GSÜSEM
GALATASARAY ÜNİVERSİTESİ
Sürekli Eğitim Uygulama ve Araştırma Merkezi

Descriptive Data Analysis and Preprocessing

Ar. Gör. Pınar Uluer
Ar. Gör. Merve Ünlü

29/02/2020

Outline

1 Data Preprocessing

2 Data Cleaning
- How to Handle Missing Data?
- How to Handle Noisy Data?

3 Data Integration

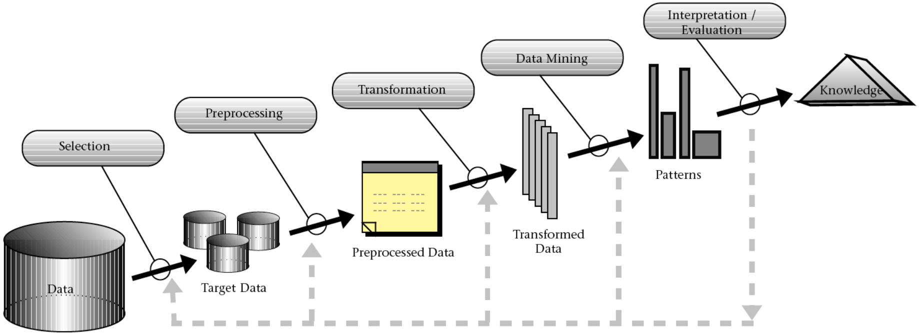
4 Data Transformation
- Normalization
- Discretization
- Data Type Conversion

5 Data Reduction
- Numerosity Reduction
- Dimensionality Reduction

Part 2 - Data Preprocessing

GSÜSEM
GALATASARAY ÜNİVERSİTESİ
Sürekli Eğitim Uygulama ve Araştırma Merkezi

Introduction



Knowledge **D**iscovery in **D**atabases Process (1996)

Introduction

Data Preprocessing

Data mining technique for transforming raw data into an understandable format

Why do we need preprocessing?

- Real-world data is generally incomplete, noisy, inconsistent
- Some algorithms can not process all types of data

Steps to be taken in data preprocessing (not ordered and some of them might be side-stepped!):

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data Quality

Data Quality

The fit for its intended use in operations, decision making and planning

In short:

Poor-quality data $\rightarrow$ inaccurate reporting and ill-conceived strategies
High-quality data $\rightarrow$ high-quality results

Data Quality Dimensions

A set of criteria used to measure the quality of data

Data Quality Dimensions

Data Quality Dimensions - Examples

- Accuracy: Address of an employee in the employee database is the real address ✓
- Completeness: A customer's first name and last name are mandatory but middle name is optional ✓
- Consistency: Employee status is terminated but pay status is active X
- Timeliness: Credit system checking realtime on the credit card account activity ✓
- Integrity: In a customer database, there should be a valid customer, addresses and relationship between them. If there is an address relationship data without a customer X
- Conformity: All the dates in a database is in the format "dd/mm/yyyy" ✓

Poor Quality or High Quality Data?

CompanyName	BrandName	PrimaryCategory	SubCategory	ChemicalName
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Nighttime Skin Care	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	NULL	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Skin Cleansers	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrohena	Sun-Related	Sunscreen	Titanium dioxide

Poor Quality or High Quality Data?

CompanyName	BrandName	PrimaryCategory	SubCategory	ChemicalName
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Nighttime Skin Care	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	NULL	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Skin Cleansers	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrohena	Sun-Related	Sunscreen	Titanium dioxide

Poor Quality or High Quality Data?

CompanyName	BrandName	PrimaryCategory	SubCategory	ChemicalName
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
Alfalfa Nail Supply, Inc.	5000	Nail	Artificial Nails	Titanium dioxide
	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Anti-Wrinkle/Anti-Aging	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Nighttime Skin Care	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	NULL	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrogena	Skin Care	Skin Cleansers	Titanium dioxide
Johnson & Johnson Consumer Companies	Neutrohena	Sun-Related	Sunscreen	Titanium dioxide

Outline

1. Data Preprocessing
2. Data Cleaning
   - How to Handle Missing Data?
   - How to Handle Noisy Data?
3. Data Integration
4. Data Transformation
5. Data Reduction

Data Cleaning

"Data scientists spend 80% of their time cleaning and manipulating data and only 20% of their time actually analyzing it."

