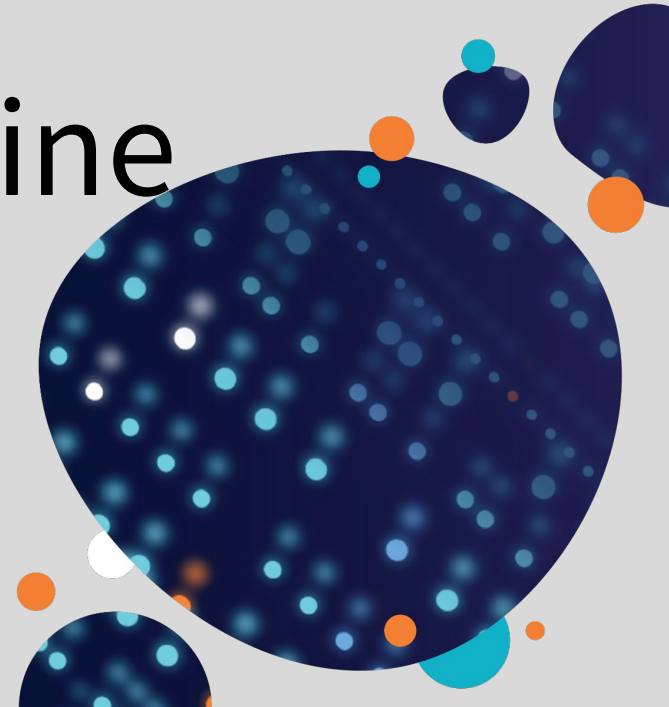
data science
student society

Basic Applied Machine Learning

Barry Xue and Garrett Lam



OVERVIEW

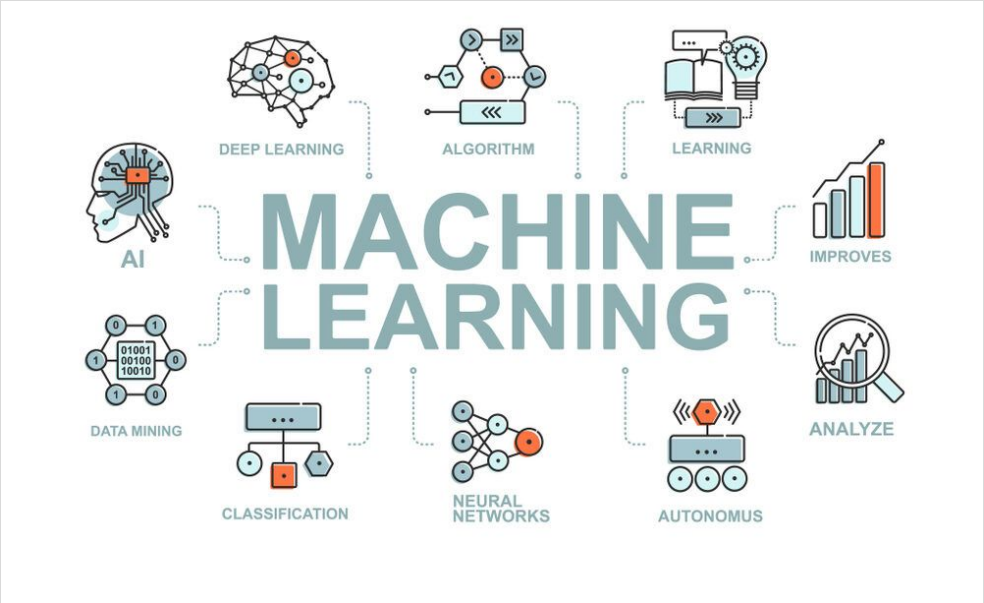
data science
student society

- What is Machine Learning?
- ML vs Statistical Modeling
- Categories of ML
- Applications of ML
- Life Cycle of a ML Application
- Jupyter Notebook Demo

WHAT IS MACHINE LEARNING?

data science
student society

A subfield of artificial intelligence (AI) that utilizes algorithms and data to learn, make predictions, and generate insights without explicitly programming it to do so



MACHINE LEARNING VS STATISTICAL MODELLING

data science student society

Statistical Modelling: formalizing a mathematical relationship between one or more random variables in the form of a mathematical equation

Machine Learning	Statistical Modeling
Makes less/no assumptions about data	Mathematically based, many assumptions must be made about data
Can use a large range of datasets (big data)	Limits to small datasets (cannot have too many attributes/features)
Streamlining the training process with algorithms	Focus on model explainability
Goal: strong predictive ability + usability	**Goal:** best estimate of data

