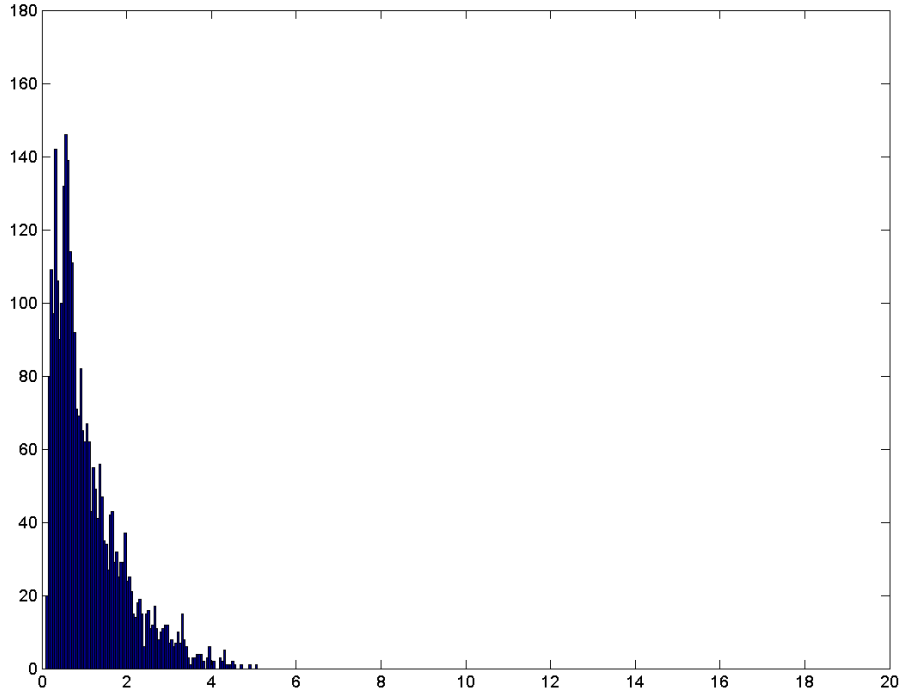
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More “stable” data
    - Aggregated data tends to have less variability

Data Preprocessing: Aggregation

- Example: Based on precipitation in Australia from the period 1982 to 1993

- Histogram for the standard deviation of average monthly precipitation and the average yearly precipitation





Data Preprocessing: Sampling

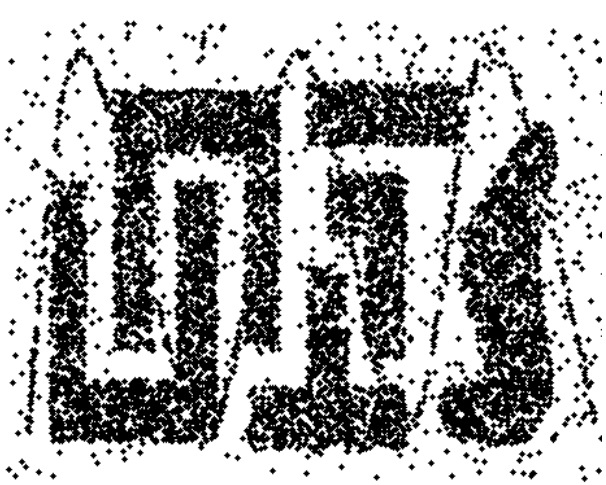
- Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed
- Statisticians use sampling because obtaining the entire set of data of interest is too expensive or time consuming
- Data miners sample because it is too expensive or time consuming to process all the data
- Using a sample will work almost as well as using the entire data set if the sample is representative
- Example: If the mean of the data objects is the property of interest, then a sample is representative if it has a mean that is close to that of the original data

Data Preprocessing: Sampling

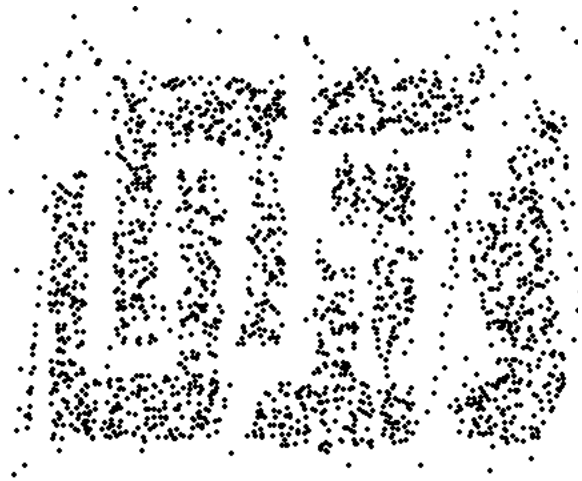
- Simple Random Sampling: There is an equal probability of selecting any particular item
  - Sampling without replacement: As each item is selected, it is removed from the population
  - Sampling with replacement: Objects are not removed from the population as they are selected for the sample.
- Stratified sampling: Split the data into several partitions; then draw random samples from each partition

