Introduction to Machine Learning

Tuan Nguyen

Introduction

- Giảng viên Khoa Toán Kinh tế, Đại học Kinh tế Quốc dân.
- Tuan https://nttuan8.com/, founder: AI For Everyone (AI4E).
- The author of deep learning series, GAN series.
- The author of e-book deep learning, https://nttuan8.com/sach-deep-learning-co-ban/
- Deep learning, python teachers.

Outline

- What is Machine Learning?
- Types of Machine Learning
- Sklearn library
- Machine learning project steps

What is ~~Machine~~ Learning?

The acquisition of knowledge or skills through experience, study, or by being taught

What is Machine Learning?

There are two main steps in Machine Learning task:

- Training: Data -> Model
- Prediction: Model -> Predict

House price prediction

Diện tích	Giá
30	448.524
32.4138	509.248
34.8276	535.104
37.2414	551.432
39.6552	623.418
42.069	625.992

Data

Training

ML algorithm → **Model**

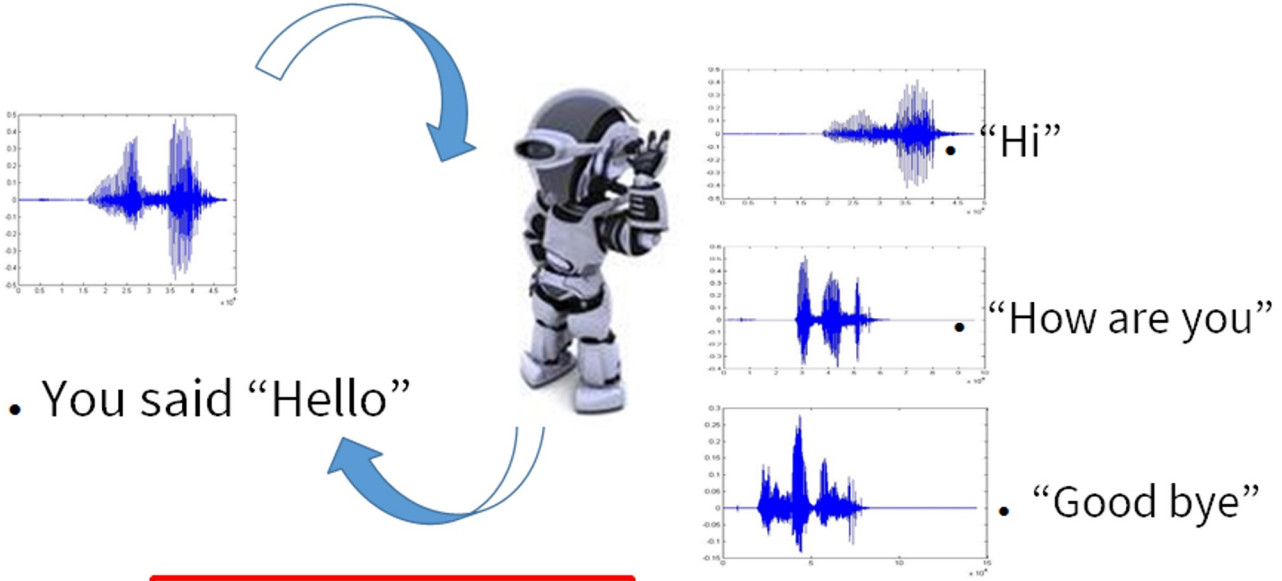
Prediction

Diện tích (60) → **Model** → Giá (800)

What is Machine Learning?

