Introduction to Machine Learning Applications

Spring 2021

Lecture-5

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Today’s agenda

- Homework-2 discussion
  - Questions
  - How to evaluate
- Overview of Machine Learning
- Data and its characteristics
  - Visualizations with Python

Machine Learning

According to Tom Mitchell (1998):

Machine Learning is the study of algorithms that

- improve their performance P
- at some task T
- with experience E

Well-defined learning task: <P, T, E>

Learning to detect objects in images

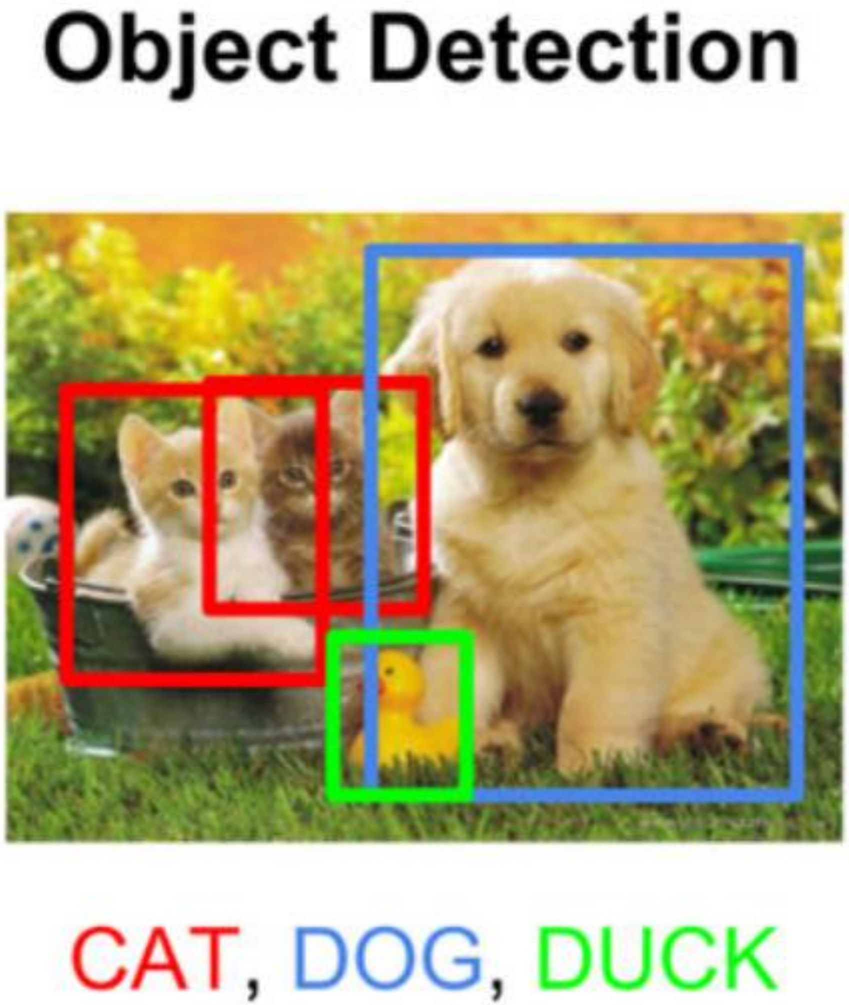


Image src: datacamp.com

Learning to classify text documents

Movie Reviews

http://www.rottentomatoes.com

http://www.cs.cornell.edu/people/pabo/movie-review-data/

Negative	Postive
most of the $\boxed{\text{problems}}$ with the film don't derive from the screenplay , but rather the $\boxed{\text{mediocre}}$ performances by most of the actors involved	the film provides some great $\boxed{\text{insight}}$ into the neurotic mindset of all comics -- even those who have reached the absolute $\boxed{\text{top}}$ of the game .

