Unsupervised Machine Learning with Python

COPYRIGHT SATISH REDDY 2021

1.1 Introduction

Copyright Satish Reddy 2021

What is Machine Learning?

Definition (adapted from Wikipedia page on Machine Learning)

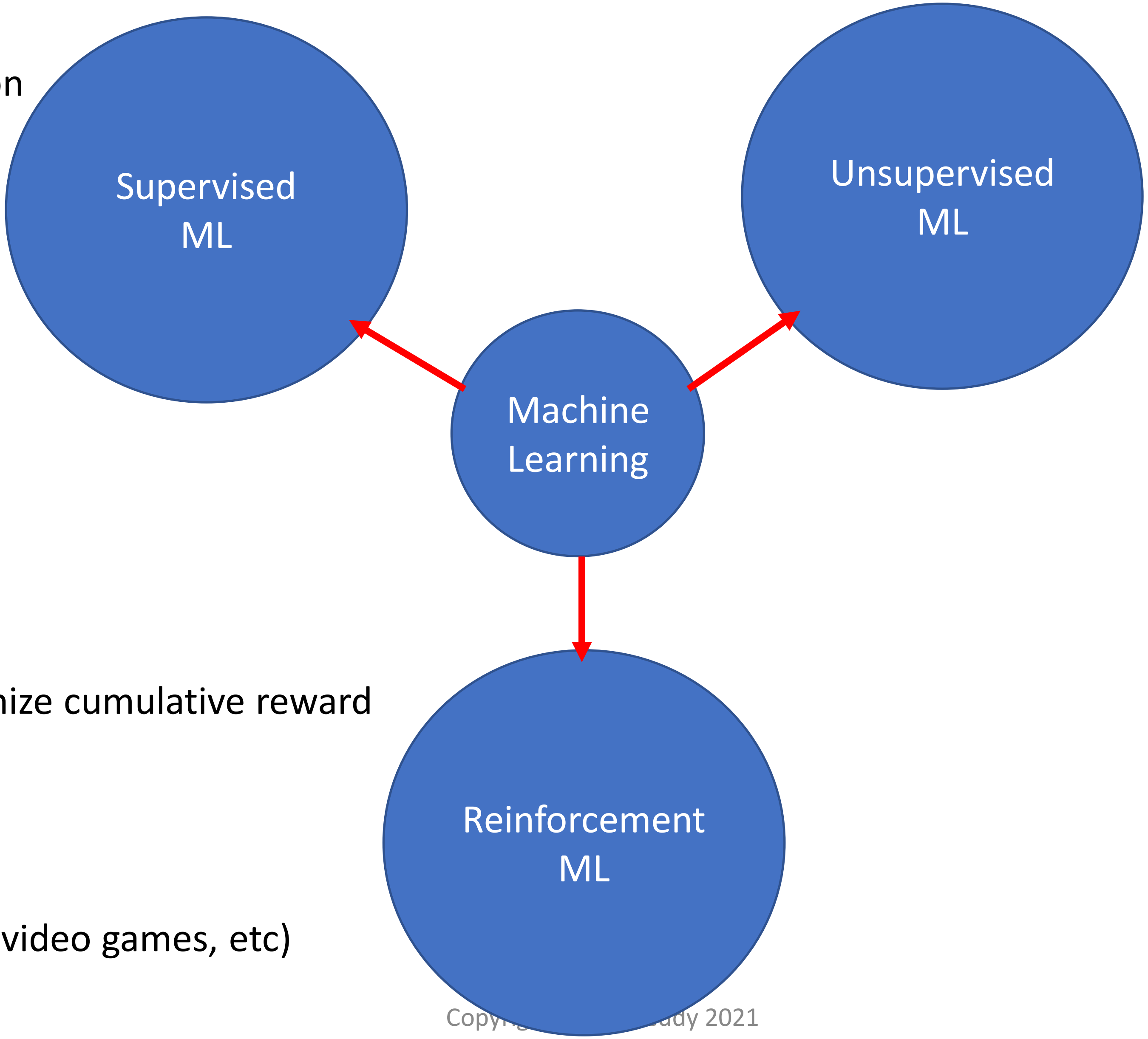
- Machine Learning is the study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead

Copyright Satish Reddy 2021

Types of Machine Learning

Fit function to data
Use function for prediction

Applications:
-House Price prediction
-Image Classification
-Spam Filtering
-Language Translation



Machine Learning

Supervised ML

Unsupervised ML

Reinforcement ML

Find patterns in data

Applications:
-Customer Segmentation
-Document Clustering
-Image Grouping
-Anomaly/fraud detection

Find strategy to maximize cumulative reward

Applications:
-Industrial Control
-Robotics
-Game Playing (chess, video games, etc)

Copyr [illegible] udy 2021

Unsupervised Machine Learning

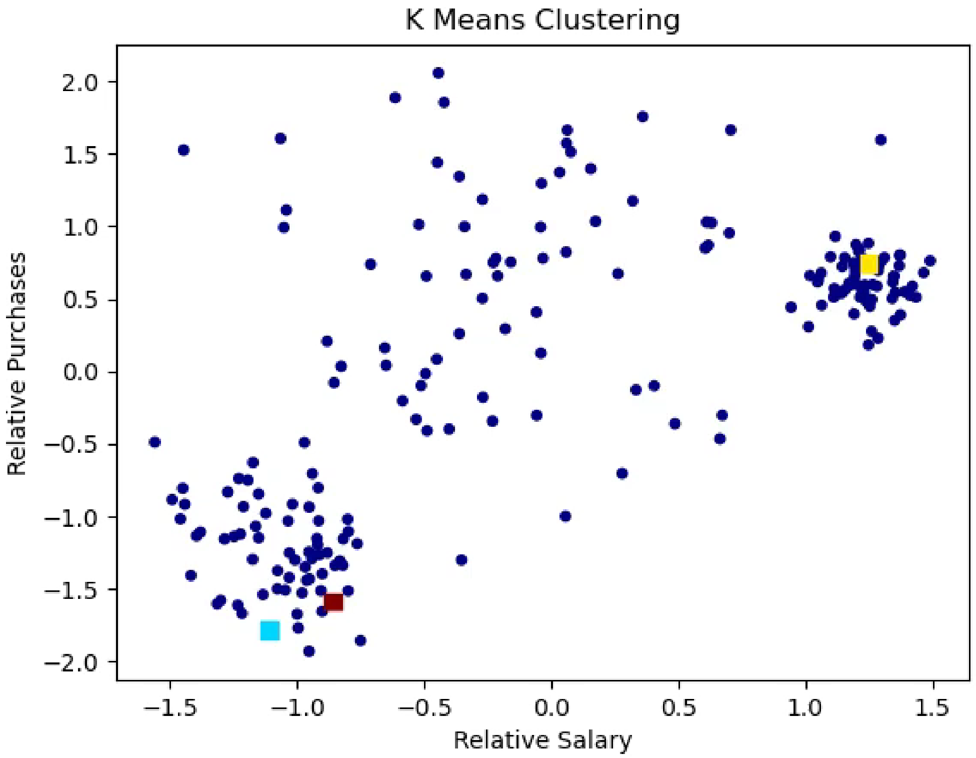
Definition (see Wikipedia page on Unsupervised Learning)