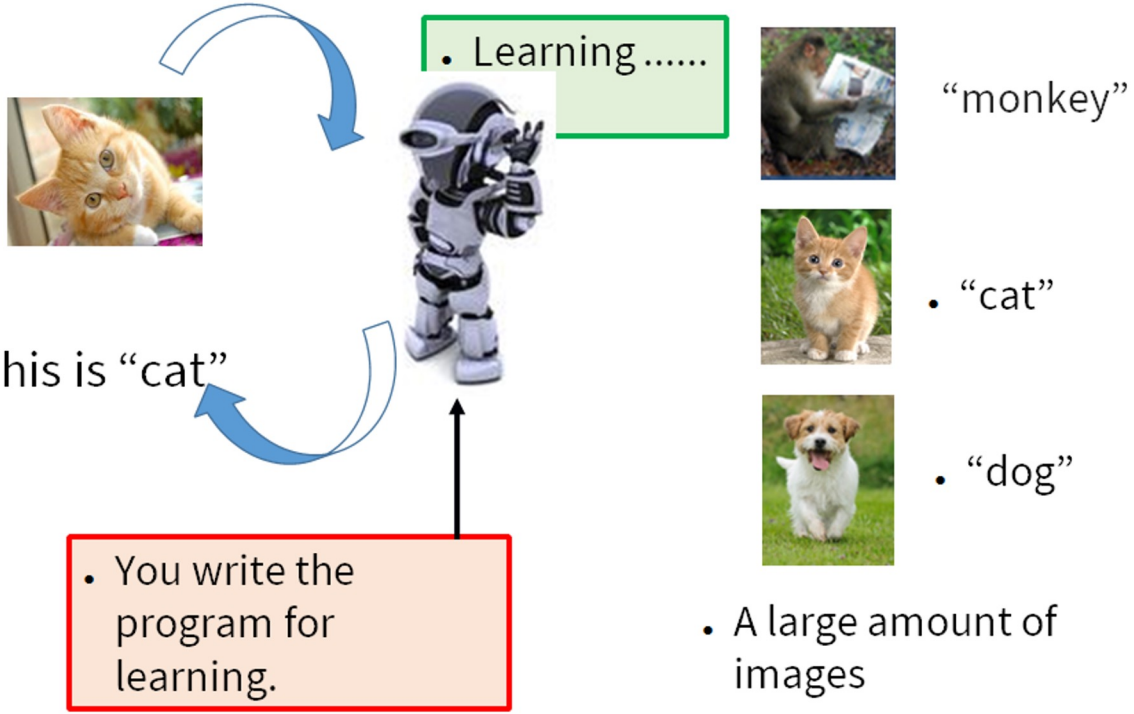
- [Arthur Samuel, 1959] Field of study that gives computers the ability to learn without being explicitly programmed
- [Kevin Murphy] algorithms that automatically detect patterns in data use the uncovered patterns to predict future data or other outcomes of interest
- [Tom Mitchell] algorithms that improve their performance (P) at some task (T) with experience (E)

What is Machine Learning?



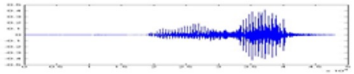
- You write the program for learning.

What is Machine Learning



Machine Learning ≈ Looking for a Function

- Speech Recognition

$f($  $)=$ “How are you”

- Image Recognition

$f($  $)=$ “Cat”

- Playing Go

$f($  $)=$ “5-5” (next move)

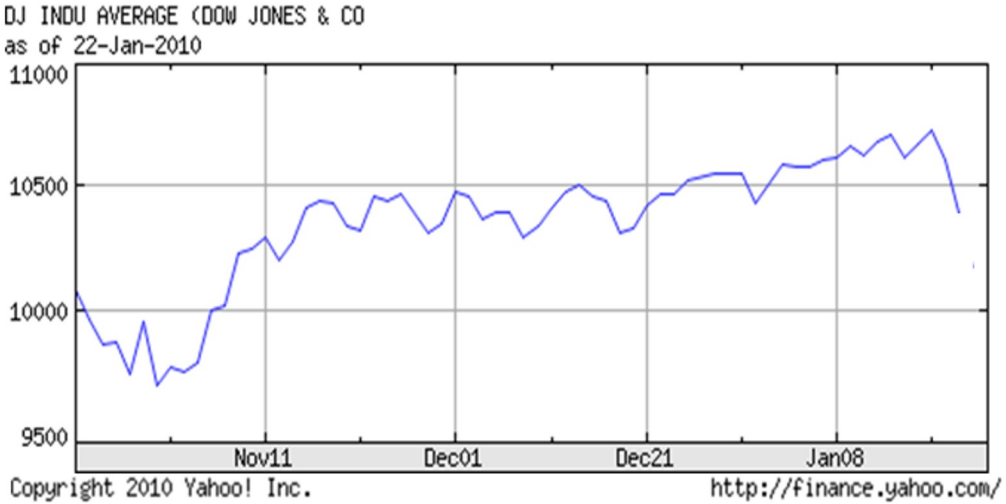
- Dialogue System

$f($ “Hi” $)=$ “Hello”

(what the user said) (system response)