Data Cleaning

The process of detecting and correcting corrupt and inaccurate records from a set

1. It removes major errors and inconsistencies
2. Accessing clean data is quick
3. Fewer errors means better results
4. The ability to map the different functions and what your data is intended to do and where it is coming from your data

Missing Data

Definition

When no data values is stored for the variable in an observation

The reasons why data goes missing:

1. Missing completely at random: Missingness can not be related to any event
   ex: Data acquisition terminated due to fire alarm
2. Missing at random: Missingness is not random, but where missingness can be fully accounted for by variables where there is complete information $\rightarrow$ Biased data!!!
   ex: Men in depression survey, women and their age/weight
3. Missing is not random: the value of the variable that's missing is related to the reason it's missing
   ex: Different pay grades

How to Handle Missing Data?

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Jason	M	English		30000
Michael	M	French	University	45000
John			High School	
Emily	F	English	High School	30000
Brad	M	German	University	
Elizabeth		English		50000

Table: Information about people and their incomes

1. Ignore the tuple

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Michael	M	French	University	45000
Emily	F	English	High School	30000

Table: Data matrix after the rows with the missing values are ignored

2. Fill in the missing values manually

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Jason	M	English		45000
Michael	M	French	University	45000
John	**M**		High School	
Emily	F	English	High School	30000
Brad	M	German	University	
Elizabeth	**F**	English		50000

Table: Data matrix after the rows are filled in by an expert

3. Use a global constant to fill in the missing value

Name	**Gender**	**Language**	**Education**	**Income**
Susan	F	French	University	40000
Jason	M	English	**Education**	45000
Michael	M	French	University	45000
John	M		High School	
Emily	F	English	High School	30000
Brad	M	German	University	
Elizabeth	F	English	**Education**	50000

Table: Data matrix after the global constant “Education” replaced the missing values

4. Use a measure of central tendency for the feature

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Jason	M	English	Education	45000
Michael	M	French	University	45000
John	M	**English**	High School	
Emily	F	English	High School	30000
Brad	M	German	University	
Elizabeth	F	English	Education	50000

Table: Data matrix after the missing values are replaced with the mode of Language column

4. Use a measure of central tendency for the feature

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Jason	M	English	Education	45000
Michael	M	French	University	45000
John	M	English	High School	**30000**
Emily	F	English	High School	30000
Brad	M	German	University	**42500**
Elizabeth	F	English	Education	50000

Table: Data matrix after the missing values are replaced with the mean of the Income column

5. Use an algorithm to predict the most probable value to fill in (such as Maximum Likelihood, Bayesian Estimation, etc.)

Name	Gender	Language	Education	Income
Susan	F	French	University	40000
Jason	M	English	Education	45000
Michael	M	French	University	45000
John	M	English	High School	**38750**
Emily	F	English	High School	30000
Brad	M	German	University	**42000**
Elizabeth	F	English	Education	50000

Table: Data matrix after the missing values are replaced with the most probable income value

Noisy Data

Definition

Unwanted data items, features or records which don't help in explaining the feature itself, or the relationship between feature and target

Reasons of Noise

- Faulty data collection
- Human or computer errors occurring at data entry
- Data transmission errors
- Limited buffer size for coordinating synchronized data transfer
- Inconsistencies in naming conventions or data codes used
- Inconsistent formats for input fields (e.g. date)

Noisy Data - Example

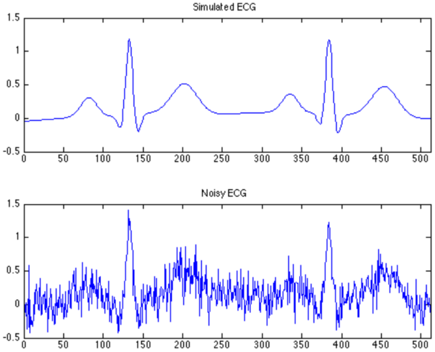


Fig: Noisy and Clean ECG Data

Color	**Weight**	**Class**
red	0.2gr	Positive
green	0.11gr	Negative
ref	0.3gr	**?**
green	**-0.1**	Negative
red	0.25gr	**Negative**
red	**0,34gr**	Positive
green	0.14gr	Negative

