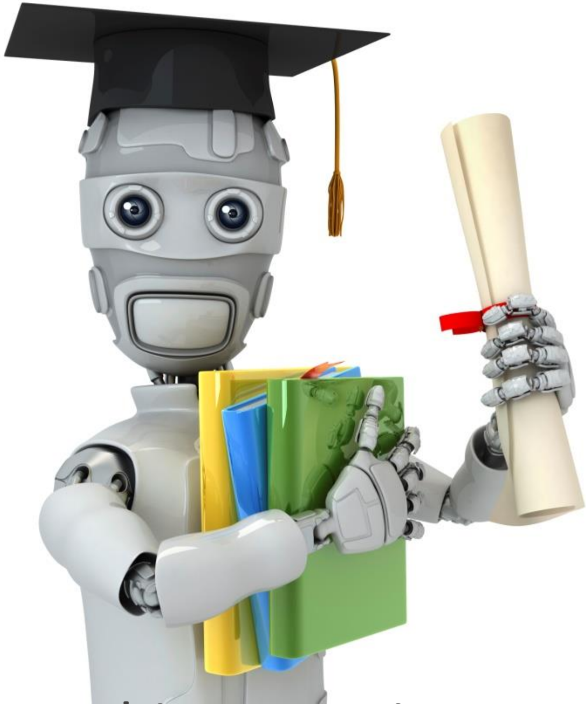


Introduction

Welcome

Machine Learning
Fouad Hadj Selem

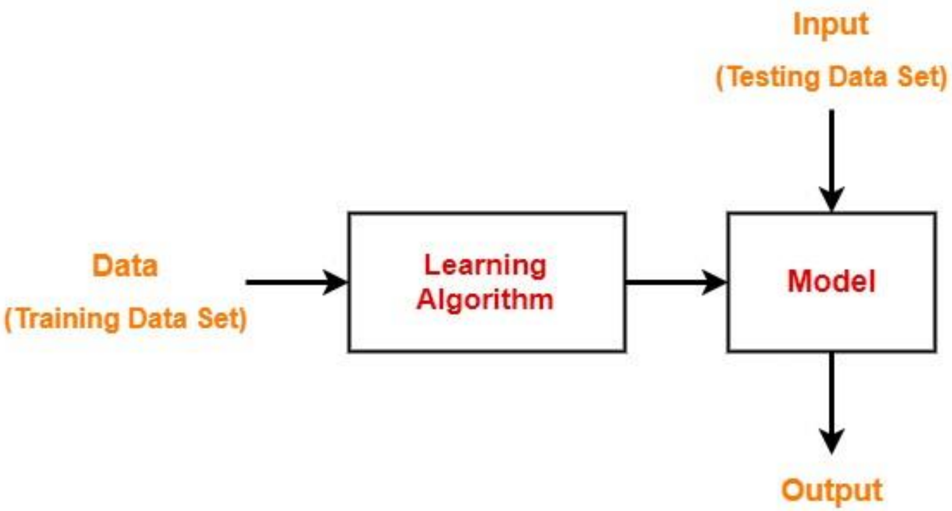
Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Introduction

In machine learning,

- There is a learning algorithm.
- Data called as training data set is fed to the learning algorithm.
- Learning algorithm draws inferences from the training data set.
- It generates a model which is a function that maps input to the output.



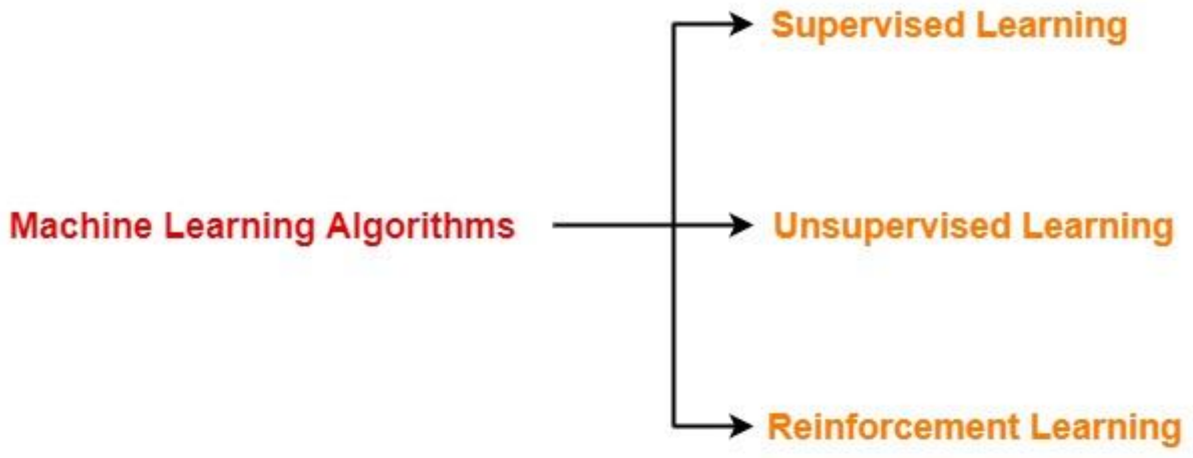
Applications

Some important applications of machine learning are-

- Spam Filtering
- Fraudulent Transactions
- Credit Scoring
- Recommendations
- Robot Navigation

Machine Learning Algorithms

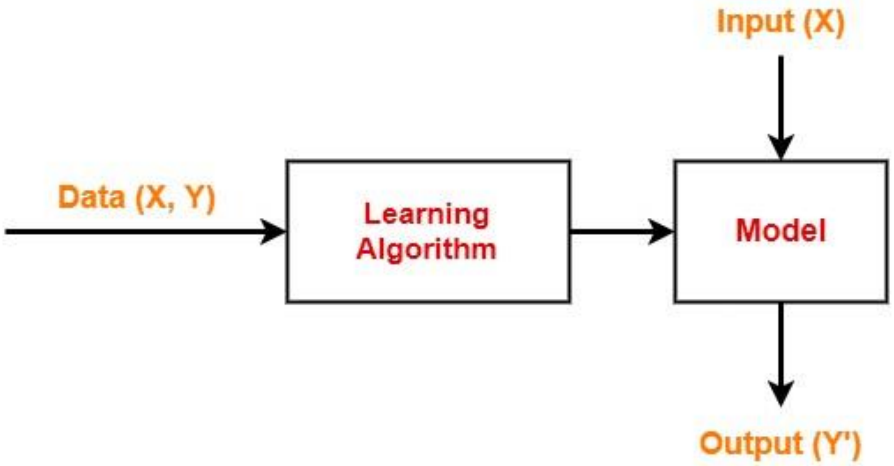
There are three types of machine learning algorithms-



Supervised Learning

In this type of machine learning algorithm,

- The training data set is a labeled data set.
- In other words, the training data set contains the input value (X) and target value (Y).
- The learning algorithm generates a model.
- Then, new data set consisting of only the input value is fed.
- The model then generates the target value based on its learning.



Example of supervised learning

Consider a sample database consisting of two columns where-

- The first column specifies mails.
- The second column specifies whether those emails are spam or not.

Mails (X)	IsSpam (Y)
Mail-1	Yes
Mail-2	No
Mail-3	No
Mail-4	No

In this training data set, emails categorized as spam or not are done by a supervisor's knowledge.

So, it is supervised learning algorithm.

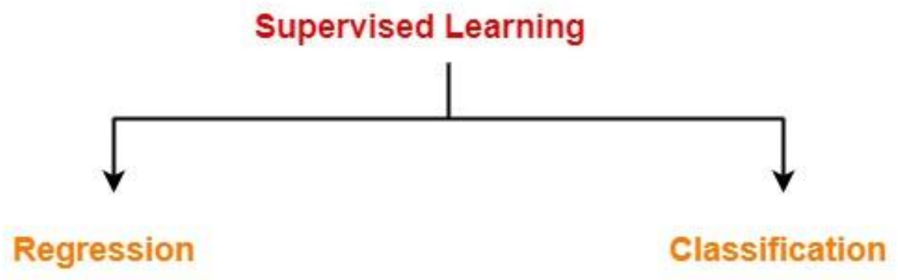
Applications of supervised learning

Some real-life applications are-

- Spam Filtering
- House Price Prediction
- Credit Scoring (high risk or a low risk customer while lending loans by the banks)
- Face Recognition etc

Types of supervised learning

There are two types of supervised learning algorithm-



1. Regression
2. Classification

Types of supervised learning

Regression-

Here,

- The target variable (Y) has continuous value.
- Example- house price prediction