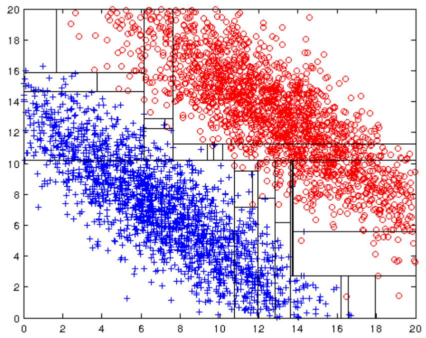
Source: datacamp.com

Learning to predict/classify

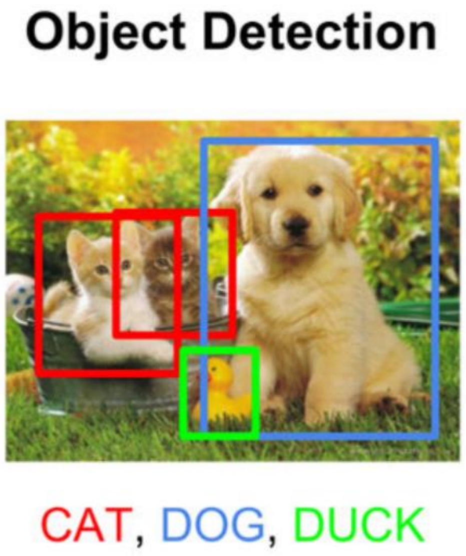
Image src: datacamp.com

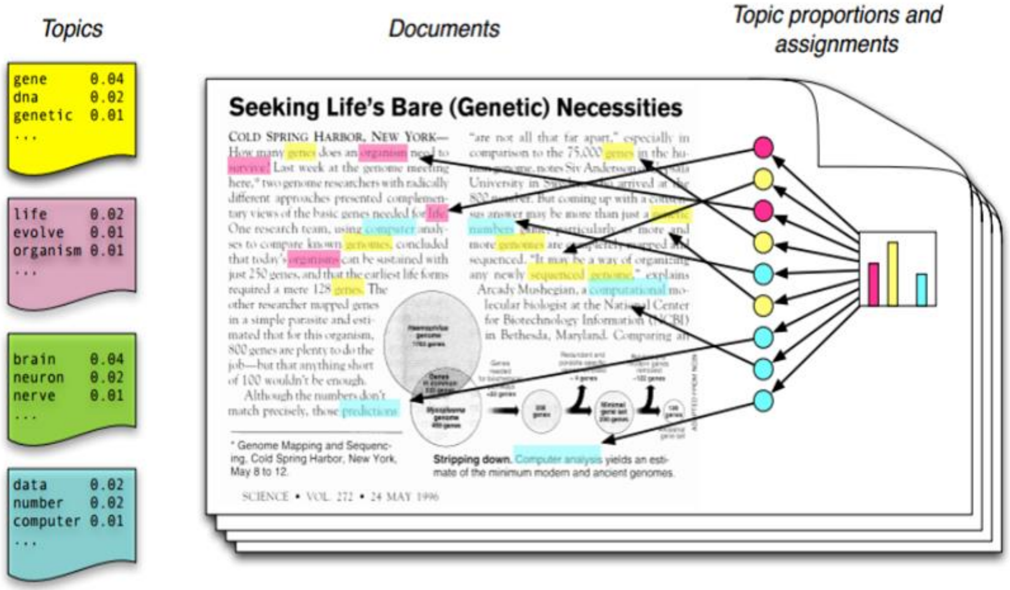
Machine Learning

- Supervised learning
- Unsupervised learning
- Bayesian networks
- Hidden markov models
- Reinforcement learning
- …..



Detecting boundaries
src: Kumar et al.