Machine Learning in Action

Stock Market Prediction



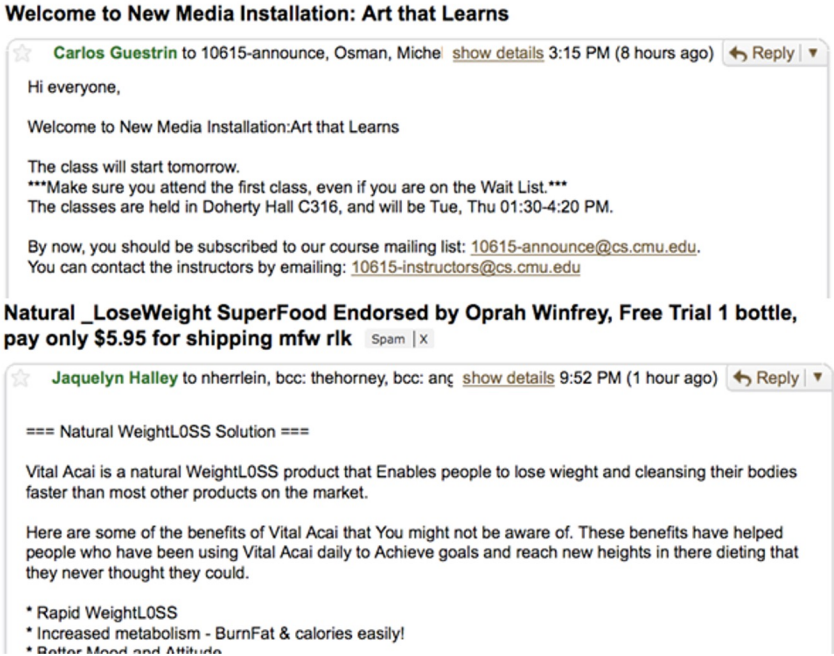
Machine Learning in Action

Document classification

Sports
Science
News

Machine Learning in Action

Spam filtering



Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

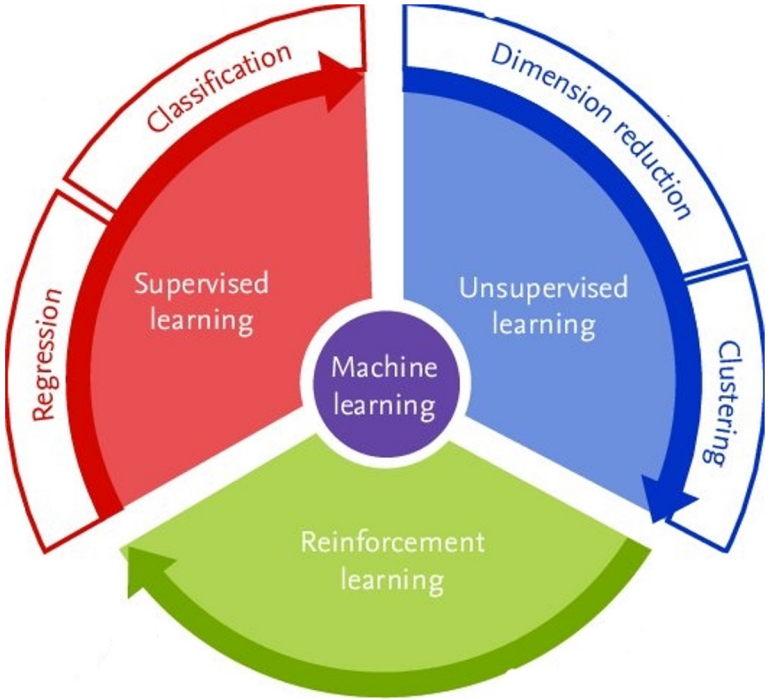
Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude

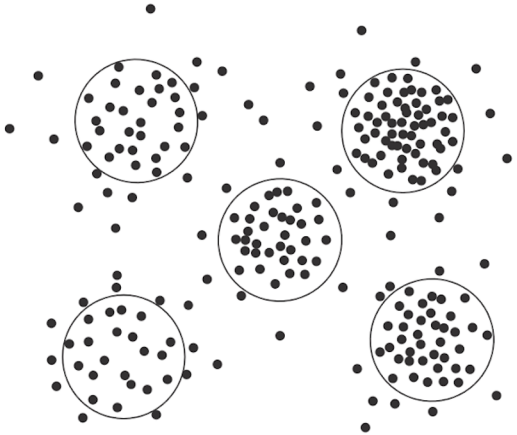
Spam/
Not spam

Types of Machine Learning

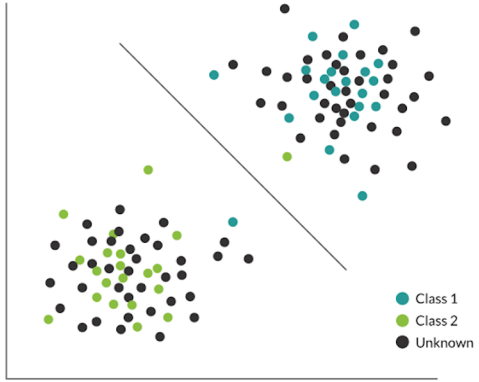


Supervised vs Unsupervised Learning

Unsupervised



Supervised



Supervised Learning

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
…

Share Price
"$ 24.50"

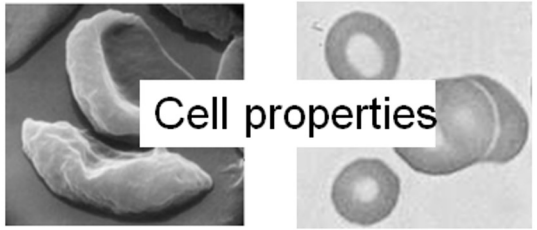
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Supervised Learning - Classification

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

"Sports"
"News"
"Science"
…



"Anemic cell"
"Healthy cell"

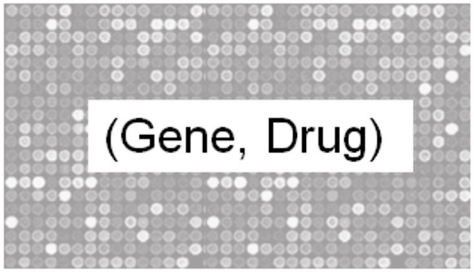
Discrete Labels

Supervised Learning - Regression

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$

Share Price
"$ 24.50"

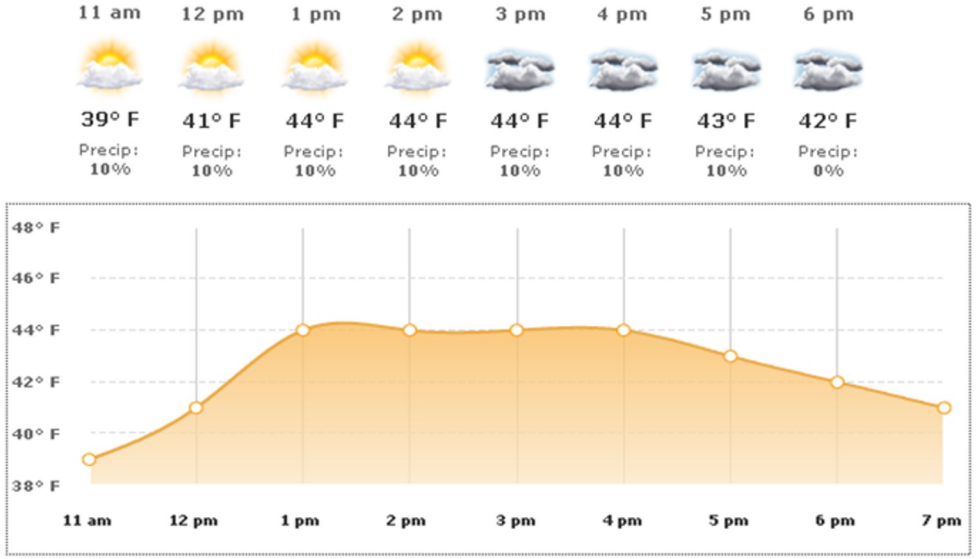


Expression level
"0.01"

Continuous Labels

Supervised Learning problems

Features? **Labels?** **Classification/Regression?**



Temperature/Weather prediction

Supervised Learning problems

Features? Labels? Classification/Regression?