Classification-

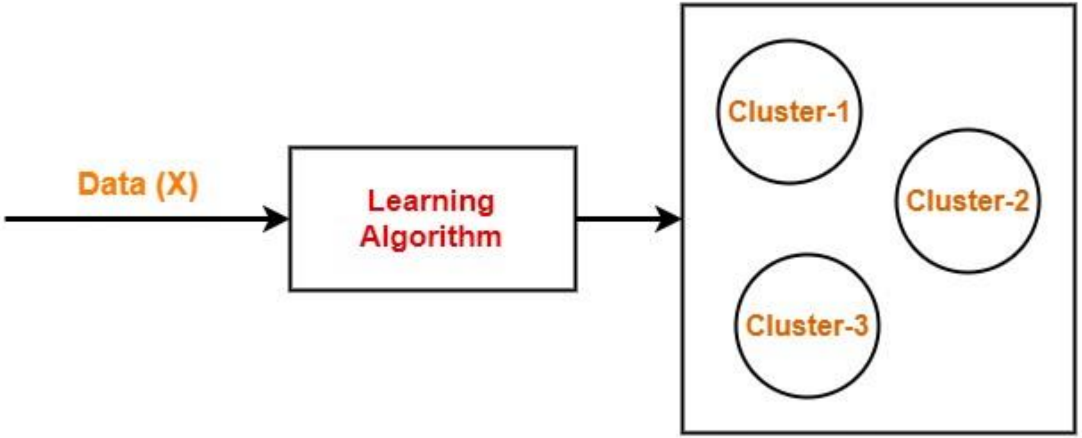
Here,

- The target variable (Y) has discrete values such as Yes or No, 0 or 1 and many more.
- Example- Credit Scoring, Spam Filtering

Unsupervised Learning

In this type of machine learning algorithm,

- The training data set is an unlabeled data set.
- In other words, the training data set contains only the input value (X) and not the target value (Y).
- Based on the similarity between data, it tries to draw inference from the data such as finding patterns or clusters.



Unsupervised Learning Applications

Applications-

Some real-life applications are-

- Document Clustering
- Finding fraudulent transactions

Reinforcement Learning

In this type of machine learning algorithm,

- The agent acts in an environment in order to maximize the rewards and minimize the penalty.
- Unlike supervised learning, no data is provided to the agent.
- The agent itself takes action or sequence of actions whether right or wrong to perform a task and learn from the experience.

Applications-

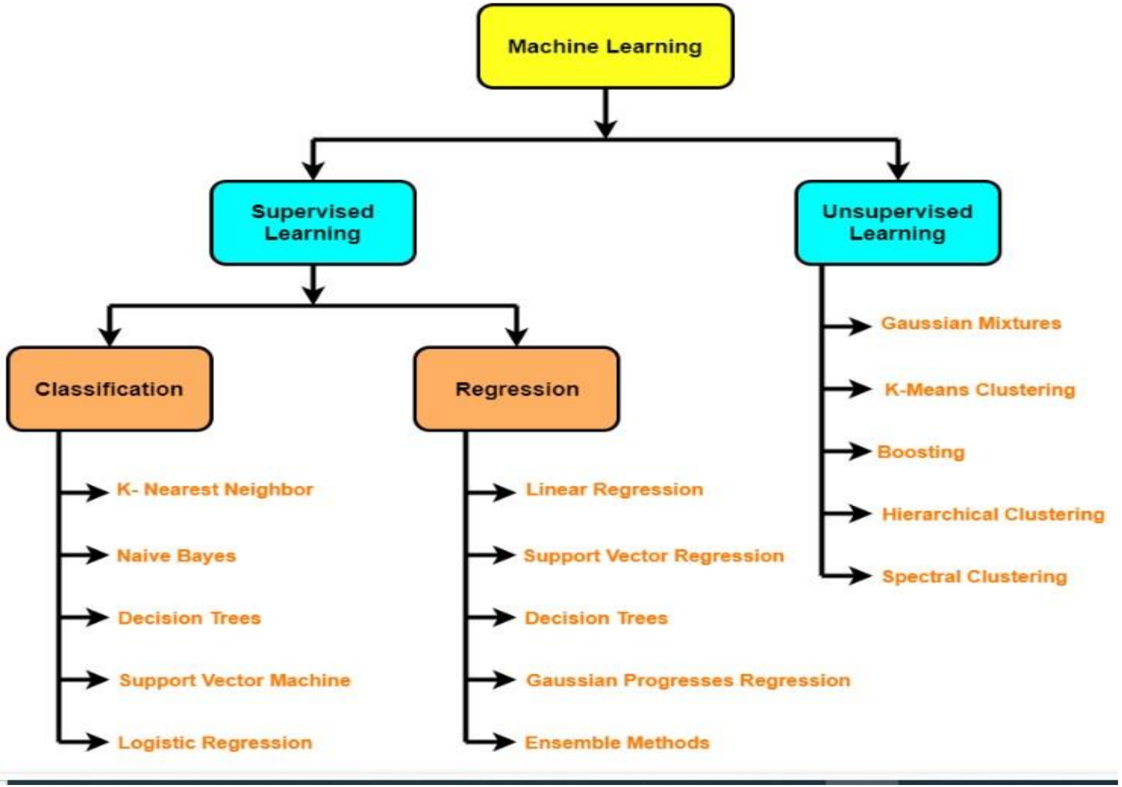
Some real-life applications are-

- Game Playing
- Robot Navigation

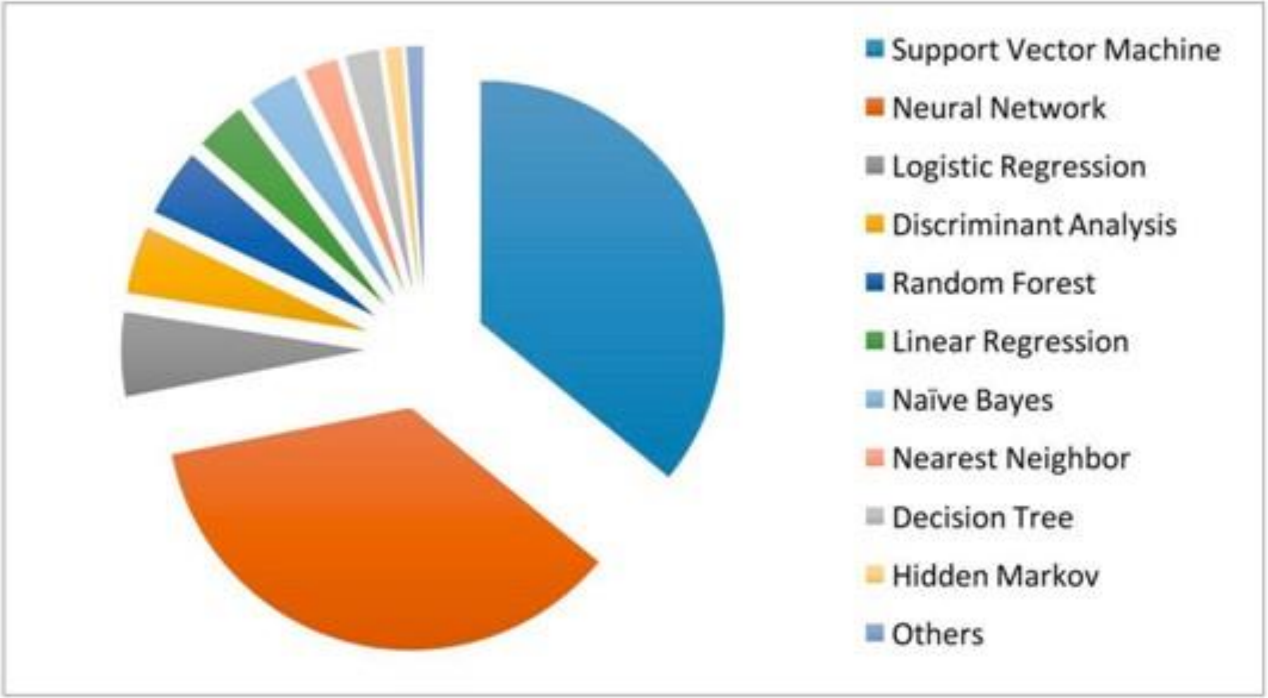
To gain better understanding about Machine Learning & its Algorithms,