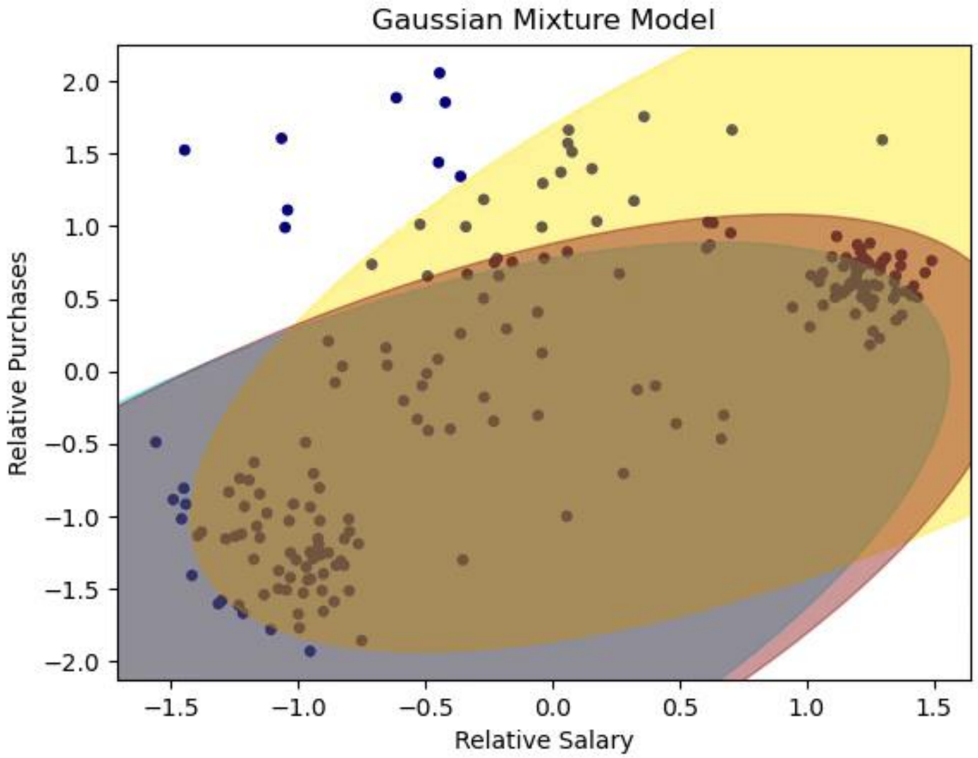
- Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a dataset with no pre-existing labels and with a minimum of human supervision.

Copyright Satish Reddy 2021

Clustering Algorithms

- Example: Find clusters in customer data
- Purpose: create specialized marketing campaigns for each cluster