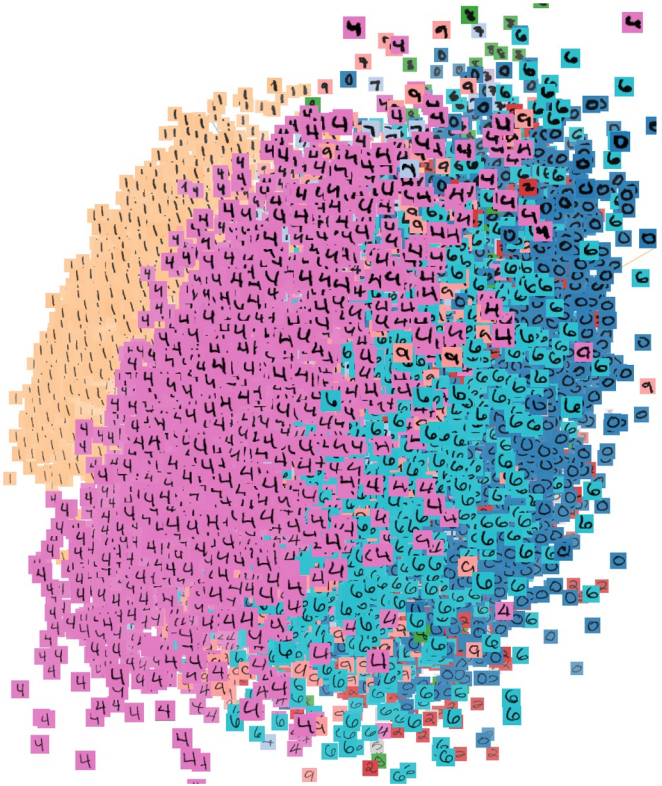
Environmental Mapping

Unsupervised Learning

Unsupervised Learning

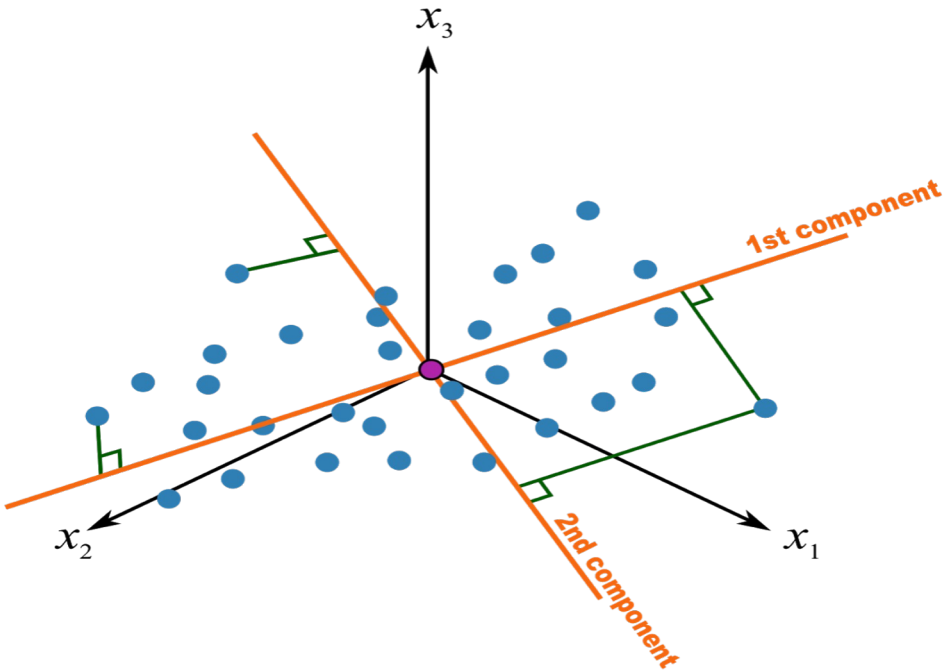
Data: X no label

Goal: Learn the structure of the data learn correlations between features

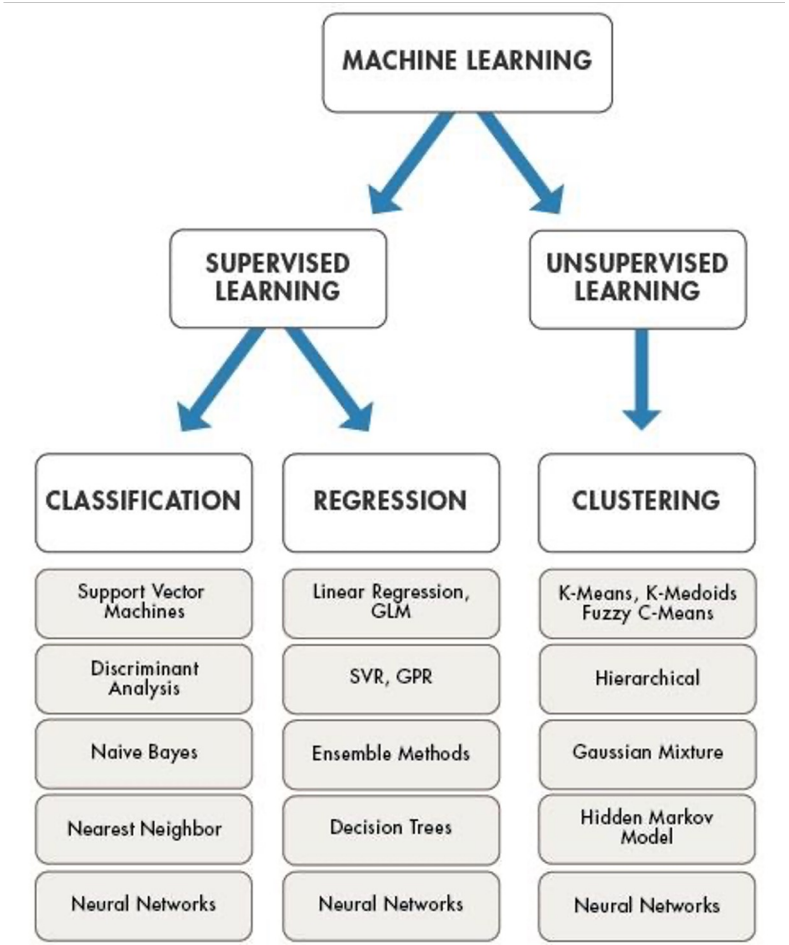


Principal Component analysis (PCA)

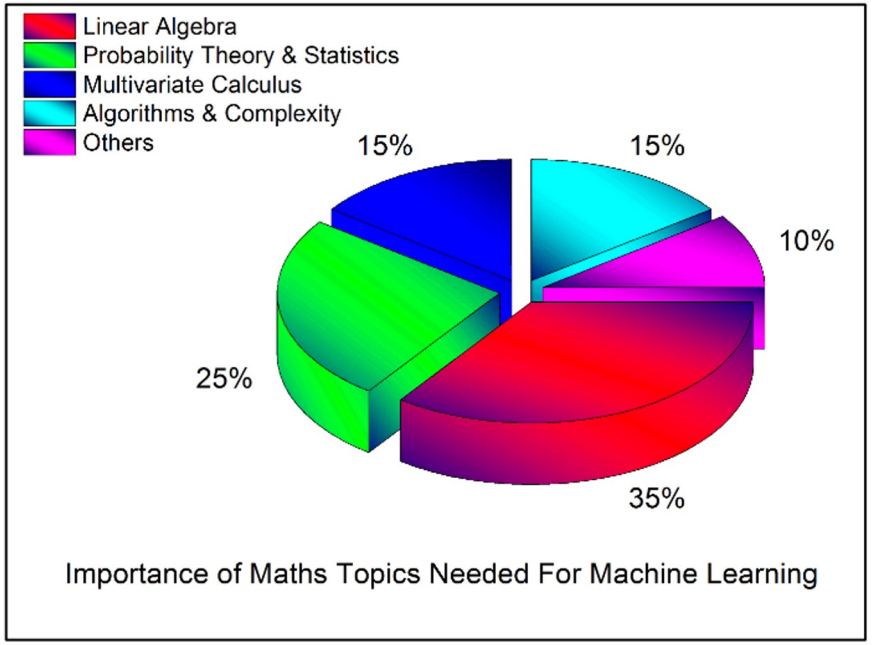
- Statistical approach for data compression and visualization
- Invented by Karl Pearson in 1901
- Weakness: linear components only.



ML Algorithms



Math for Machine Learning



Machine learning project steps

- Define problem
- Summarize data
- Preprocessing data
- Evaluate algorithms
- Improve results
- Present results

1. Define problem

Load everything we need to start working on your problem.