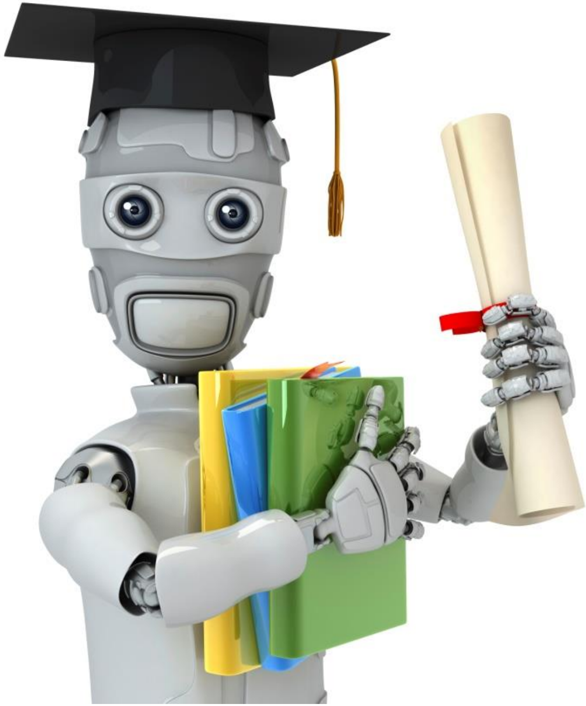
Overview

The following chart provides the overview of learning algorithms-



Current trends in Machine Learning



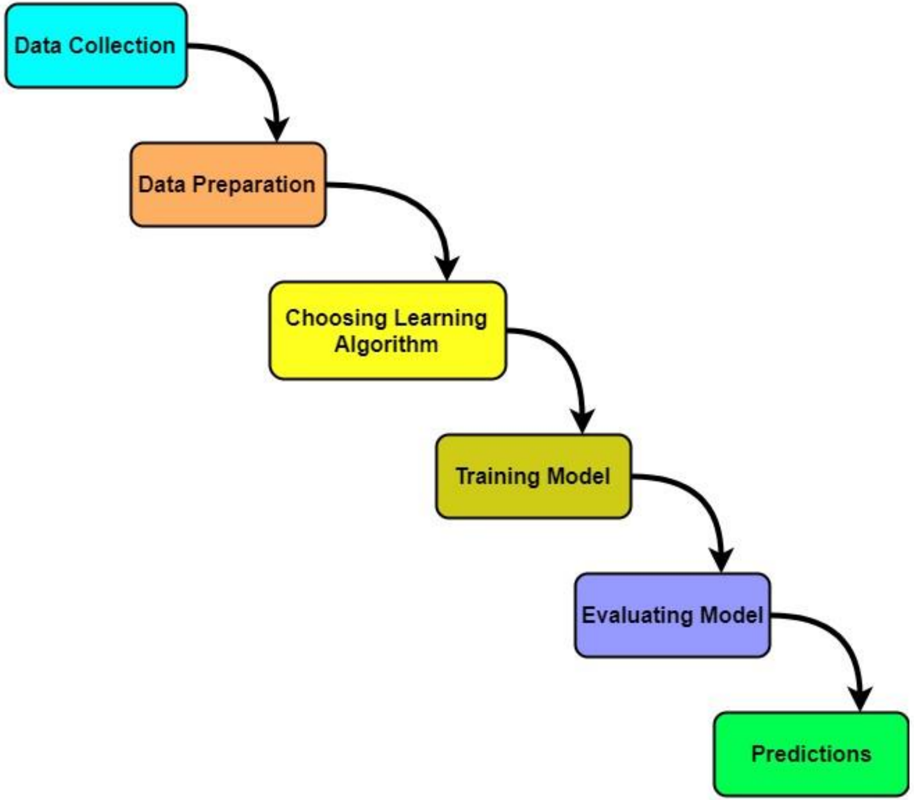


Machine Learning

Introduction

Workflow

ML workflow


Data Collection
Data Preparation
Choosing Learning Algorithm
Training Model
Evaluating Model
Predictions

Data collection

Let us discuss each stage one by one.

1. Data Collection-

In this stage,

- Data is collected from different sources.
- The type of data collected depends upon the type of desired project.
- Data may be collected from various sources such as files, databases etc.
- The quality and quantity of gathered data directly affects the accuracy of the desired system.

Data preparation

2. Data Preparation-

In this stage,

- Data preparation is done to clean the raw data.
- Data collected from the real world is transformed to a clean dataset.
- Raw data may contain missing values, inconsistent values, duplicate instances etc.
- So, raw data cannot be directly used for building a model.

Different methods of cleaning the dataset are-

- Ignoring the missing values
- Removing instances having missing values from the dataset.
- Estimating the missing values of instances using mean, median or mode.
- Removing duplicate instances from the dataset.
- Normalizing the data in the dataset.

This is the most time consuming stage in machine learning workflow.

Selection

3. Choosing Learning Algorithm-

In this stage,

- The best performing learning algorithm is researched.
- It depends upon the type of problem that needs to solved and the type of data we have.
- If the problem is to classify and the data is labeled, classification algorithms are used.
- If the problem is to perform a regression task and the data is labeled, regression algorithms are used.
- If the problem is to create clusters and the data is unlabeled, clustering algorithms are used.

The following chart provides the overview of learning algorithms-

Training

4. Training Model-

In this stage,

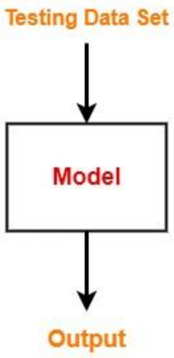
- The model is trained to improve its ability.
- The dataset is divided into training dataset and testing dataset.
- The training and testing split is order of 80/20 or 70/30.
- It also depends upon the size of the dataset.
- Training dataset is used for training purpose.
- Testing dataset is used for the testing purpose.
- Training dataset is fed to the learning algorithm.
- The learning algorithm finds a mapping between the input and the output and generates the model.

Evaluation

5. Evaluating Model-

In this stage,

- The model is evaluated to test if the model is any good.
- The model is evaluated using the kept-aside testing dataset.
- It allows to test the model against data that has never been used before for training.
- Metrics such as accuracy, precision, recall etc are used to test the performance.
- If the model does not perform well, the model is re-built using different hyper parameters.
- The accuracy may be further improved by tuning the hyper parameters.

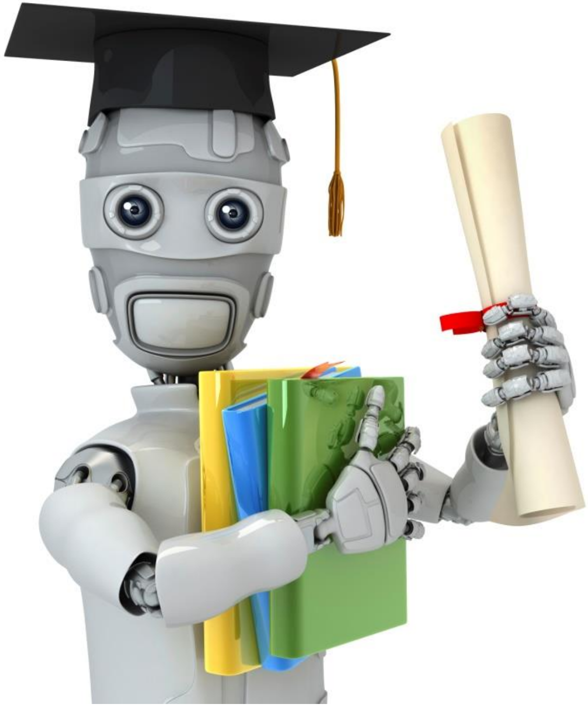


Prediction

6. Predictions-

In this stage,

- The built system is finally used to do something useful in the real world.
- Here, the true value of machine learning is realized.



Machine Learning

Illustration

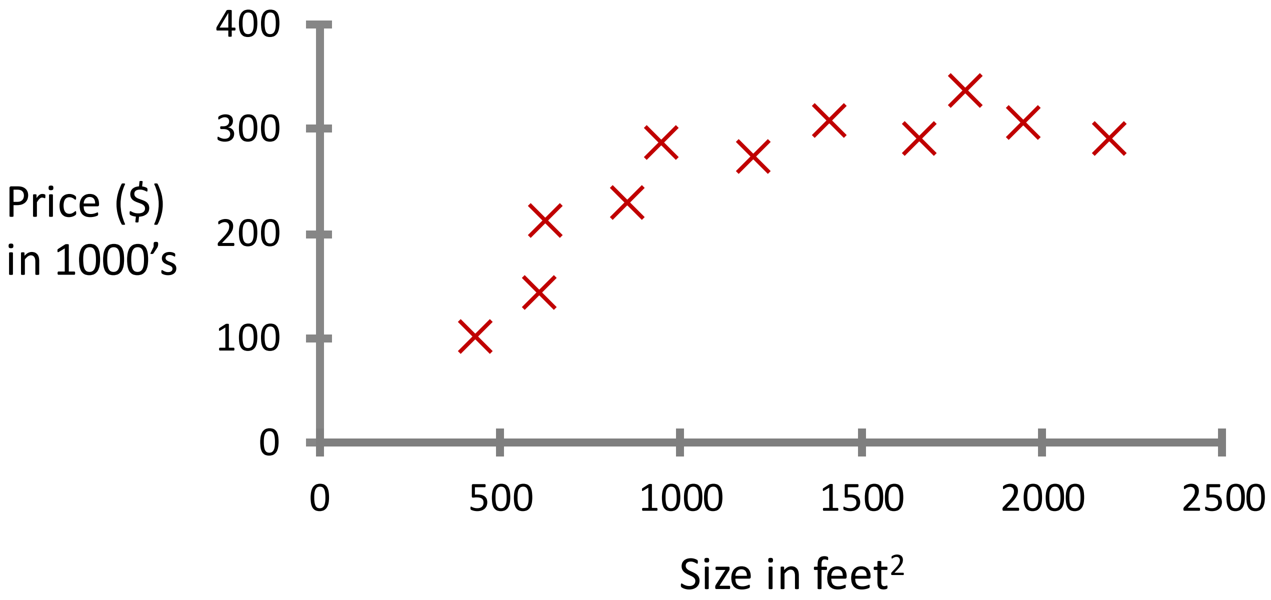
Supervised Learning

"A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

- ◯ Classifying emails as spam or not spam.
- ◯ Watching you label emails as spam or not spam.
- ◯ The number (or fraction) of emails correctly classified as spam/not spam.
- ◯ None of the above—this is not a machine learning problem.