Table: Noisy Data Example

Method 1: Binning

Smoothing a sorted data value based on its *neighborhood* $\rightarrow$ the values around it
$\approx$ one dimensional clustering

Ex: Sorted data for price (in dollars):
4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

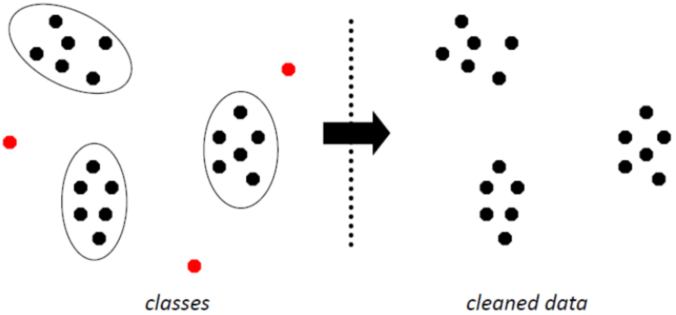
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Method 2: Outlier Analysis

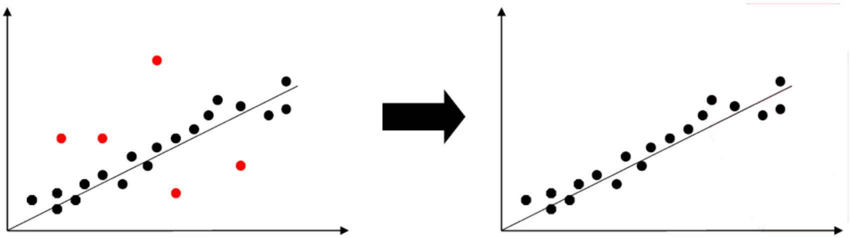
The process of finding data objects with behaviors that are very different from expectation → e.g. Outliers may be detected by clustering



Method 3: Regression

A technique that relates data values to a function such as *linear regression*:

$$y = ax + b$$



Outline

1. Data Preprocessing
2. Data Cleaning
3. Data Integration
4. Data Transformation
5. Data Reduction

Data Integration

Definition

The process of merging data coming from multiple data stores

- Multiple databases, data cubes or data files
- High volume data process
- Complex and fast query processing
- Advanced data summarization and storage
- **Problems:** Entity identification problem, redundancy, detection and resolution of data value conflicts

Outline

1 Data Preprocessing
2 Data Cleaning
3 Data Integration
4 Data Transformation
   - Normalization
   - Discretization
   - Data Type Conversion
5 Data Reduction

Data Transformation

Definition

The process of converting data from one format or structure into another

Why? Different features in the data set may have values in different ranges

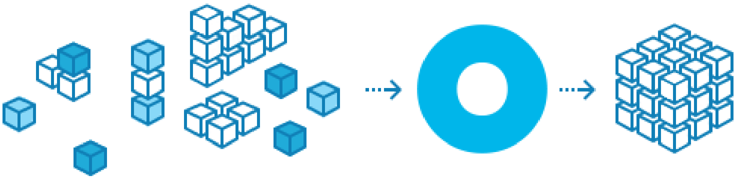


Fig: Image courtesy of `https://www.onedot.com/data-transformation`

Normalization - Example

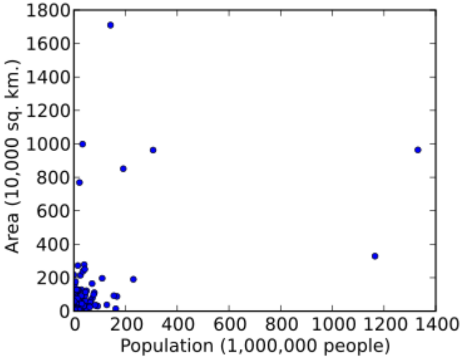


Fig: Raw data

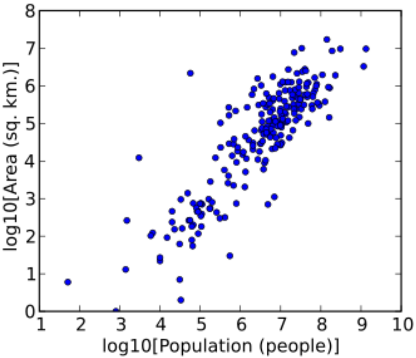


Fig: Log10 normalized data

Normalization

Bringing all the columns into same range

- **Min-Max Normalization** tries to move the values towards the mean of the column