Data Preprocessing: Sampling

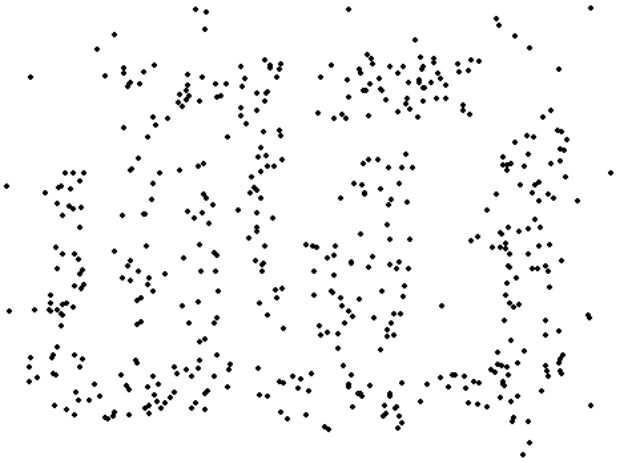
- Once a sampling technique has been selected, it is still necessary to choose the sample size



8000 points



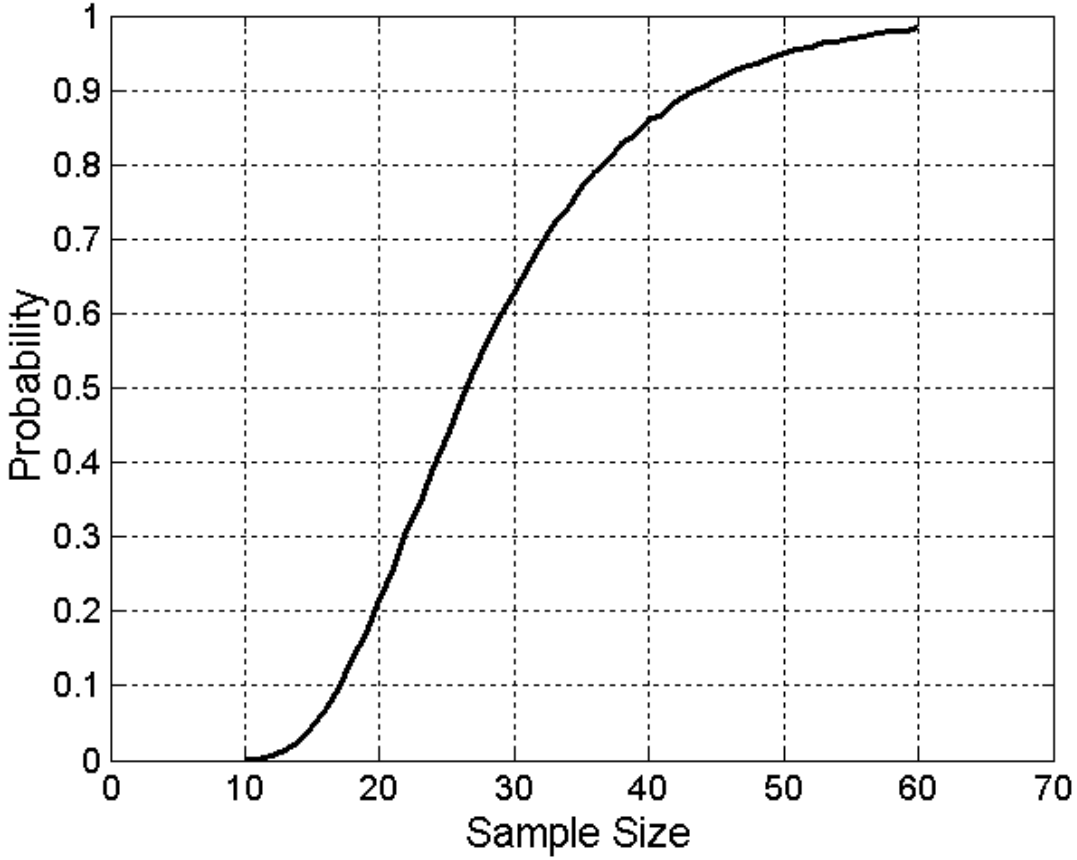
2000 Points



500 Points

Data Preprocessing: Sampling

- Example: What sample size is necessary to get at least one object from each of 10 equal-sized groups?

