Introduction to Machine Learning Applications

Spring 2023

Data preprocessing

Minor Gordon

gordom6@rpi.edu

Rensselaer

Main steps of data preprocessing

- Aggregation
- Sampling
- Dimensionality reduction (future lecture)
- Feature subset selection
- Feature creation
- Discretization and binarization
- Attribute transformation

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

Aggregation example

Date	Value
01/10/2020	10
01/27/2020	2
02/10/2020	4
02/19/2020	13
03/05/2020	19
03/21/2020	11
04/10/2020	15
04/16/2020	19
05/03/2020	8
05/18/2020	10
05/31/2020	7

Aggregate using sum (or any other metric that fits the problem)

Month	Value
January 2020	12
February 2020	17
March 2020	30
April 2020	34
May 2020	25

Sampling

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

Sampling

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

Types of 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.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Random sampling example

Date	Value
01/10/2020	10
01/27/2020	2
02/10/2020	4
02/19/2020	13
03/05/2020	19
03/21/2020	11
04/10/2020	15
04/16/2020	19
05/03/2020	8
05/18/2020	10
05/31/2020	7

Random sampling (n=3)

Date	Value
02/10/2020	4
05/18/2020	10
01/10/2020	10
04/16/2020	19
05/03/2020	8

Stratified sampling example

Date	Value
01/10/2020	10
01/27/2020	2
02/10/2020	4
02/19/2020	13
03/05/2020	19
03/21/2020	11
04/10/2020	15
04/16/2020	19
05/03/2020	8
05/18/2020	10
05/31/2020	7

Bin-based sampling

Date	Value
01/10/2020	10
02/19/2020	13
03/21/2020	11
04/16/2020	19
05/03/2020	8

Feature subset selection

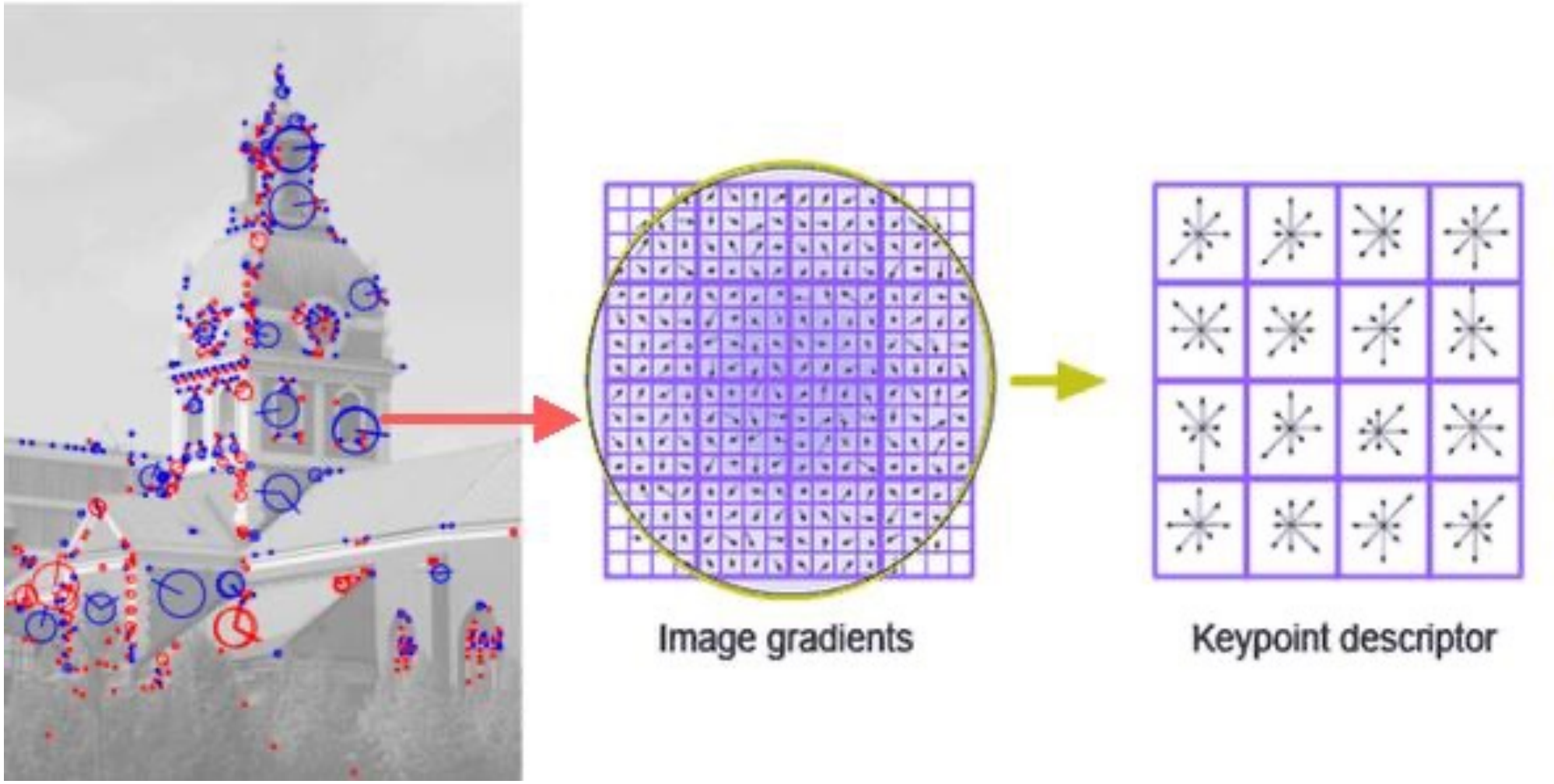
- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the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