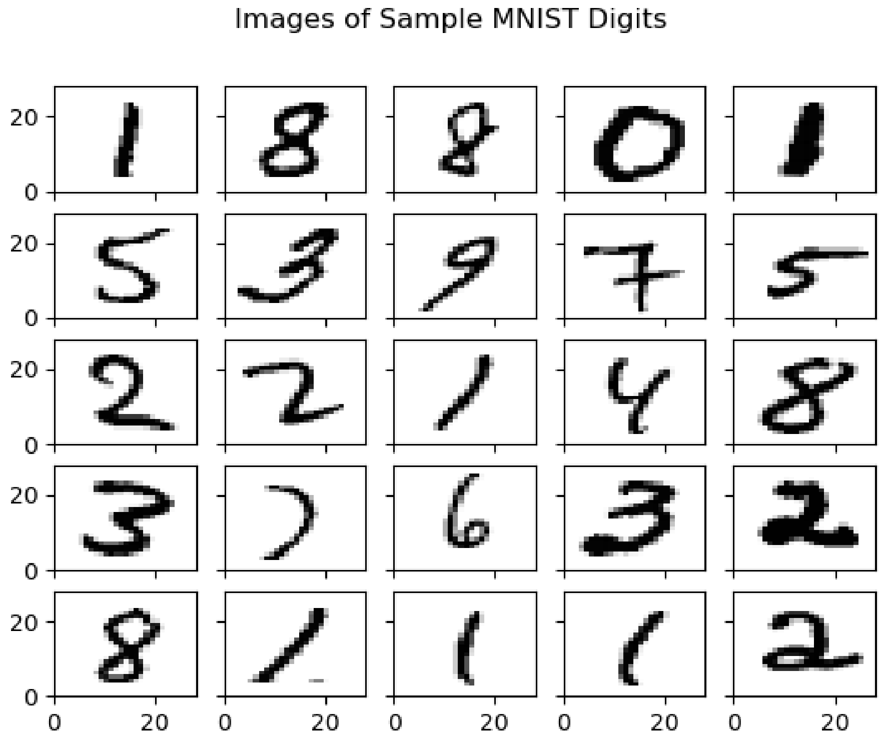


Copyright Satish Reddy 2021

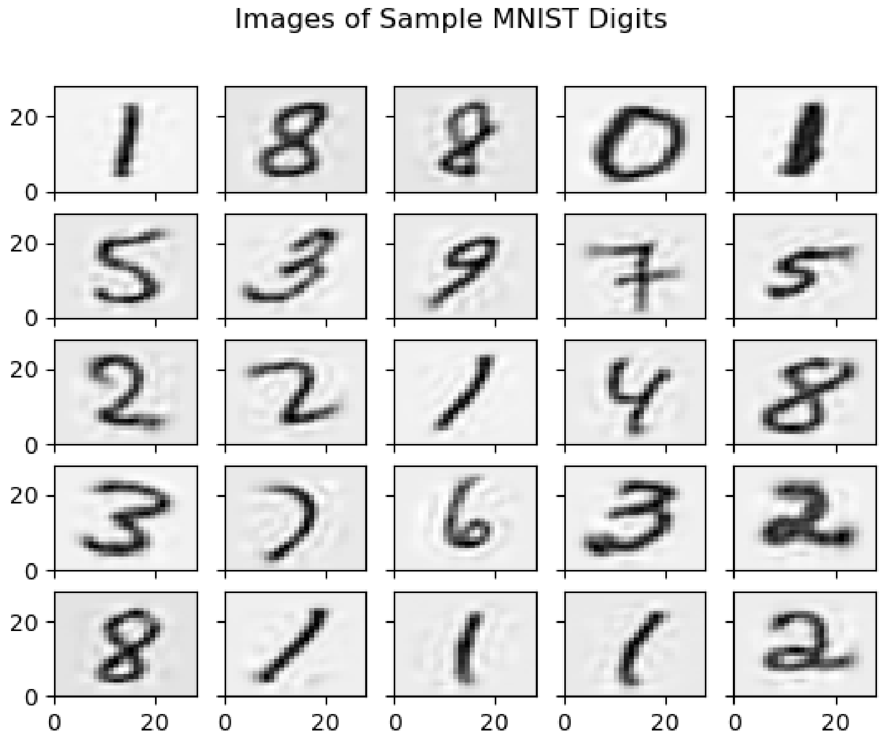
Dimension Reduction Algorithm

- Example: Reduce dimension of images (while retaining essential features)
- Purpose: use to speed up calculations