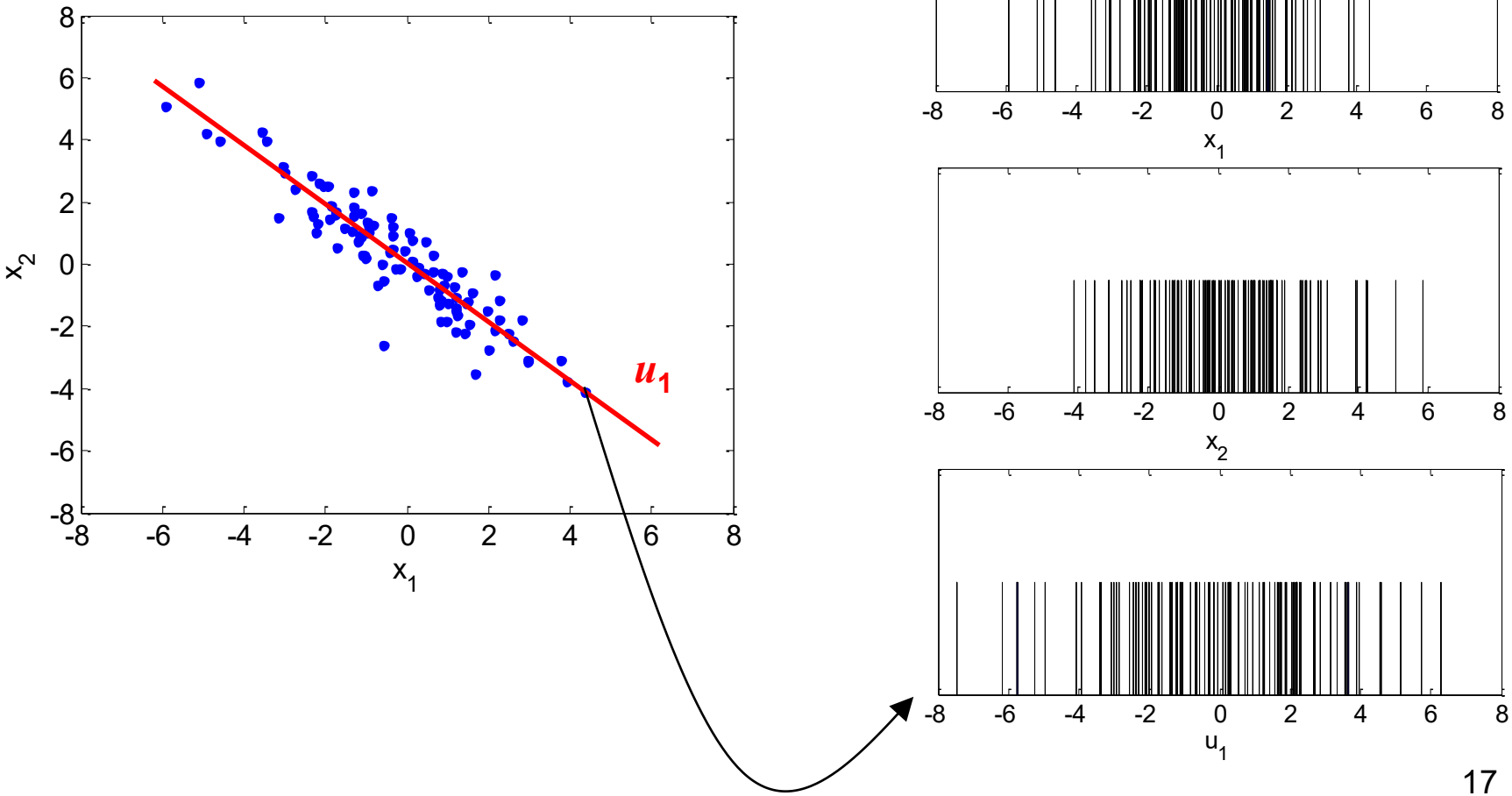


Data Preprocessing: Dimensionality Reduction

- Datasets can have a large number of features
- Example: A set of documents, where each document is represented by a vector whose components are the frequencies with which each word occurs in the document
- The benefit of dimensionality reduction is that many data mining algorithms work better if the dimensionality is lower
- This is partly because dimensionality reduction can eliminate irrelevant features and reduce noise
- The amount of time and memory required by the data mining algorithm is reduced with a reduction in dimensionality