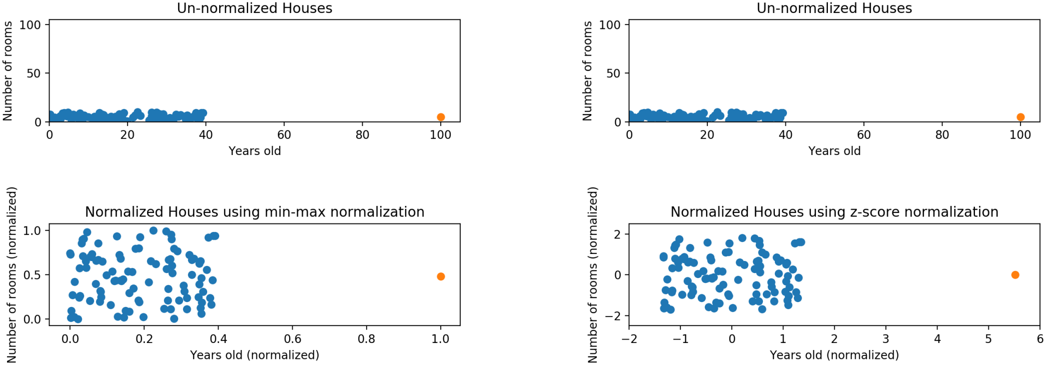
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **z-score Normalization** transforms the data by converting the values to a common scale with an average of zero and a standard deviation of one

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation

Normalization



Normalization

- **Decimal Scaling Normalization** transforms the data by moving the decimal points of values of a feature. The number of decimal points moved depends on the maximum absolute value of feature

$$z = \frac{x}{10^j}$$

where j is the smallest integer such that $\max(|x|) < 1$

Discretization

Definition

Reduce the number of values a continuous variable assumes by grouping them into a number of intervals or bins

Why?

- Some algorithms can only work with discrete data
- Discretization can be used to reduce the data size

Discretization - Example

Before discretization:

Ages: 10, 11, 13, 14, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75

$$\underbrace{10, 11, 13, 14, 17, 19}_{\text{Young}} \quad \underbrace{30, 31, 32, 38, 40, 42}_{\text{Mature}} \quad \underbrace{70, 72, 73, 75}_{\text{Old}}$$

After discretization:

Ages: Young Mature Old

Data Type Conversion

Definition

Converting a feature type into another in the meantime representing the original distribution accurately

- Numerical to Categorical:
  As in discretizing, converting ages into *young, mature, old*
- Categorical to Numerical:
  As in binarization, converting Positive/Negative into 1 and 0 $\rightarrow$ Label encoding

Be careful!
Which algorithm do you use to process your data?
Which type of data does it take as input?

Example - 1

Focus point: Magazine subscription

Client ID	Name	Address	Subcription Date	Magazine
23134	Bemol	Rue du Moulin, Paris	7/10/96	Car
23134	Bemol	Rue du Moulin, Paris	12/5/96	Music
23134	Bemol	Rue du Moulin, Paris	25/7/95	Cartoon
31435	Bodinoz	Rue Verte, Nancy	11/11/11	Cartoon
43342	Airinaire	Rue de la Source, Brest	30/5/95	Sport
25312	Talonion	Rue du Marché, Paris	25/02/98	NULL
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du Moulin, Paris	11/11/11	House

Example - 2

Client ID	Name	Address	Subcription Date	Magazine
23134	Bemol	Rue du Moulin, Paris	7/10/96	Car
23134	Bemol	Rue du Moulin, Paris	12/5/96	Music
23134	Bemol	Rue du Moulin, Paris	25/7/95	Cartoon
31435	Bodinoz	Rue Verte, Nancy	NULL	Cartoon
43342	Airinaire	Rue de la Source, Brest	30/5/95	Sport
~~25312~~	~~Talonion~~	~~Rue du Marché, Paris~~	~~25/02/98~~	~~NULL~~
43241	Manvussa	NULL	14/04/96	Sport
23130	Bemolle	Rue du Moulin, Paris	NULL	House