Housing price prediction.

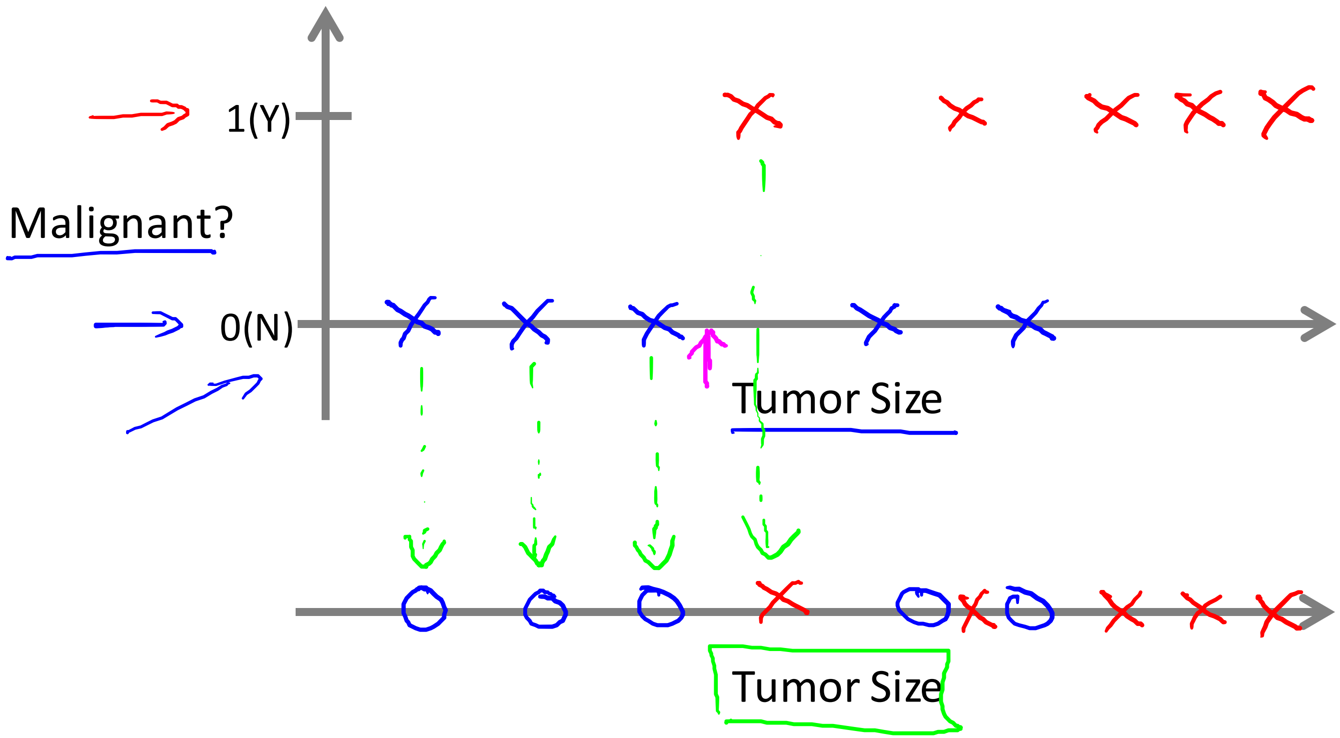


<u>Supervised Learning</u>

"right answers" given

<u>Regression:</u> Predict continuous valued output (price)

Breast cancer (malignant, benign)



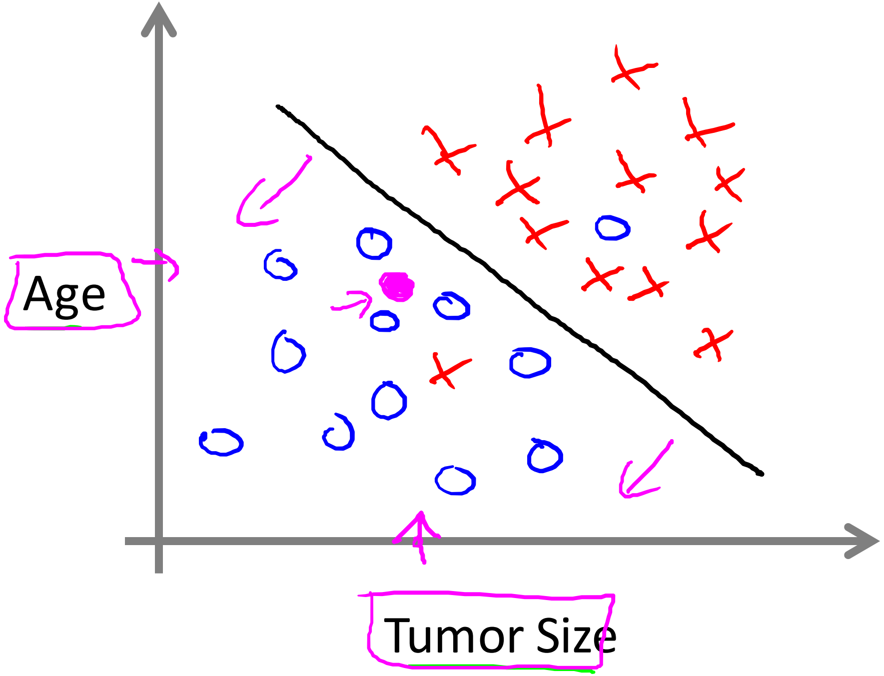
Classification

Discrete valued output (0 or 1)

0, 1, 2, 3

benign

type 1 cancer



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

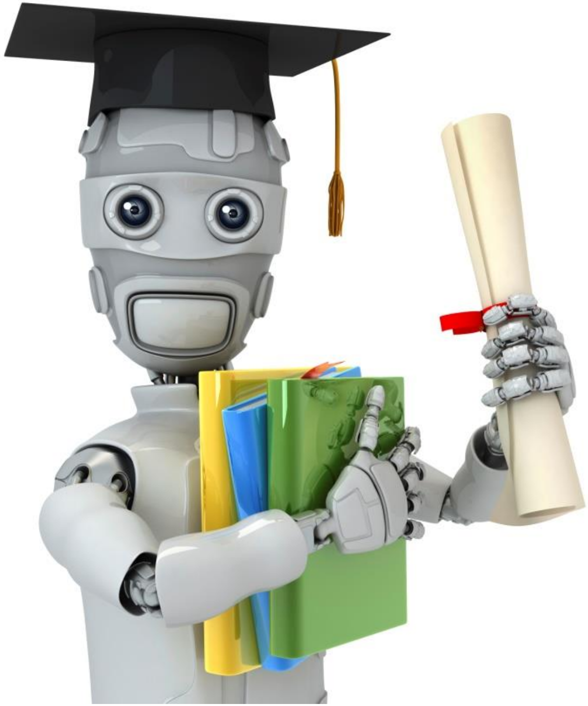
…

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

- ○ Treat both as classification problems.
- ○ Treat problem 1 as a classification problem, problem 2 as a regression problem.
- ○ Treat problem 1 as a regression problem, problem 2 as a classification problem.
- ○ Treat both as regression problems.