Data Preprocessing: Dimensionality Reduction

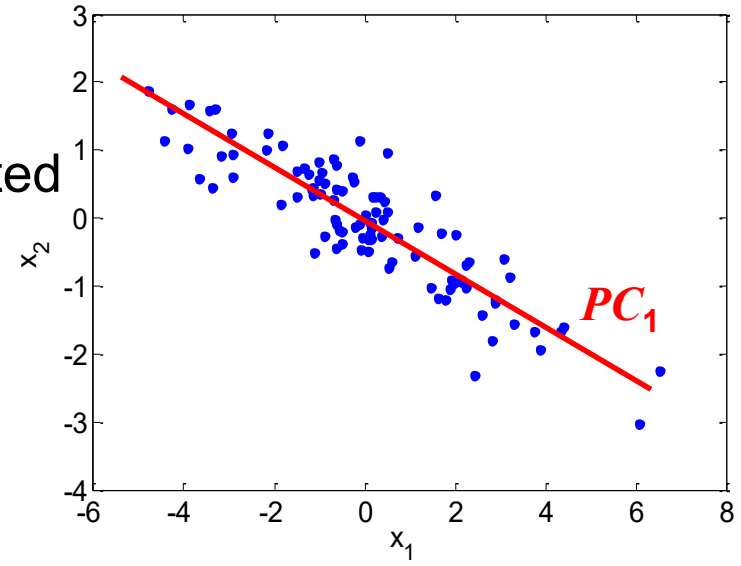
- Projecting 2-dimensions to 1-dimenison: The goal is to do such projection while preserving the properties of the points



Principal Component Analysis (PCA)

- Identifies principal components such that the variance of the data is maximized

- Let D be the dimensionality of the data and M be the dimensionality of the projected space, where $M < D$



- The M principal components are orthonormal

- Consider the case M = 1, and let the direction of the projected space be $\mathbf{u}_1$ where $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$

Principal Component Analysis (PCA)

- Each data point $\mathbf{x}_n$ is then projected onto the new principal component space to $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n$
- Let the mean of the data points be $\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$
- The mean of the projected data is then $\mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}$
- The covariance of the data points is $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}$
- The variance of the projected data is then $\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$

Principal Component Analysis (PCA)

- The objective of PCA is to maximize the projected variance $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ subject to the constraint $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$

- PCA can be formulated then as maximizing the following objective function

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right)$$

- Taking derivative w.r.t. $\mathbf{u}_1$ and equate with zero

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

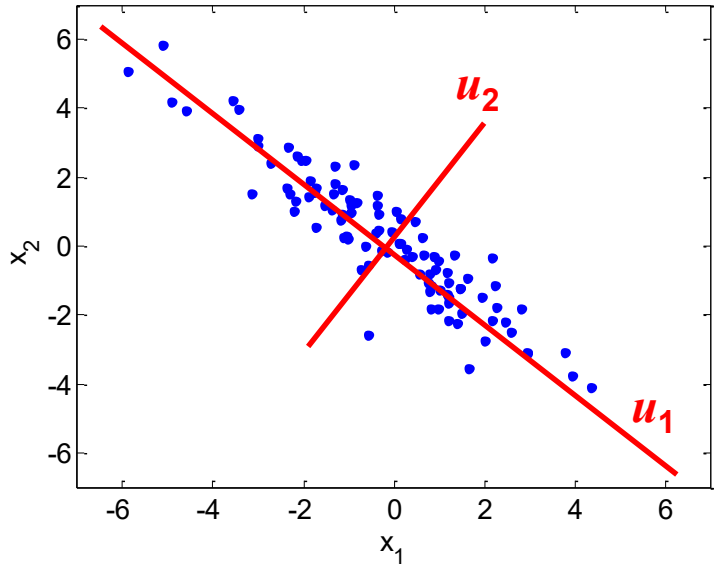
- Multiply by $\mathbf{u}_1^{\mathrm{T}}$ from the left

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1$$

- Therefore, maximizing the objective function is equivalent to finding the eigenvector equivalent to the largest eigenvalue

Principal Component Analysis (PCA)

- Example



$$\lambda_1 = 7.58 \qquad \lambda_2 = 0.29$$

$$u_1 = \begin{bmatrix} 0.7075 \\ -0.7067 \end{bmatrix} \qquad u_2 = \begin{bmatrix} 0.7067 \\ 0.7075 \end{bmatrix}$$