Text analysis using LDA

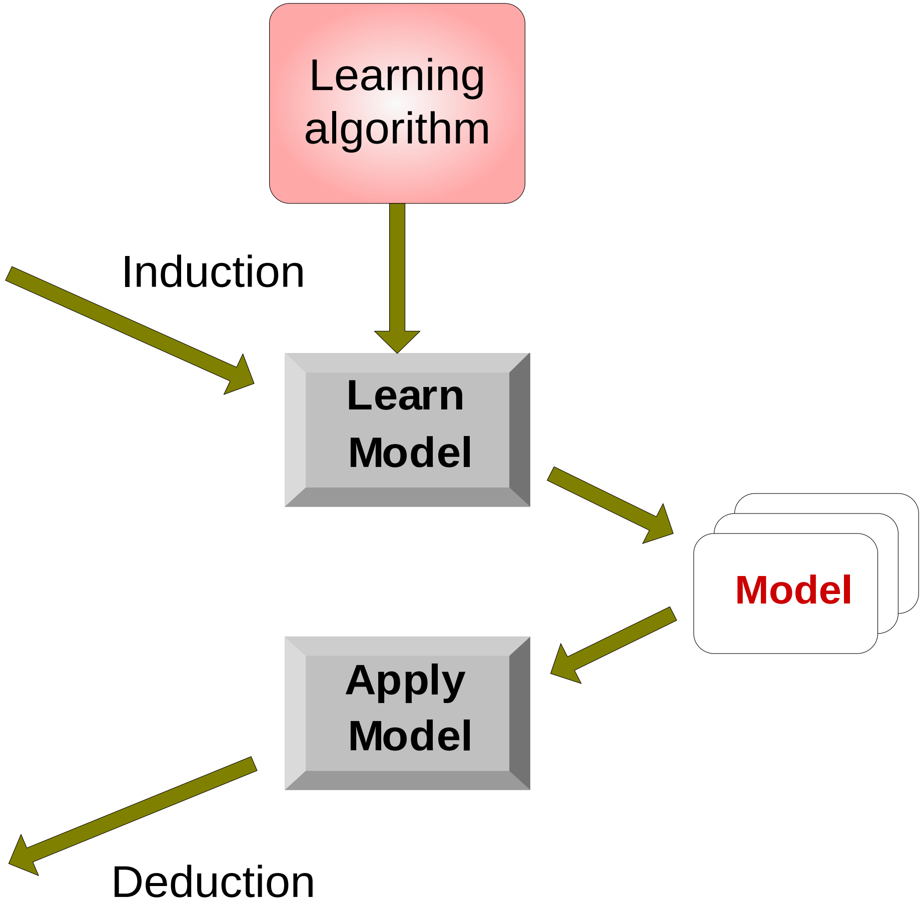
Classification

- Given a collection of records or transactions – training data:
  - Each record is expressed as a tuple – (*x*, *y*) where x is the attribute set and y is the class label
  - *x* – attribute, independent variable, input
  - *y* – class label, dependent variable, output
- Task:
  - Build a model that maps each attribute set *x* to the class label *y*

Classification Model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

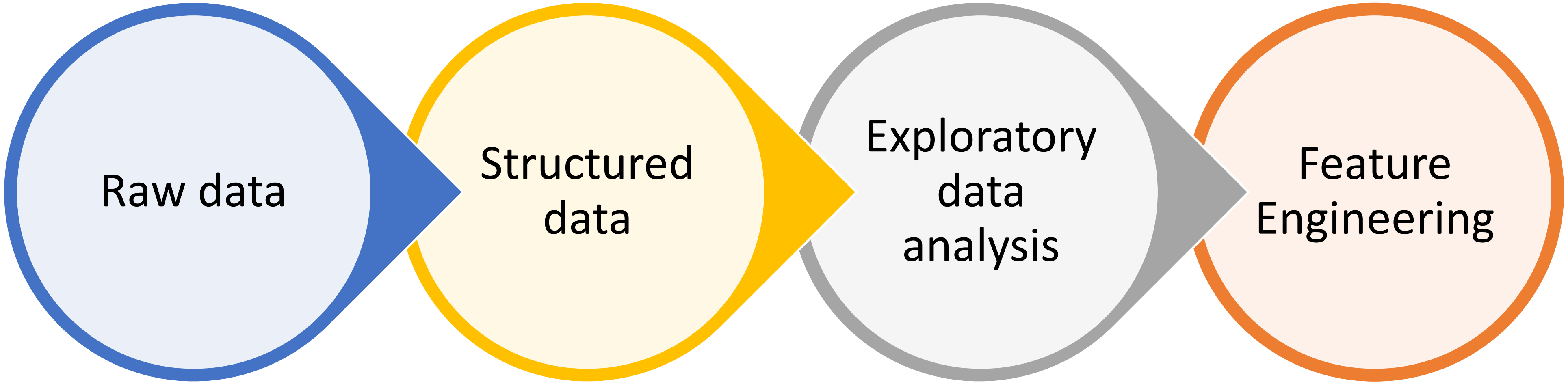
Test Set

Src: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Clustering

- Main aim is to segment data into meaningful segments or detect patterns
- There is no target (outcome) variable to predict or classify
- Hence, we don’t have a model to train using training data like in Classification

Snapshot of data preprocessing



What is data?