CATEGORIES OF MACHINE LEARNING

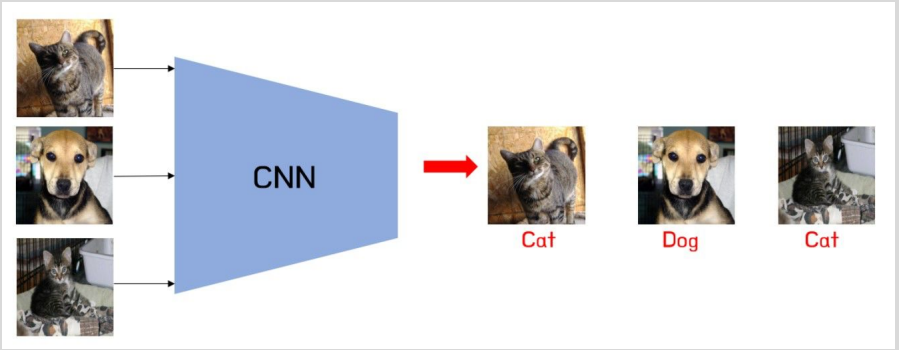
data science student society

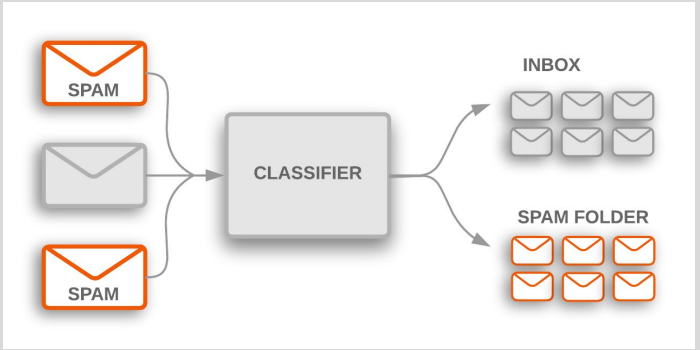
1. Supervised Learning

Uses a trained dataset to train an algorithm into classifying data or predicting outcomes accurately

- Training and testing data is given
- **Goal:** extrapolate knowledge learned from training data to the testing data

Classification is a supervised learning task, examples include:





CATEGORIES OF MACHINE LEARNING

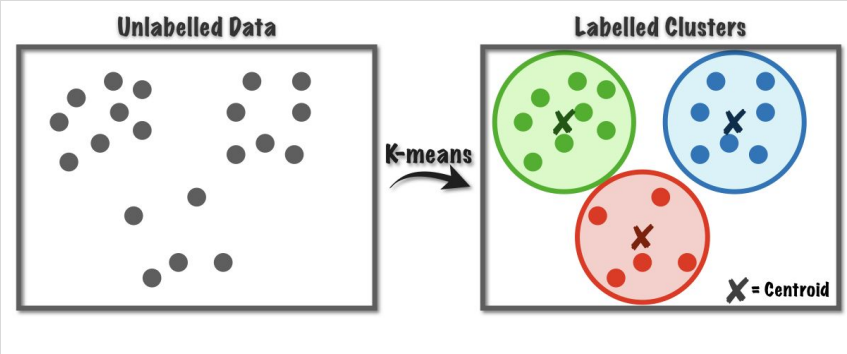
data science student society

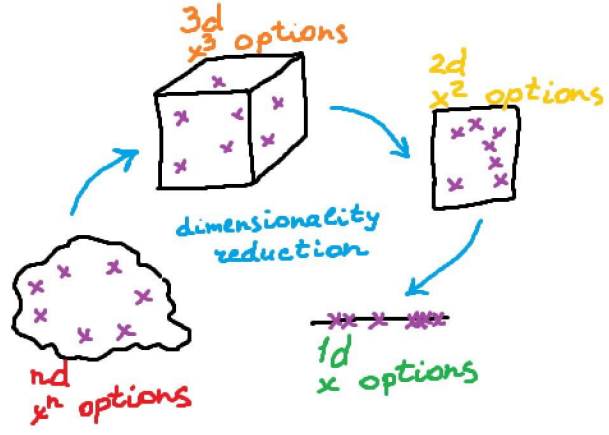
2. Unsupervised Learning

Uses generic machine learning algorithms to analyze and discover hidden patterns in data

- Only testing data is given
- **Goal:** learn patterns and rules about the testing data

Clustering is an unsupervised learning task, examples include:





CATEGORIES OF MACHINE LEARNING

data science student society

3. Reinforcement Learning

A technique that utilizes an agent to learn in an environment by using positive and negative feedback from its own actions and experiences