Example - 3

Client ID	Sport	Cartoon	Car	House	Music
23134	0	1	1	1	1
31435	0	1	0	0	0
43342	1	0	0	0	0
43241	1	0	0	0	0

Example - 4

Customer information data matrix coming from another source

Client ID	Client	Date of Birth	Salary	Owner	Car
23134	Bemol	13/01/50	20 000 $	Yes	Yes
31435	Bodinoz	21/05/70	12 000 $	No	Yes
43342	Airinaire	15/06/63	9 000 $	No	No
43241	Manvussa	27/03/47	15 000 $	No	Yes

Example - 5

Client	Sport	Cartoon	Car	House	Music	Age	Salary	Owner	Paris?	Years
23134	0	1	1	1	1	50	20	Yes	1	4
31435	0	1	0	0	0	30	12	No	0	NULL
43342	1	0	0	0	0	37	9	No	0	5
43241	1	0	0	0	0	53	15	No	NULL	4

Outline

1. Data Preprocessing
2. Data Cleaning
3. Data Integration
4. Data Transformation
5. Data Reduction
   - Numerosity Reduction
   - Dimensionality Reduction

Data Reduction

Motivation

To have a reduced representation of the data set that is much smaller in volume, yet closely maintaining the integrity of the original data

- Numerosity reduction: the process of replacing the original data volume by alternative, smaller forms of data representation
- Dimensionality Reduction: the process of reducing the number of features under consideration

Sampling

Definition

Selecting a subset of the data to be analyzed
The key idea is to have a representative sample of the data

- **Random Sampling**: There is an equal probability of choosing an element from the data set
- **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, an object can be picked several times

Dimensionality Reduction

Why do we need it?

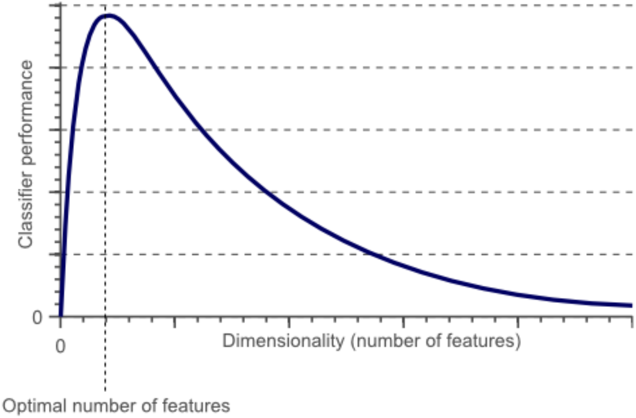
- Space required to store the data is reduced as the number of dimensions comes down
- Less dimensions lead to less computation/training time
- Some algorithms do not perform well with large dimensions
- It takes care of multicollinearity by removing redundant features
- It helps in visualizing data

Curse of Dimensionality

Definition

As the number of features or dimensions grows, the amount of data required for generalization grows exponentially

Hughes Phenomenon: as the number of features increases, the classifier's performance increases as well until it reaches the optimal number of features



Feature Selection

Definition

The process of selecting a subset of relevant features (variables, predictors) to be used in model construction

Why?

- Reduces overfitting
- Improves accuracy
- Reduces training time

- Wrapper Methods: Forward, Backward, Recursive Features Selection
- Filter Methods: Anova, LDA, Chi-Square, Correlation
- Embedded Methods: Ridge Regression, LASSO Regression

Feature Selection - Wrapper Methods

Criteria is "usefulness"

For all subset of features, train a model, check its performance → Add or remove features

(-) Overfitting risk when the number of observations is insufficient
(-) Computation time