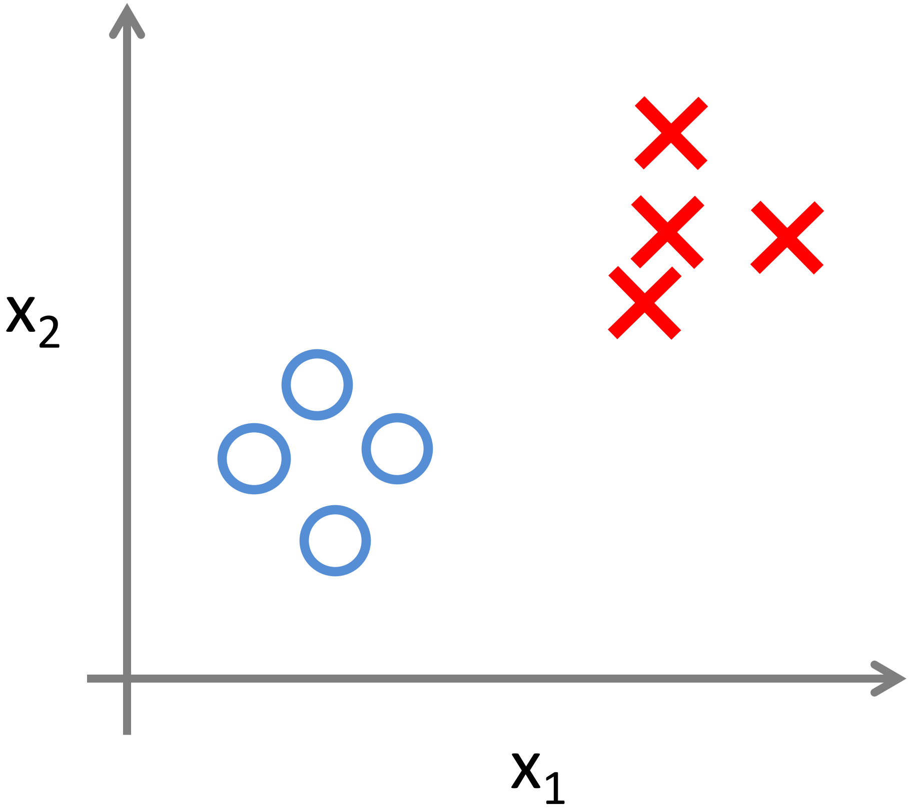


Machine Learning

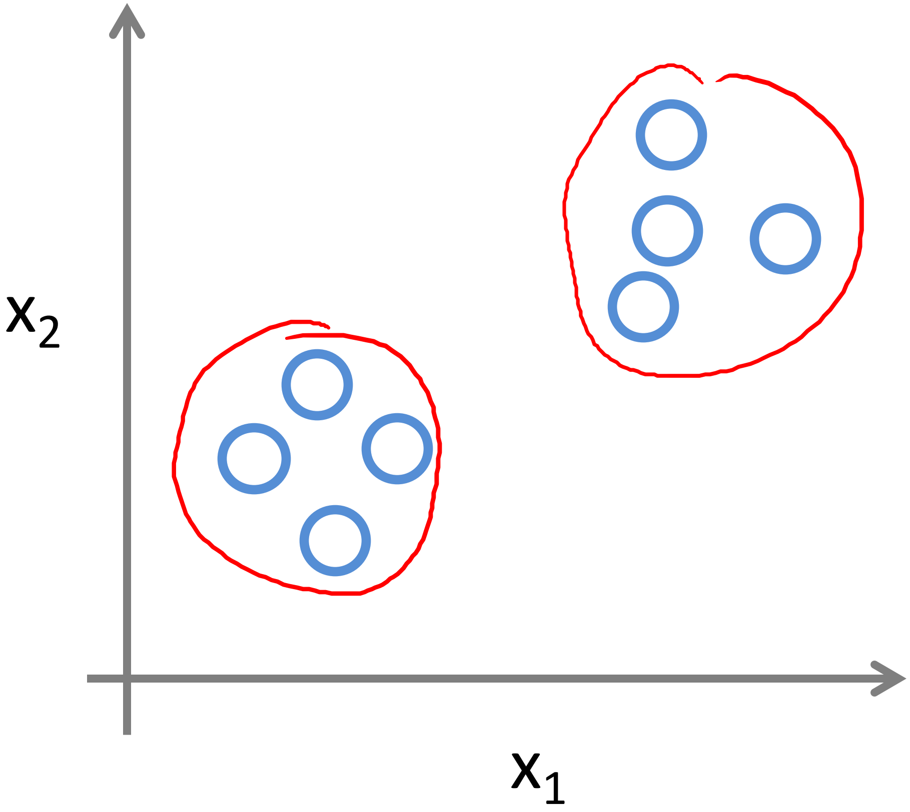
Illustration

Unsupervised Learning

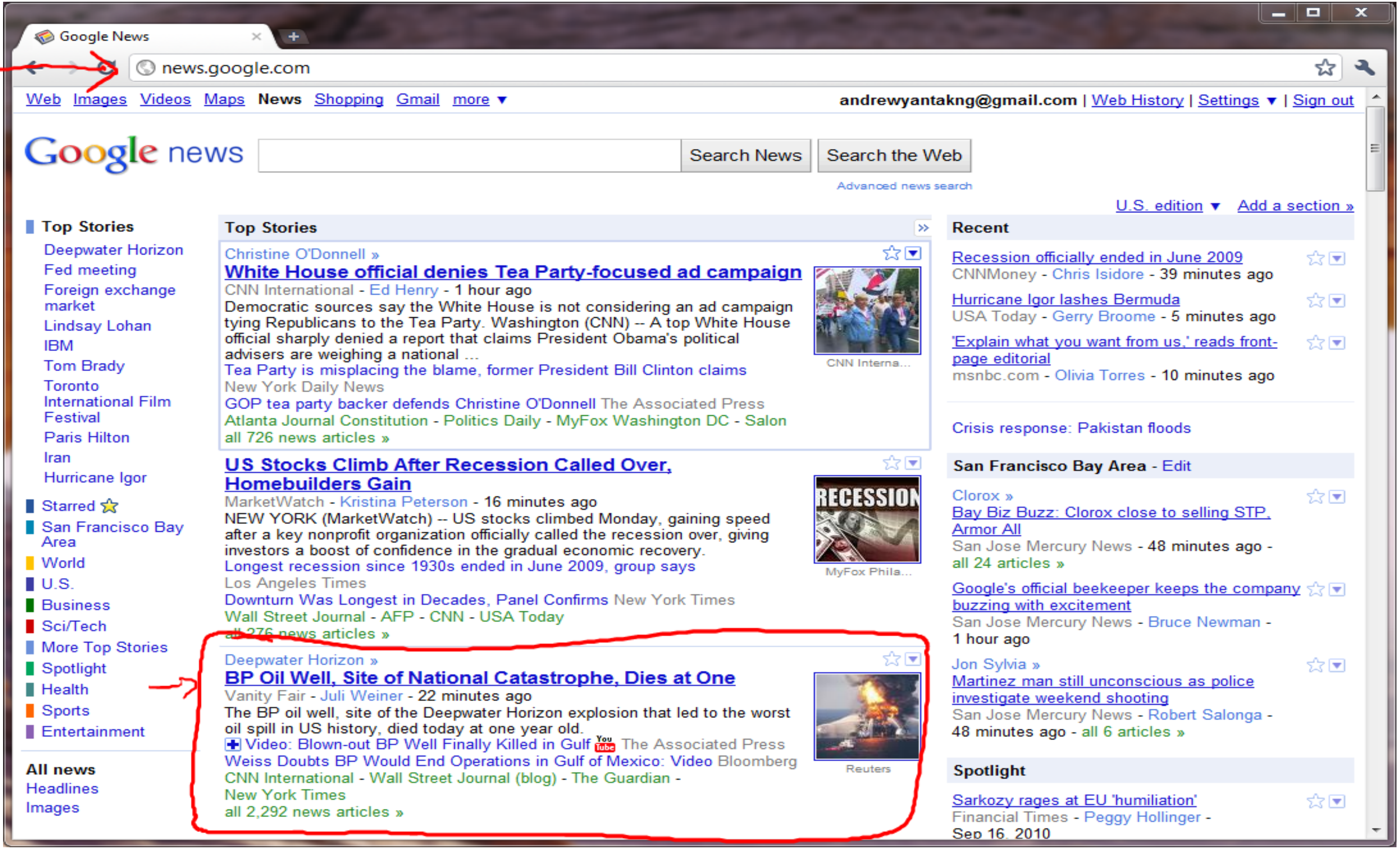
Supervised Learning

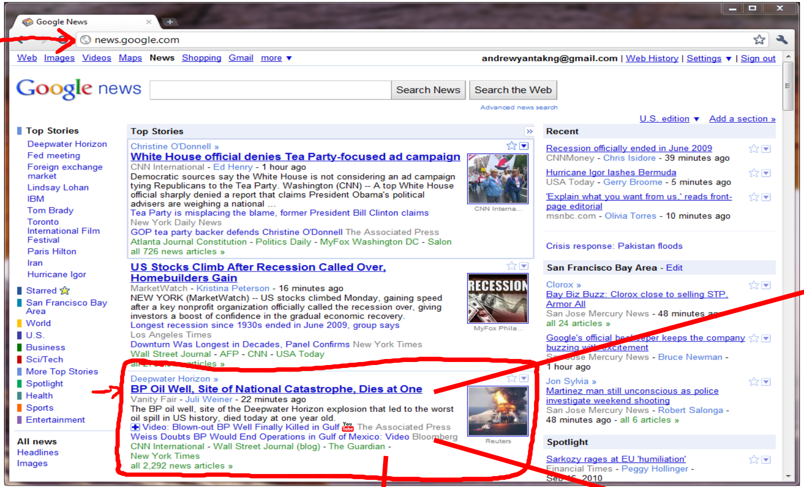


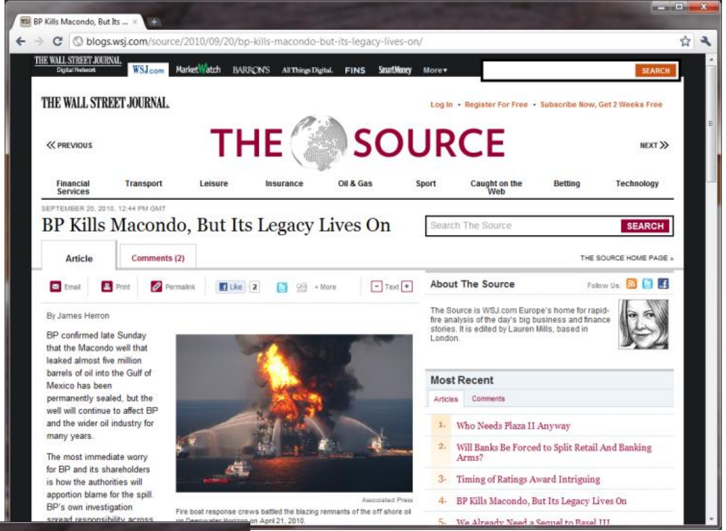
Unsupervised Learning

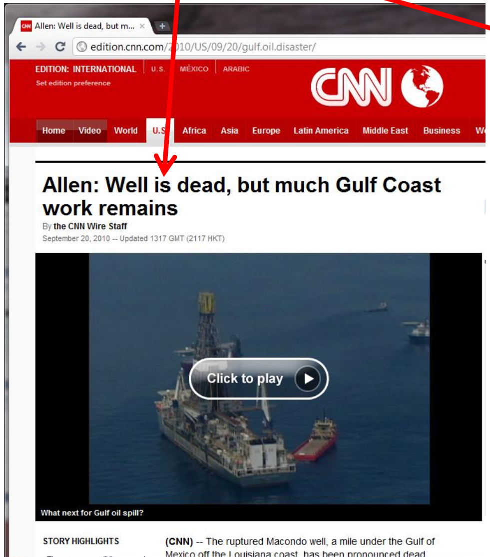


Andrew Ng


Google News
news.google.com
Web Images Videos Maps News Shopping Gmail more
andrewyantakng@gmail.com | Web History | Settings | Sign out
Google news
Search News
Search the Web
Advanced news search
U.S. edition
Add a section »
Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor
Starred
San Francisco Bay Area
World
U.S.
Business
Sci/Tech
More Top Stories
Spotlight
Health
Sports
Entertainment
All news
Headlines
Images
Top Stories
Christine O'Donnell »
White House official denies Tea Party-focused ad campaign
CNN International - Ed Henry - 1 hour ago
Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national ...
Tea Party is misplacing the blame, former President Bill Clinton claims
New York Daily News
GOP tea party backer defends Christine O'Donnell The Associated Press
Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon
all 726 news articles »
CNN Interna...
US Stocks Climb After Recession Called Over, Homebuilders Gain
MarketWatch - Kristina Peterson - 16 minutes ago
NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.
Longest recession since 1930s ended in June 2009, group says
Los Angeles Times
Downturn Was Longest in Decades, Panel Confirms New York Times
Wall Street Journal - AFP - CNN - USA Today
all 276 news articles »
RECESSION
MyFox Phila...
Deepwater Horizon »
BP Oil Well, Site of National Catastrophe, Dies at One
Vanity Fair - Juli Weiner - 22 minutes ago
The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.
Video: Blown-out BP Well Finally Killed in Gulf The Associated Press
Weiss Doubts BP Would End Operations in Gulf of Mexico: Video Bloomberg
CNN International - Wall Street Journal (blog) - The Guardian - New York Times
all 2,292 news articles »
Reuters
Recent
Recession officially ended in June 2009
CNNMoney - Chris Isidore - 39 minutes ago
Hurricane Igor lashes Bermuda
USA Today - Gerry Broome - 5 minutes ago
'Explain what you want from us,' reads front-page editorial
msnbc.com - Olivia Torres - 10 minutes ago
Crisis response: Pakistan floods
San Francisco Bay Area - Edit
Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago - all 24 articles »