- Collection of **data objects** and their **attributes**

According to Tan et al.,

- An **attribute** is a property or characteristic of an object
  - Also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Also known as tuple, record, point, case, sample, etc.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

More views of data

- Data may have parts
- The different parts of data may have relationships
- More generally, data may have structure
- Data can be incomplete

Attribute values

- Attribute values are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: Height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

Types of Attributes

- Nominal
  - Examples: ID numbers, zip codes, eye color
- Ordinal
  - Examples: Rankings (expertise level on a scale of 1-10), grades, height {tall, medium, short}
- Interval
  - Examples: Calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - Examples: Temperature in Kelvin, length, time, counts

Discrete and Continuous attributes

- Discrete Attribute:
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute:
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Important characteristics of data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Main steps of data preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- 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

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

Curse of dimensionality

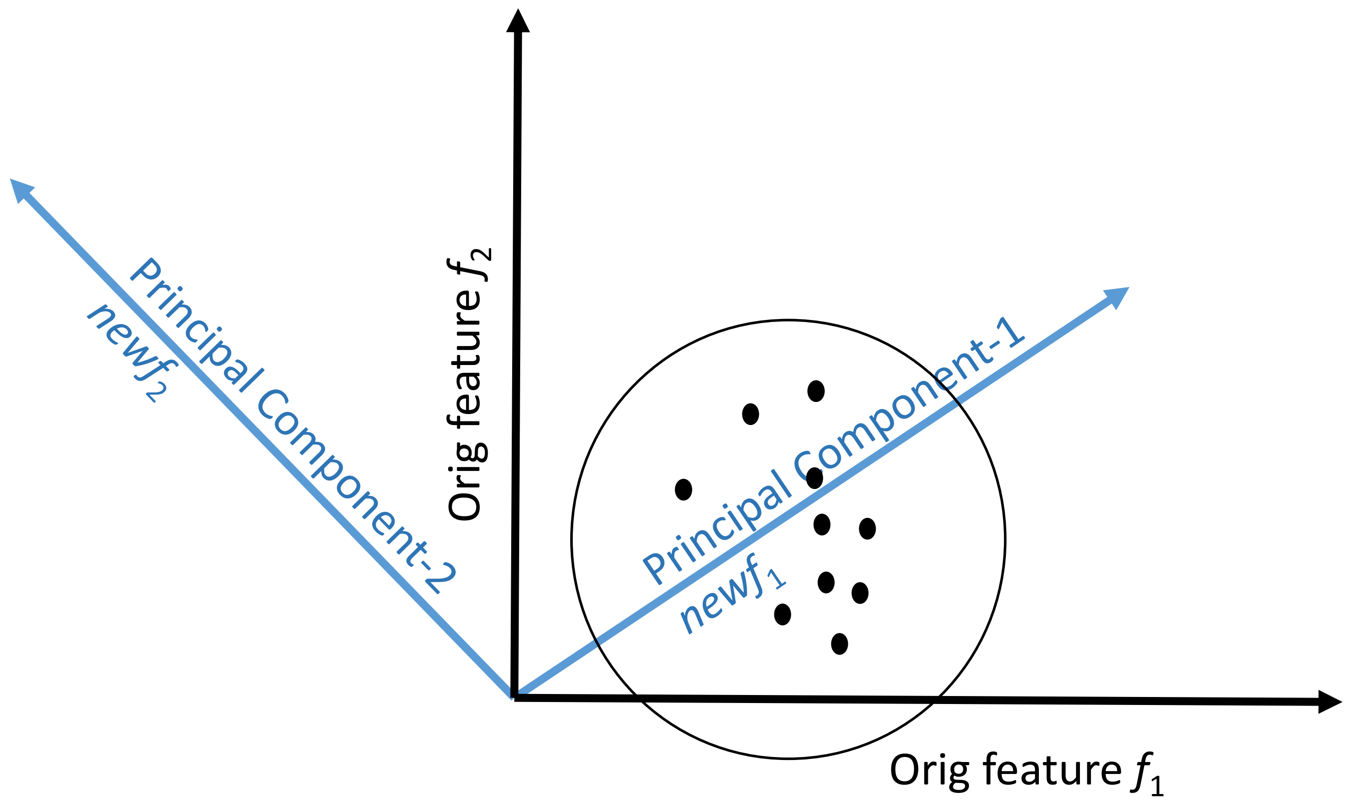
When dimensionality increases, data becomes increasingly sparse in the space that it occupies