Original: 784 dimensions

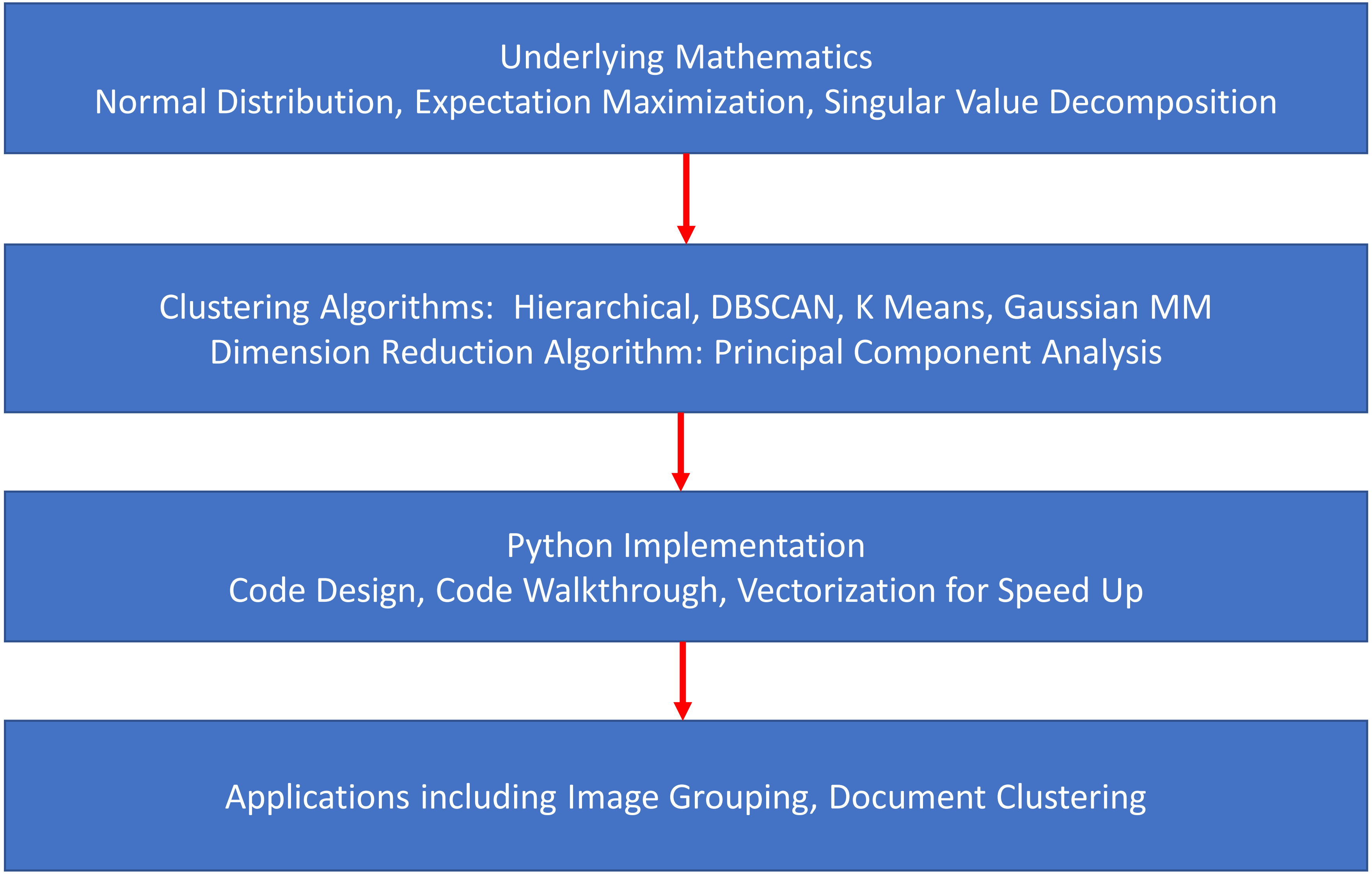


Reconstructed: using 78 principal components/dimensions



Copyright Satish Reddy 2021

What is in this Course?



Underlying Mathematics
Normal Distribution, Expectation Maximization, Singular Value Decomposition

Clustering Algorithms: Hierarchical, DBSCAN, K Means, Gaussian MM
Dimension Reduction Algorithm: Principal Component Analysis

Python Implementation
Code Design, Code Walkthrough, Vectorization for Speed Up

Applications including Image Grouping, Document Clustering

Copyright Satish Reddy 2021

Course Approach

1. Video Lectures
2. Jupyter Notebook Demos
    - Present examples of python packages and functions used in course
3. Code Walkthroughs
    - Explain code design and python implementation of algorithms
4. Exercises with Solutions
    - Give student opportunities for additional practice

Copyright Satish Reddy 2021

Course Outline:

- Section 01: Introduction
- Section 02: Python Demos
- Section 03: Review of Mathematical Concepts
- Section 04: Hierarchical Clustering
- Section 05: DBSCAN
- Section 06: K Means Clustering
- Section 07: Gaussian Mixture Model
- Section 08: Comparison of Clustering Algorithms
- Section 09: Principal Component Analysis
- Section 10: Case Studies
- Section 11: Summary and Thank You

Copyright Satish Reddy 2021

Unsupervised Machine Learning with Python

COPYRIGHT SATISH REDDY 2021

1.2 About this Course

Copyright Satish Reddy 2021

Course Prerequisites

Linear Algebra

- Should be familiar with vectors, transpose, matrices, matrix multiplication, inverses, and determinants

Probability and Statistics

- Should be familiar with basic probability and statistics, including normal distributions

Python Programming

- Should be able to write and run Python 3 programs in Jupyter notebooks and in the command window

Multivariable Calculus (Optional)

- Students familiar with multivariable calculus will be able to follow derivation of Gaussian Mixture Model approach presented in separate PDF document

Copyright Satish Reddy 2021

Audience for this Course

This course is suitable for:

1. Students without any previous experience with unsupervised machine learning
2. Students who have knowledge of the subject and would like a refresher and/or gain a more detailed understanding of math, algorithms, and code development

Copyright Satish Reddy 2021

How to Get the Most from this Course

Learning is not a spectator activity – active participation is required!!!