Google's official beekeeper keeps the company buzzing with excitement
San Jose Mercury News - Bruce Newman - 1 hour ago
Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »
Spotlight
Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger - Sep 16, 2010

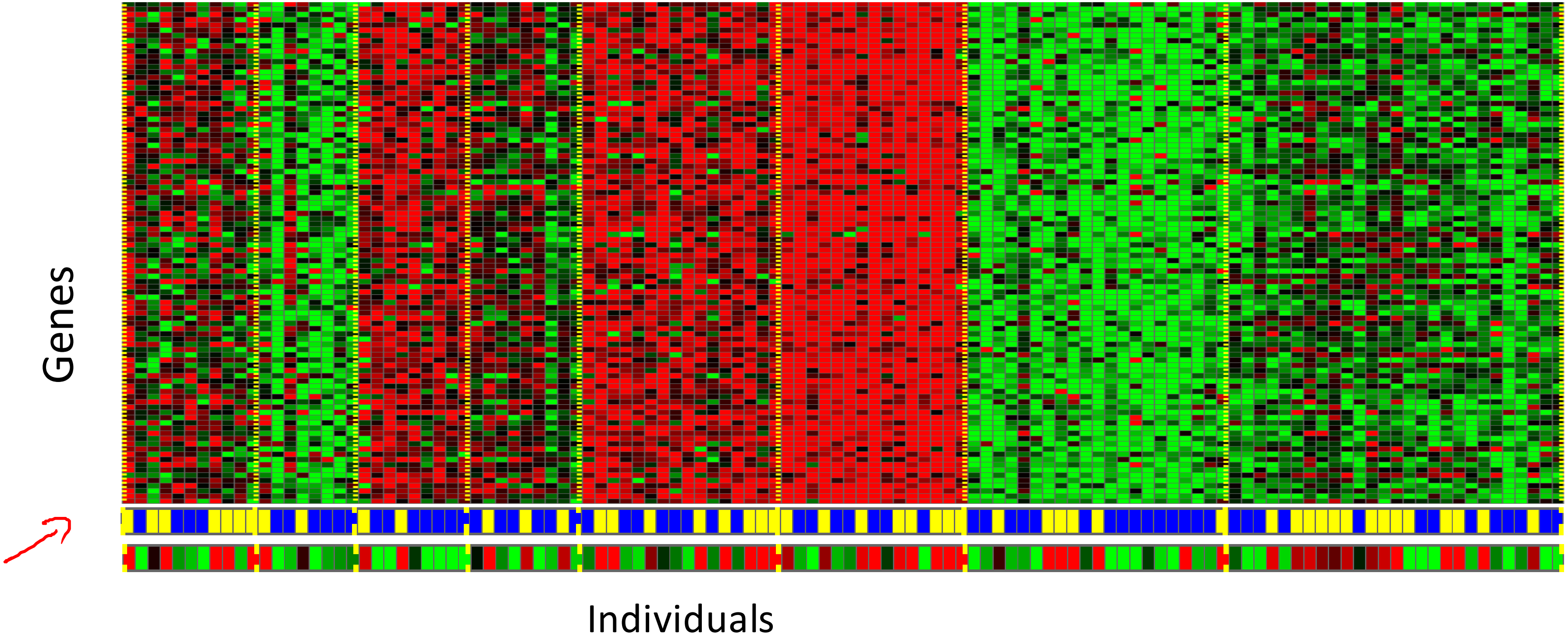

Google News
news.google.com
Web Images Videos Maps News Shopping Gmail more
andrewyantakng@gmail.com | Web History | Settings | Sign out
Google news
Search News
Search the Web
Advanced news search
U.S. edition
Add a section »
Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor
Starred
San Francisco Bay Area
World
U.S.
Business
Sci/Tech
More Top Stories
Spotlight
Health
Sports
Entertainment
All news
Headlines
Images
Top Stories
Christine O'Donnell »
White House official denies Tea Party-focused ad campaign
CNN International - Ed Henry - 1 hour ago
Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national ...
Tea Party is misplacing the blame, former President Bill Clinton claims
New York Daily News
GOP tea party backer defends Christine O'Donnell The Associated Press
Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon
all 726 news articles »
US Stocks Climb After Recession Called Over, Homebuilders Gain
MarketWatch - Kristina Peterson - 16 minutes ago
NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.
Longest recession since 1930s ended in June 2009, group says
Los Angeles Times
Downturn Was Longest in Decades, Panel Confirms New York Times
Wall Street Journal - AFP - CNN - USA Today
RECESSION
Deepwater Horizon »
BP Oil Well, Site of National Catastrophe, Dies at One
Vanity Fair - Juli Weiner - 22 minutes ago
The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.
Video: Blown-out BP Well Finally Killed in Gulf The Associated Press
Weiss Doubts BP Would End Operations in Gulf of Mexico: Video Bloomberg
CNN International - Wall Street Journal (blog) - The Guardian - New York Times
all 2,292 news articles »
Reuters
Recent
Recession officially ended in June 2009
CNNMoney - Chris Isidore - 39 minutes ago
Hurricane Igor lashes Bermuda
USA Today - Gerry Broome - 5 minutes ago
'Explain what you want from us,' reads front-page editorial
msnbc.com - Olivia Torres - 10 minutes ago
Crisis response: Pakistan floods
San Francisco Bay Area - Edit
Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago
all 24 articles »
Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »
Spotlight
Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger -


BP Kills Macondo, But Its ...
blogs.wsj.com/source/2010/09/20/bp-kills-macondo-but-its-legacy-lives-on/
THE WALL STREET JOURNAL.
Log In • Register For Free • Subscribe Now, Get 2 Weeks Free
THE SOURCE
PREVIOUS
NEXT
Financial Services
Transport
Leisure
Insurance
Oil & Gas
Sport
Caught on the Web
Betting
Technology
SEPTEMBER 20, 2010, 12:44 PM GMT
BP Kills Macondo, But Its Legacy Lives On
SEARCH
Article
Comments (2)
By James Herron
BP confirmed late Sunday that the Macondo well that leaked almost five million barrels of oil into the Gulf of Mexico has been permanently sealed, but the well will continue to affect BP and the wider oil industry for many years.
The most immediate worry for BP and its shareholders is how the authorities will apportion blame for the spill. BP's own investigation
Associated Press
About The Source
The Source is WSJ.com Europe's home for rapid-fire analysis of the day's big business and finance stories. It is edited by Lauren Mills, based in London.
Most Recent
1. Who Needs Plaza II Anyway
2. Will Banks Be Forced to Split Retail And Banking Arms?
3. Timing of Ratings Award Intriguing
4. BP Kills Macondo, But Its Legacy Lives On

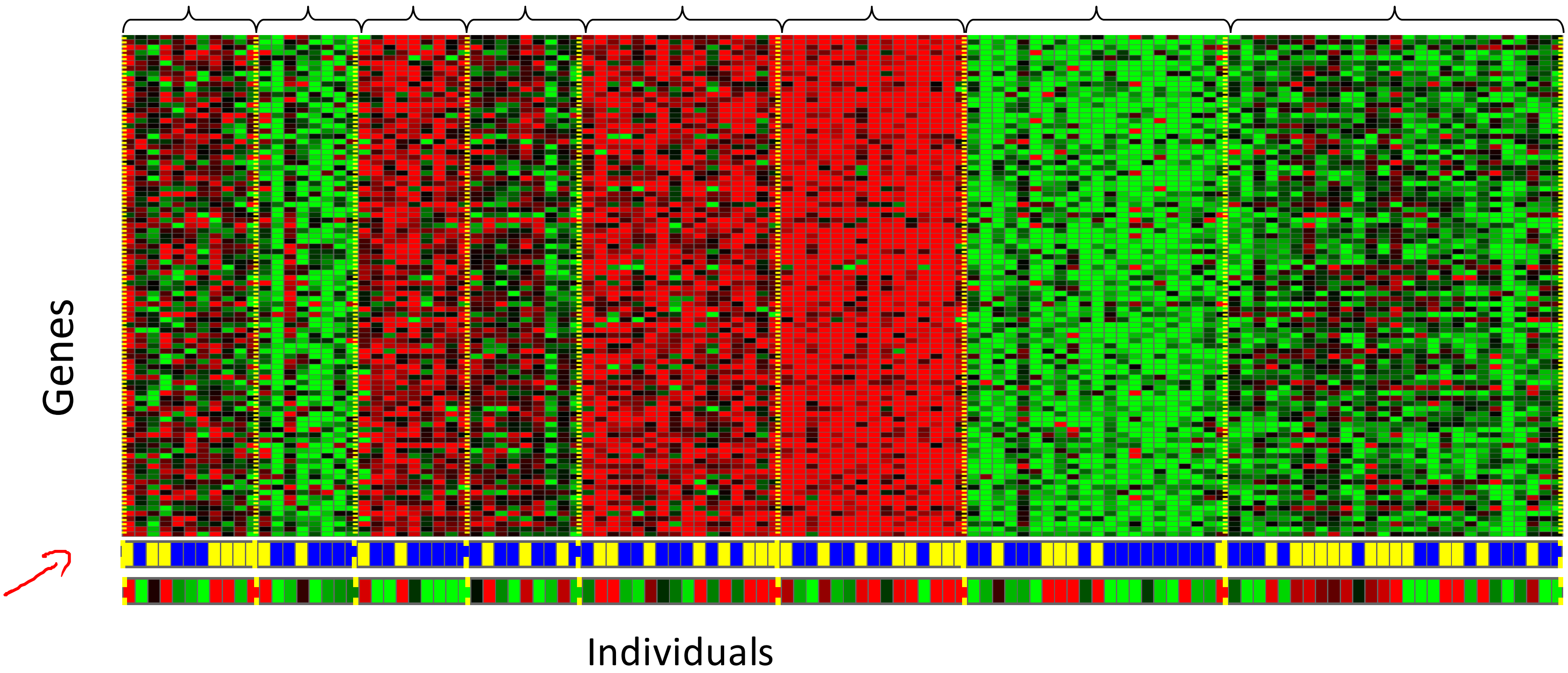

Allen: Well is dead, but m...
edition.cnn.com/2010/US/09/20/gulf.oil.disaster/
EDITION: INTERNATIONAL
U.S.
MÉXICO
ARABIC
Set edition preference
CNN
Home
Video
World
U.S.
Africa
Asia
Europe
Latin America
Middle East
Business
Allen: Well is dead, but much Gulf Coast work remains
By the CNN Wire Staff
September 20, 2010 -- Updated 1317 GMT (2117 HKT)
Click to play
What next for Gulf oil spill?
STORY HIGHLIGHTS
(CNN) -- The ruptured Macondo well, a mile under the Gulf of Mexico off the Louisiana coast, has been pronounced dead


BP oil spill cost hits nearl...
www.guardian.co.uk/environment/2010/sep/20/bp-oil-spill-dee
Mobile site
Sign in
Register
Text larger
smaller
About us
guardian.co.uk
News
Sport
Comment
Culture
Business
Money
Life & style
Business
BP
BP oil spill cost hits nearly $10bn
BP has set up a $20bn compensation fund after the Deepwater Horizon disaster, which has so far paid out 19,000 claims totalling more than $240m
Julia Kollewe
guardian.co.uk, Monday 20 September 2010 08.33 BST
Article history
BP's costs for the Deepwater Horizon disaster have hit $10bn. Photograph: Ho/Reuters

Andrew Ng



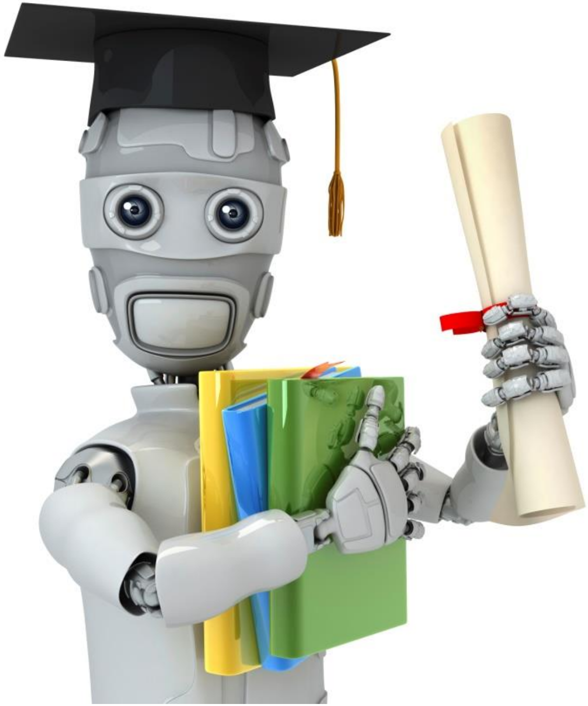
[Source: Daphne Koller]



[Source: Daphne Koller]

Of the following examples, which would you address using an <u>unsupervised</u> learning algorithm? (Check all that apply.)

- ☐ Given email labeled as spam/not spam, learn a spam filter.
- ☐ Given a set of news articles found on the web, group them into set of articles about the same story.
- ☐ Given a database of customer data, automatically discover market segments and group customers into different market segments.
- ☐ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.



Machine Learning

Data Preprocessing

Data Preprocessing: Data cleansing

Missing values

- A missing value occurs when an attribute is not recorded for a unit (they are usually coded, i.e. 999, NA)
- Many reasons : nonresponse, error, mistake.
- It may concern all the attributes of the unit or some of them.
- Missing values can reveal important information about the data ==¿ they can false the whole DM analysis
  Golden rule : all the efforts must be done during the data acquisition.

Data Preprocessing: Data cleansing

Missing values (elimination imputation)

However, one sometimes eliminates from the analysis the units with missing values or imputes them, i.e. fills them up with artificial values.