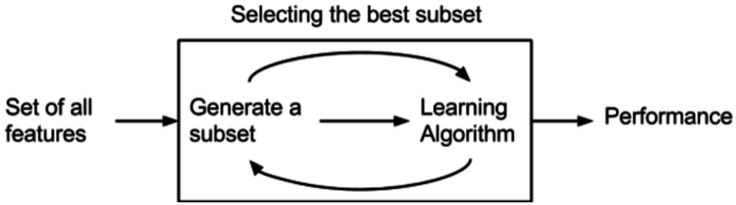


Fig: Wrapper Methods Feature Selection

Feature Selection - Wrapper Methods

- **Forward Selection** starts with no features, at each iteration adding the feature which best improves the model until the addition does not improve the performance
- **Backward Elimination** starts with all features, at each iteration remove the least significant feature which is the worst attribute according to the evaluation metric
- **Recursive Feature Elimination** performs a greedy search to find the best performing feature subset

Feature Selection - Filter Methods

Criteria is "relevance"

The features are selected regardless of any model based on the relation between the features and the variable to predict $\rightarrow$ it suppress the least interesting features

(+) Effective in computation time & robust to overfitting

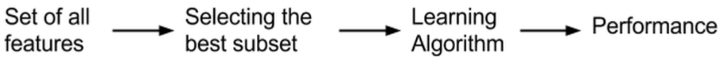


Fig: Filter Methods

Embedded Methods

Combine the advantages of both previous methods
A learning algorithm performing feature selection and classification simultaneously

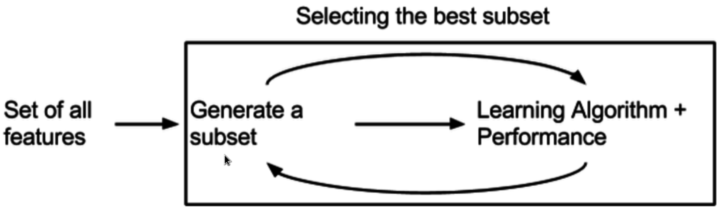


Fig: Embedded Methods

Feature Projection

Motivation

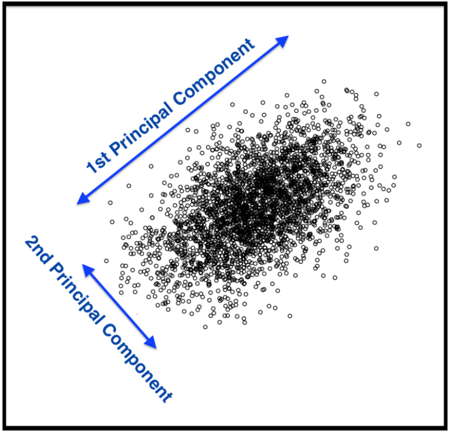
To transform the data in the high-dimensional space to a space of fewer dimensions

We will look closely at:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA)

An unsupervised method to find a low-dimensional representation of the data that retains as much information as possible : **Principal component, what is it?**

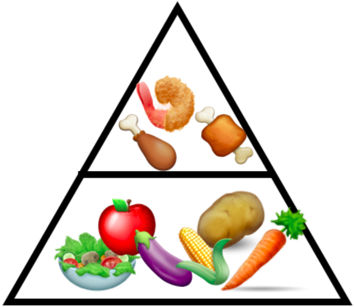


PCA - Example

Problem: Imagine that you are a nutritionist trying to explore the nutritional content of food.
What is the best way to differentiate food items?

$\rightarrow$ By vitamin content?
$\rightarrow$ By protein levels?
$\rightarrow$ Or perhaps a combination of both?

(Example source: Quora)



How to differentiate food items?

1. Vitamin C: present in vegetables, absent in meat
2. (Vitamin C - Fat): Fat present in meat, absent in vegetables (measured in $\neq$ units $\rightarrow$ normalization)
3. (Vitamin C + Fiber) - Fat: Varying amount of Fiber in veggies

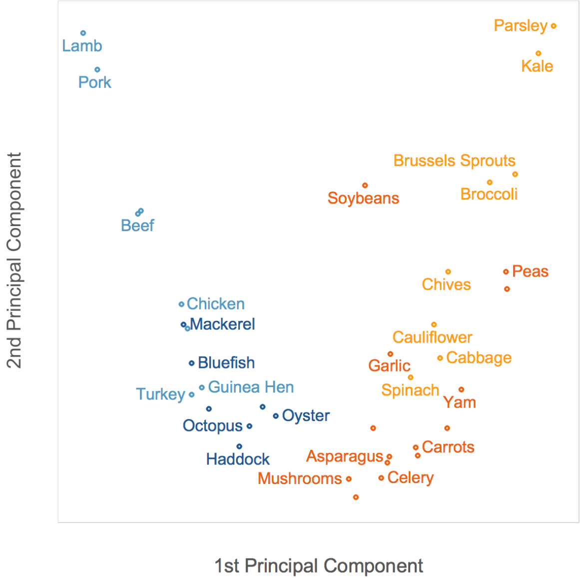
Data from the United States Department of Agriculture, analysis based on 4 nutrition variables:

- Fat,
- Protein,
- Fiber,
- Vitamin C

PCA

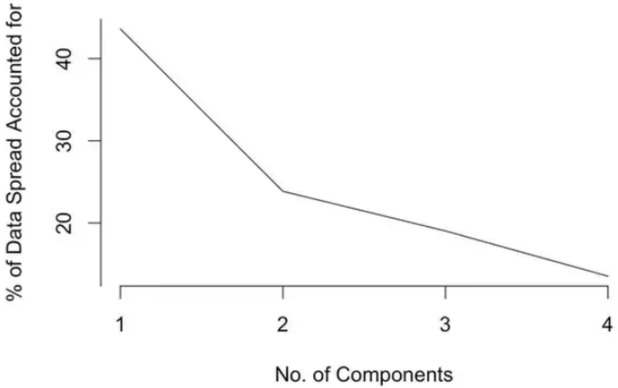
	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22

Fig: Resulting Principal Components after PCA


2nd Principal Component
1st Principal Component
Lamb
Pork
Beef
Chicken
Mackerel
Bluefish
Turkey
Guinea Hen
Octopus
Oyster
Haddock
Parsley
Kale
Brussels Sprouts
Soybeans
Broccoli
Chives
Peas
Cauliflower
Garlic
Cabbage
Spinach
Yam
Carrots
Asparagus
Mushrooms
Celery

PCA - Scree Plot

How many principal components to use? How to decide?



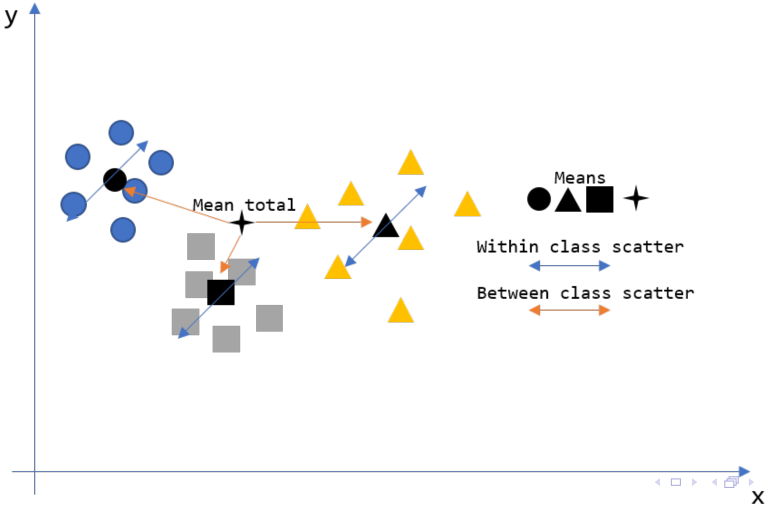
Linear Discriminant Analysis (LDA)

Definition

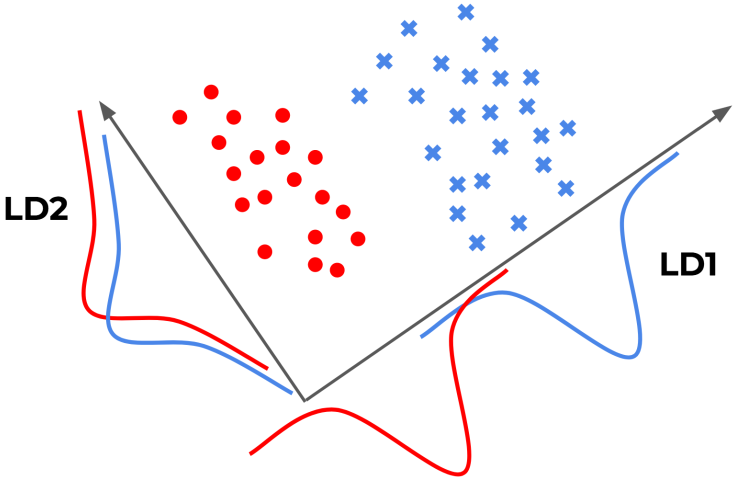
A supervised method to find a low-dimensional representation of the data that retains as much information as possible