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - 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 Example



Feature subset Selection

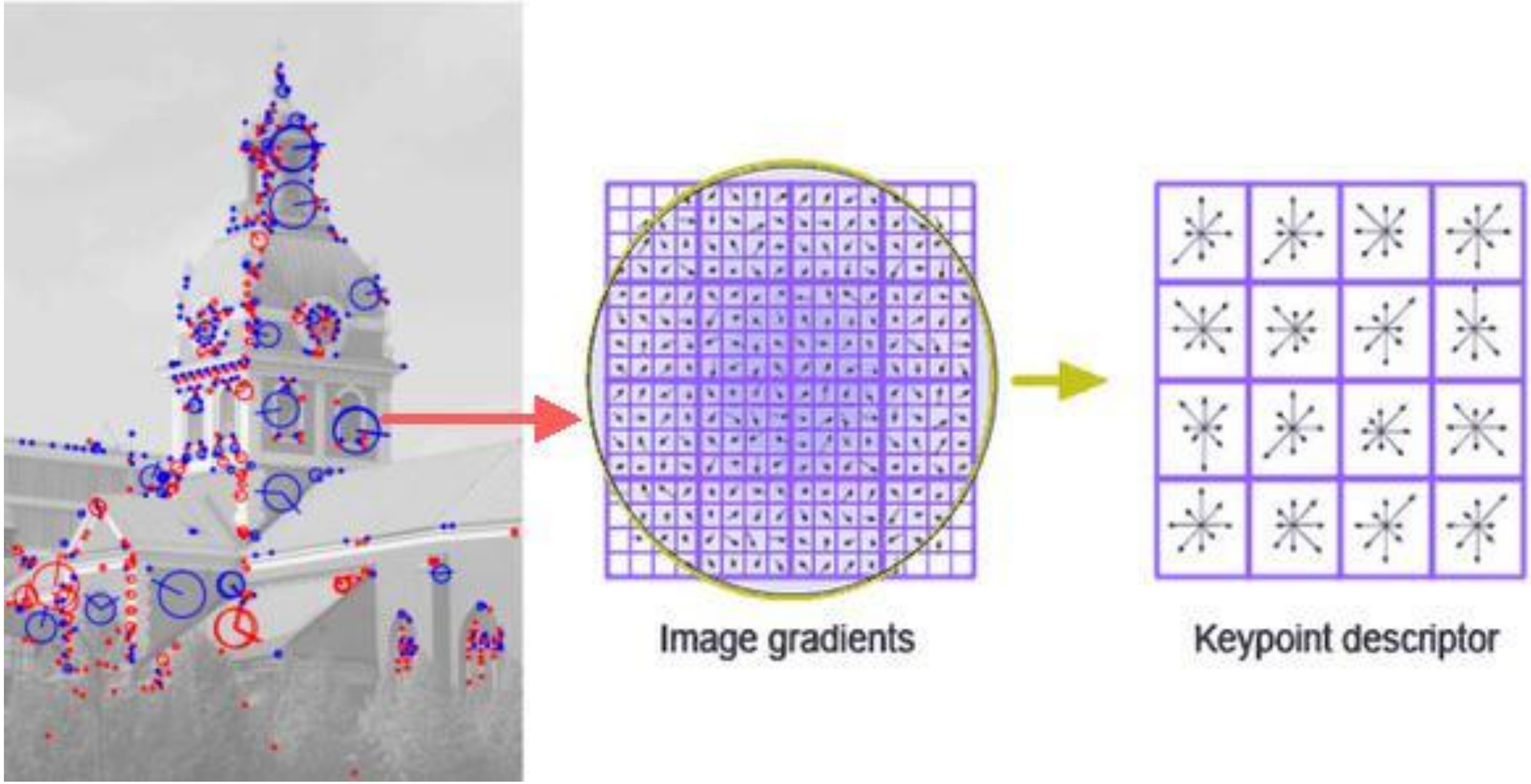
- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the 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.

Data Manipulation and Visualization using Seaborn

- Python library to generate graphs that provide a lot of good insights
- Jupyter notebook

- Next lecture focuses on:
  - Handling textual data especially crawling online web content
  - More examples to process data and visualize the data