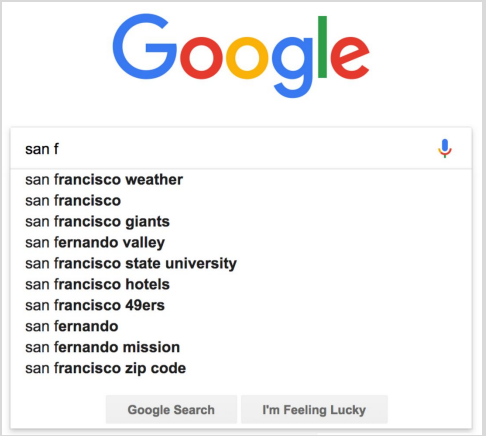
- **Goal:** discover actions that maximizes reward of the agent

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Learning with a labeled training set. Email spam detector with training set of already labeled emails.	Discovering patterns in unlabeled data. Cluster similar documents based on the text content.	Learning based on feedback or reward. Learn to play chess by winning or losing.

APPLICATIONS OF MACHINE LEARNING

data science student society

- Netflix Recommendations (Recommender Systems)
- Google Search (Natural Language Processing)
- Facial Recognition (Computer Vision)



NETFLIX RECOMMENDATIONS

data science student society

Netflix uses the Netflix Recommendation Algorithm (NRE) to automate recommendations for users:

- Combination of multiple ML algorithms
- Two-tiered row based ranking system

Data used:

- Most viewed genres
- Rewatched titles
- User ratings (Thumbs up/down)
- “My List” section
- Viewing habits

GOOGLE SEARCH

data science
student society

Google uses transformer based architecture neutral networks called **Bidirectional Encoder Representations from Transformers (BERT)**, and more recently **Siamese Multi-Depth Transformer-based Hierarchical (SMITH) Encoder**

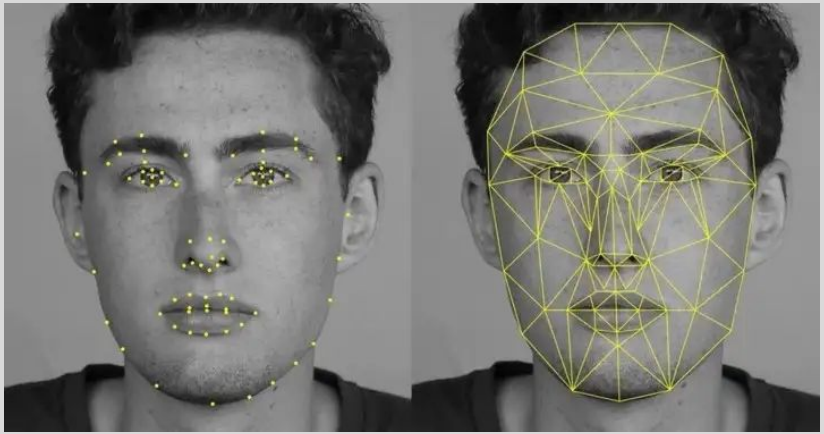
- Both NLP algorithms: BERT short queries, SMITH long queries
- Uses semantics, word segmentation, most popular global searches, and browsing history

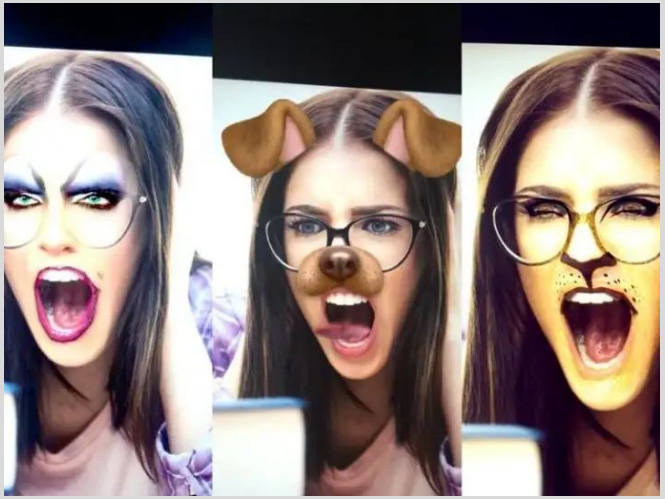
SNAPCHAT FILTERS

data science
student society

Snapchat uses face detection, facial landmarks, and image processing to produce filters