- Python modules, classes and functions that you intend to use.
- Loading your dataset from CSV

1. Define problem

```
# Load libraries
from pandas import read_csv
from pandas.tools.plotting import scatter_matrix
from matplotlib import pyplot
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
# Load dataset
filename = 'iris.data.csv'
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = read_csv(filename, names=names)
```

2. Summary data

Visualize and understand the data

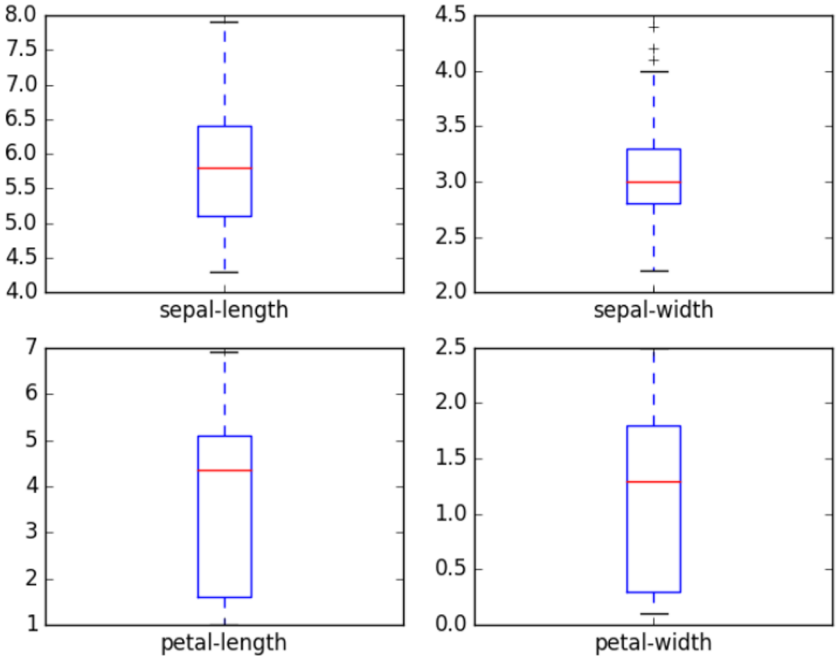
- Descriptive statistics such as summaries.
- Data visualizations such as plots with Matplotlib, ideally using convenience functions from Pandas.

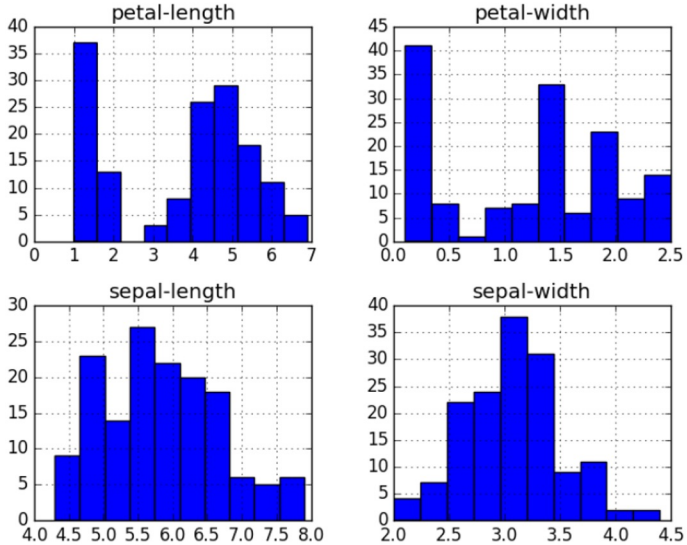
2. Summary data

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

class	
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

2. Summary data





3. Preprocessing data

Preprocess the data before fitting into the model

- Cleaning data by removing duplicates, marking missing values and even imputing missing values
- Feature selection where redundant features may be removed and new features developed
- Data transforms where attributes are scaled or redistributed in order to best expose the structure of the problem later to learning algorithms

4. Evaluate algorithms

Choose machine learning algorithms

- Separating out a validation dataset to use for later confirmation of the skill of your developed model.
- Defining test options using scikit-learn such as cross validation and the evaluation metric to use.
- Spot-checking a suite of linear and nonlinear machine learning algorithms.
- Comparing the estimated accuracy of algorithms.

Choose ML algorithm

- Logistic Regression (LR).
- Linear Discriminant Analysis (LDA).
- k-Nearest Neighbors (KNN).
- Classification and Regression Trees (CART).
- Gaussian Naive Bayes (NB).
- Support Vector Machines (SVM)

5. Improve results

Improve the algorithm in the previous step

- Search for a combination of parameters for each algorithm using scikit-learn that yields the best results.
- Combine the prediction of multiple models into an ensemble prediction using ensemble techniques.

6. Present results

Make prediction on unseen data

- Using an optimal model tuned by scikit-learn to make predictions on unseen data.
- Creating a standalone model using the parameters tuned by scikit-learn
- Saving an optimal model to file for later use

Q&A

The end