1. calculate the distance between the mean of different classes
   $\rightarrow$ between-class variance
2. calculate the distance between the mean and sample of each class
   $\rightarrow$ within class variance
3. construct the lower dimensional space which maximizes the between class variance and minimizes the within class variance

LDA


y
x
Mean total
Means
Within class scatter
Between class scatter

LDA - Example

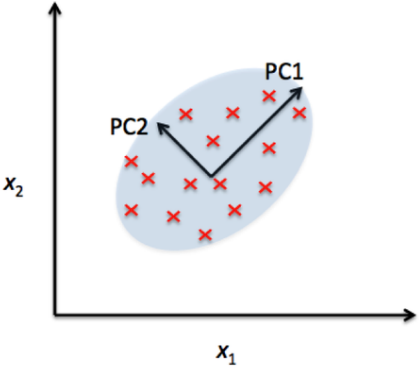


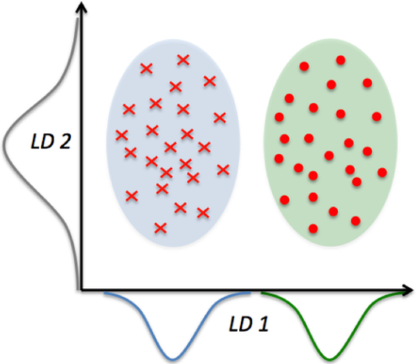
PCA versus LDA

Remarks

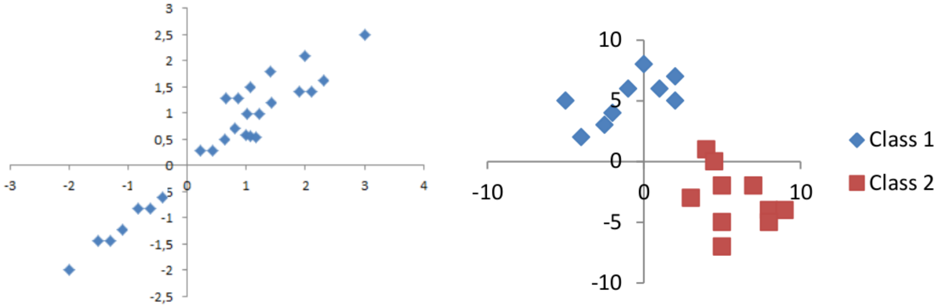
- Both linear transformation techniques
- LDA is a supervised whereas PCA is unsupervised $\rightarrow$ PCA ignores class labels
- PCA performs better in case where number of samples per class is less
- LDA works better with large dataset having multiple classes; class separability is an important factor while reducing dimensionality

PCA versus LDA

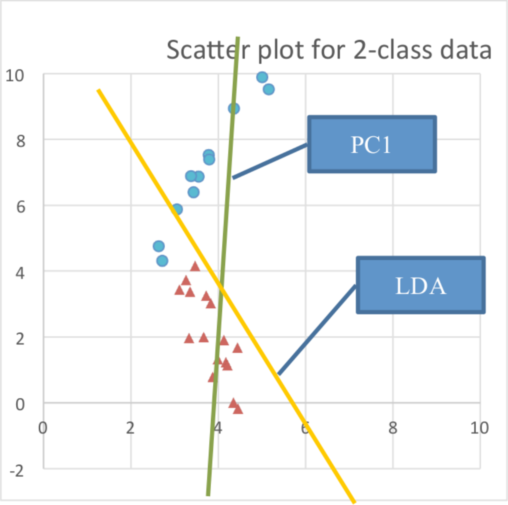




PCA versus LDA



PCA versus LDA



Coming Up Next

Python practice:

- What libraries and methods to use?
- How to preprocess the data we have?
- How to analyze the data?
- How to visualize it?
- Principal component analysis
- Linear discriminant analysis

Any questions so far?