Feature creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Feature creation example: SIFT features



Credits: https://www.codeproject.com/Articles/619039/Bag-of-Features-Descriptor-on-SIFT-Features-with-O

Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values

Discretization example

Date	Value
01/10/2020	1.354
01/27/2020	1.83
02/10/2020	2.63
02/19/2020	9.242
03/05/2020	6.43
03/21/2020	9.23
04/10/2020	1.32
04/16/2020	1.756
05/03/2020	0.344
05/18/2020	3.33
05/31/2020	5.014

Assuming the range of value is [0,10) continuous

Assume [0,6): label1
[6,10): label2

Date	Value
01/10/2020	Label1
01/27/2020	Label1
02/10/2020	Label1
02/19/2020	Label2
03/05/2020	Label2
03/21/2020	Label2
04/10/2020	Label1
04/16/2020	Label1
05/03/2020	Label1
05/18/2020	Label1
05/31/2020	Label2

Binarization

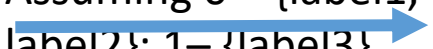
- Binarization maps a continuous or categorical attribute into one or more binary variables

- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Binarization example

Date	Value
01/10/2020	Label1
01/27/2020	Label1
02/10/2020	Label3
02/19/2020	Label2
03/05/2020	Label2
03/21/2020	Label2
04/10/2020	Label1
04/16/2020	Label3
05/03/2020	Label1
05/18/2020	Label3
05/31/2020	Label2

Assuming 0 – {label1, label2}; 1– {label3}



Date	Value
01/10/2020	0
01/27/2020	0
02/10/2020	1
02/19/2020	0
03/05/2020	0
03/21/2020	0
04/10/2020	0
04/16/2020	1
05/03/2020	0
05/18/2020	1
05/31/2020	0

Attribute transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value 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 attributes in terms of frequency 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

Attribute transformation using normalization

Original data = [0.5, 1.0, 0.5]

Computation = [0.5/(0.5+1.0+0.5),1.0/(0.5+1.0+0.5),0.5/(0.5+1.0+0.5)]

= [0.5/2.0, 1.0/2.0, 0.5/2.0]

Normalized data = [0.25, 0.5, 0.25] – sum of the list is 1.