If the missing values are completely at random and a very low fraction of n :

- hot deck imputation (very bad idea !)
- list-wise elimination

More complex imputation schemes include :

- mean / median value imputation for a quantitative attribute
- mode imputation for a qualitative attribute
- regression imputation

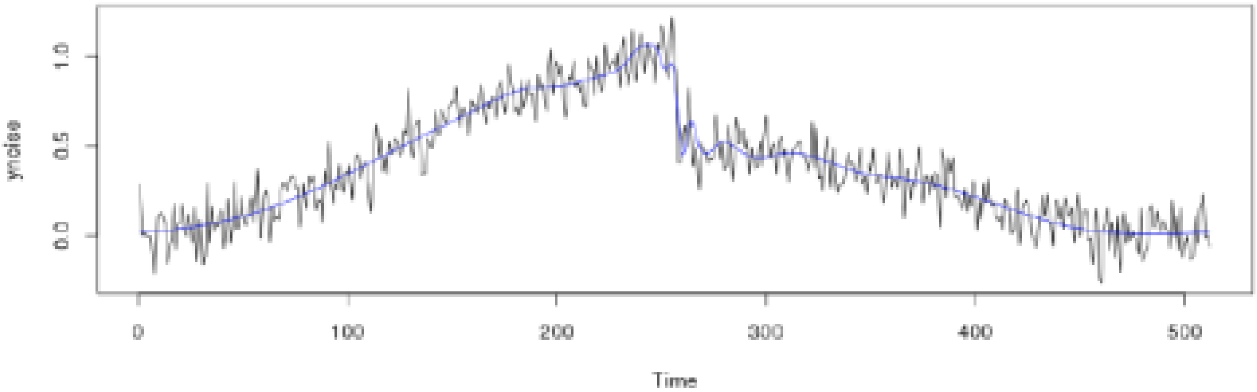
Also, you may rely on robust algorithms that will work even with NA .

Data Preprocessing: Data cleansing

Noisy data

- Noise can be seen as unstructured (randomly) unwanted data
- Can be due to low quality technology in acquisition or transmission (i.e. cheap microphones or cables).
- Noise may difficult data mining (or fake it)

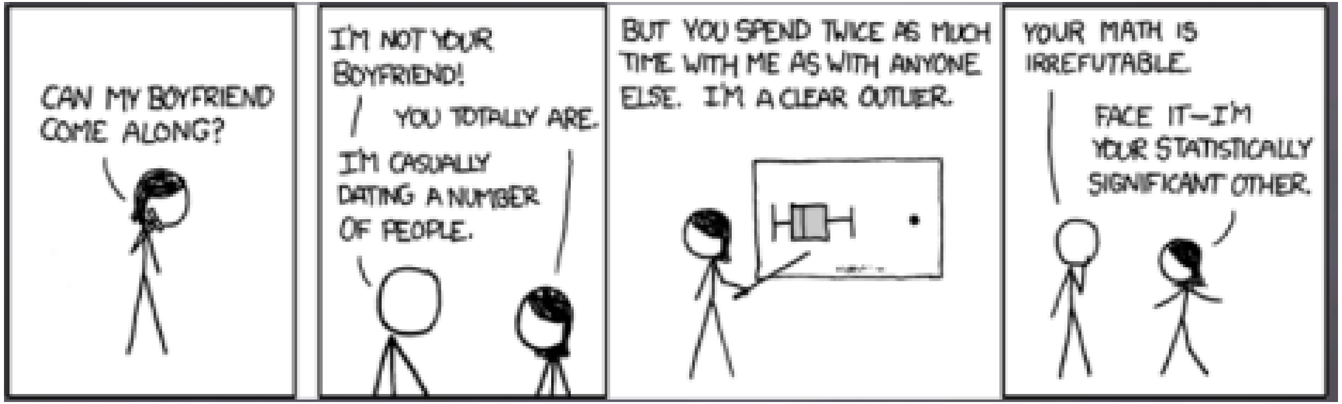
Noise can be reduct using smoothing filters or thresholding.



Data Preprocessing: Data cleansing

Outliers vs Influential units

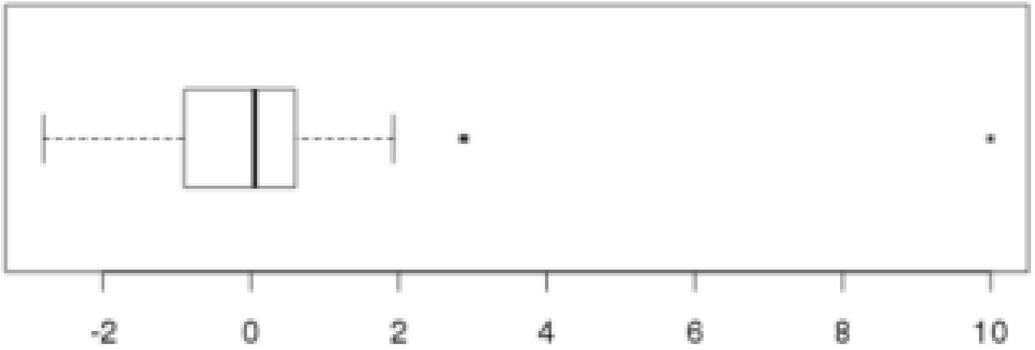
- An outlier is an unit that have a different probability structure from the pack. It may be due to measurement error or heavy-tailed distributions (i.e. high kurtosis) .
- An unit is influential if its deletion noticeably alters the result of the analysis.

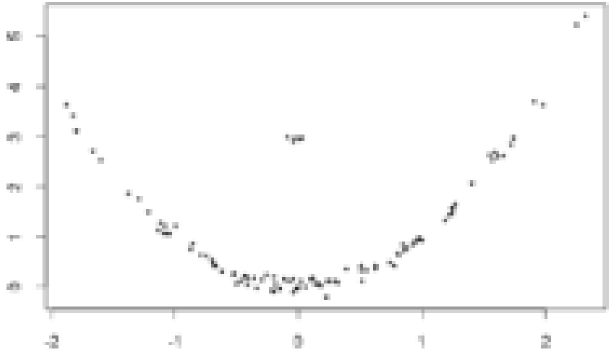


Data Preprocessing: Data cleansing

Outliers vs Influential units

Detection of outliers using graphical tools





A frequently used rule-of-the thumb is to assume that observations laying outside Q2 1 :5 IQR are outliers (be very careful with this kind of rule).

Data Preprocessing: Data transformation

Why would someone choose to transform the data ?

- Some techniques may need to transform the data in order to make it dimensionless, e.g. correlation coefficient, or to rend some hypothesis more reasonable (e.g. log for stabilize variance)
- Sometimes we are interest on categories instead of numerical scales (e.g. low, mid, high income instead of actual nominal income)
- Some techniques can not handle categorical values with more than two categories (i.e. binary variables)
- The target variables were not recorded but you have a proxy
- De-noising (check section 2.c)
- Aggregation : in order to change resolution of data

Data Preprocessing: Data transformation

Min-Max normalization

- If X is the original attribute we compute a new attribute X^* by computing

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Linear transformation
- Maps data from the original range $[X_{min}, X_{max}]$ to $[0, 1]$

z-core normalization

- If X is the original attribute we compute a new attribute z_X by computing

$$z_X = \frac{X - \bar{X}}{s_X^2}$$

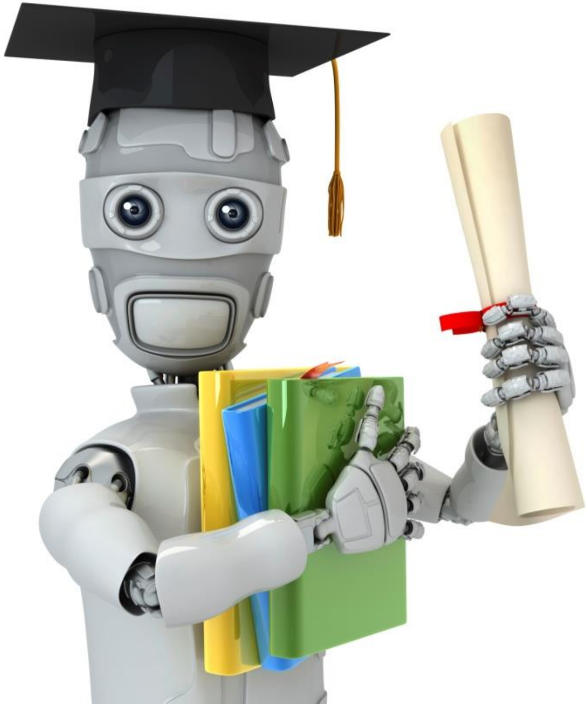
- Associated to the normalization of a normal random variable
- We use it implicitly when we compute correlations.
- The z-core normalizes an attribute to have zero mean and unitary standard deviation. (Prove it !)

Data Preprocessing: Data reduction

Reduction can be performed by

- selecting instances (i.e. rows of the data matrix)
- selecting features/attributes (i.e. columns of the data matrix)
- combining instances : e.g. data aggregation
- combining features : e.g. PCA (coming soon !)

The reduction may be wanted to reduce the computational time of some algorithms.



Machine Learning

See you next chapter