1. Take notes as you go through the material
2. Do the programming
   - Code design videos describe structure
   - Code walk through videos show the implementation
   - Students can do design and code development without watching these videos or can review videos first, then create their own design and do their own implementation
3. Do the exercises
   - Solutions provided
4. Ask questions on forums

Copyright Satish Reddy 2021

Why Code from Scratch?

- Why code from scratch when there are many Python packages for Unsupervised ML?
- It is my fundamental belief that to truly understand what is going on in an algorithm one must code from scratch!

Copyright Satish Reddy 2020

Unsupervised Machine Learning with Python

COPYRIGHT SATISH REDDY 2021

1.3 Resources and Set Up

Copyright Satish Reddy 2021

Course Github Site

https://github.com/satishchandrareddy/UnsupervisedML/

- Code
  - Codes and drivers for unsupervised machine learning algorithms
- Examples
  - Jupyter notebooks examples complementing lectures
- Exercises
  - PDF file of exercises and solutions
  - Solution files (Jupyter notebooks and python program files)
- Presentations
  - PDF files of presentations
- Resources
  - UnsupervisedML_Resources PDF has references and links to additional resources
  - ExpectationMaximization PDF with derivation details for Gaussian Mixture Model

Copyright Satish Reddy 2021

Resources and Set Up

- Instructor will use Windows 10 machine
  - Course material not specific to Windows – can use MacOS or Linux
- All code examples written in Python
- Course will run programs using
  - Jupyter Notebook: demos
  - Command Window: code development walkthroughs
- Should have text editor compatible with Python for writing and editing programs
  - Examples: Atom, Sublime, Notepad+, etc
  - Instructor will use Sublime, but you can use your favourite editor

Copyright Satish Reddy 2021

Packages used in Course

Component	Version	Description/Comments
Python	3.8.3	
NumPy	1.18.5	Package for scientific computing. In this course, numpy array is the principal container used to hold data
Matplotlib	3.2.2	Package for plotting
pandas	1.0.5	Package containing data structures and data analysis tools. We will its functionality to load data from csv file.
scikit-learn	0.23.1	Package for supervised and unsupervised learning We will used its functionality for (a) generating datasets (b) text processing
IPython	7.16.1	Package of tools for interactive use of Python. We will use for animation in Jupyter notebooks.
wordcloud	1.8.1	Package for generating word clouds (useful for document clustering)
copy, pathlib, time		These packages are part of the python release

Copyright Satish Reddy 2021

Anaconda Platform

If you don’t have Python on your machine, probably best to install the Anaconda Platform

- Anaconda Platform is distribution of Python for scientific computing
- We will use Anaconda prompt window and Jupyter notebooks to run programs
- Anaconda installation comes with packages: numpy, matplotlib, pandas, sklearn, IPython (we only need to install wordcloud)
- You don’t need to use the Anaconda Platform. You are free to use any python platform, as long as you can install required packages, run python in the command window, and use Jupyter notebooks.

Copyright Satish Reddy 2021

Package Installation

- Anaconda distribution comes with packages: numpy, matplotlib, pandas, sklearn, IPython
- Install wordcloud manually
  - Open Anaconda prompt window
  - Make sure you are in base environment
  - Issue command: conda install wordcloud
- For simplicity, I have not created a separate conda environment on my machine for this course. You are free to create an environment.

Copyright Satish Reddy 2021

1.3 Resources and Course Set Up Demo

1) Download and unzip resources from Course Github Site

https://github.com/satishchandrareddy/UnsupervisedML/

2) Installation of Anaconda

https://www.anaconda.com

3) Installation of wordcloud:

- In Anaconda Prompt Window (base environment) issue command: conda install wordcloud

4) Test in Anaconda Prompt Window

- Open Anaconda Prompt window
- Change to directory UnsupervisedML/Code/Clustering
- Issue command: python commandwindowtest.py

5) Test in Jupyter Notebook

- Open Jupyter Notebook
- Change to directory UnsupervisedML/Examples/Section01
- Open JupyterNotebookTest.ipynb

Copyright Satish Reddy 2021