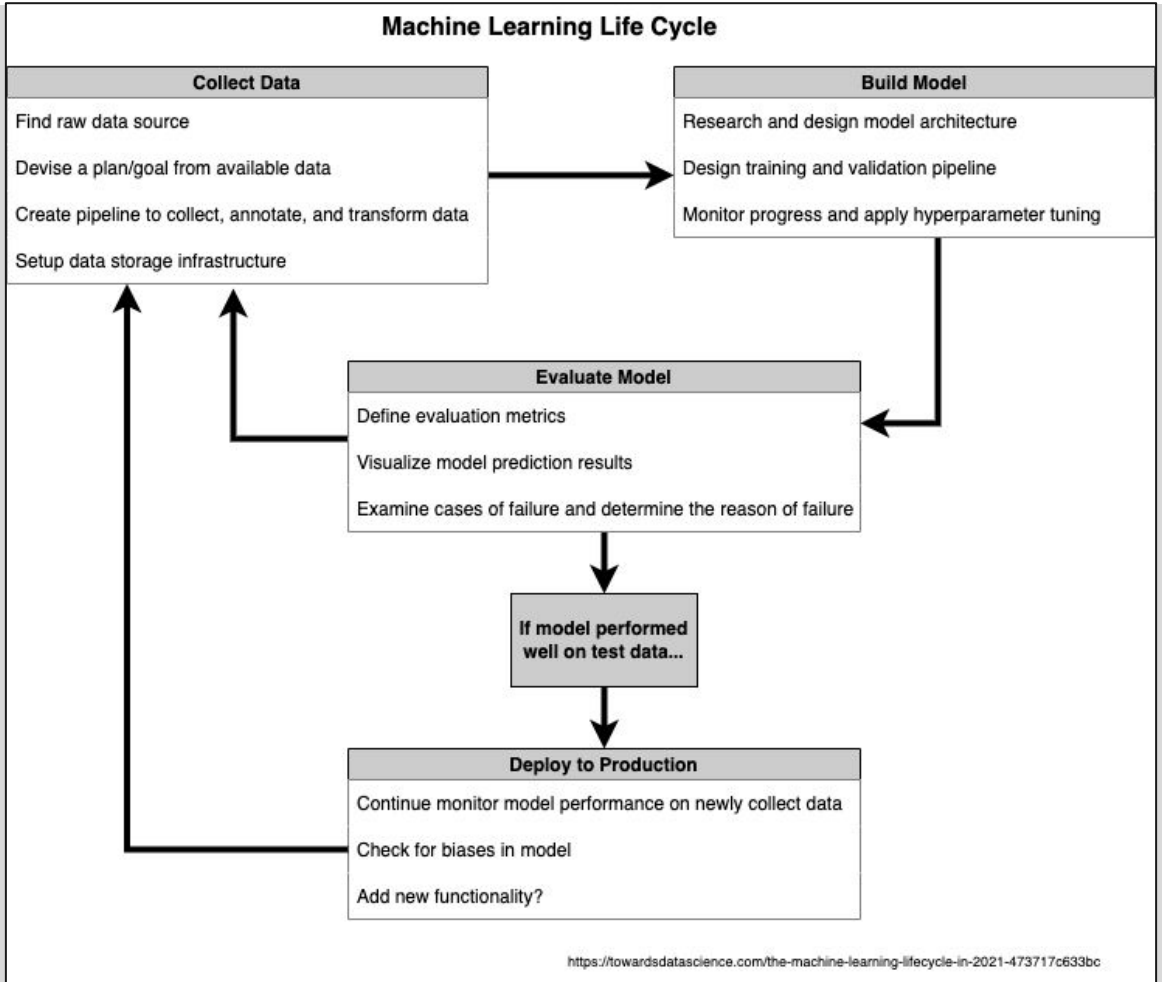
- **Face Detection:** Support Vector Machine (SVM)
- **Facial Landmark:** Regression Trees
- **Image Processing:** Active Shape Model





data science
student society

“Models are a small part of building a ML application”



data science
student society

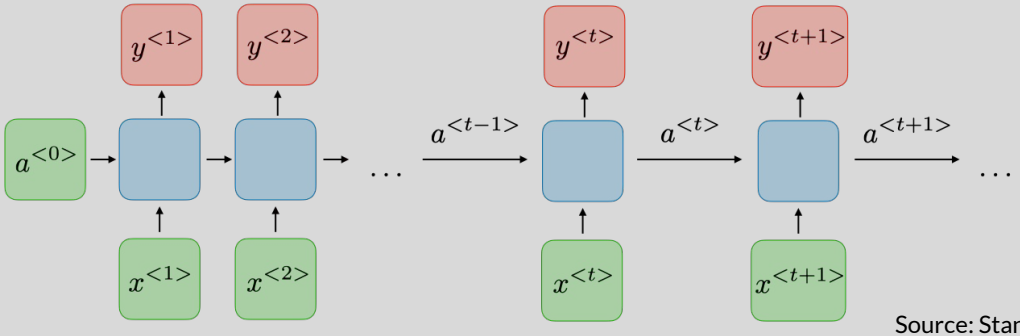
DEMO

Stock Market Trend Prediction

Objective: Using the Moving Average and a LSTM RNN model to predict a stock's closing prices.

What you will learn:

- Moving average filter
- Recurrent Neural Network
- Long short-term memory (LSTM)



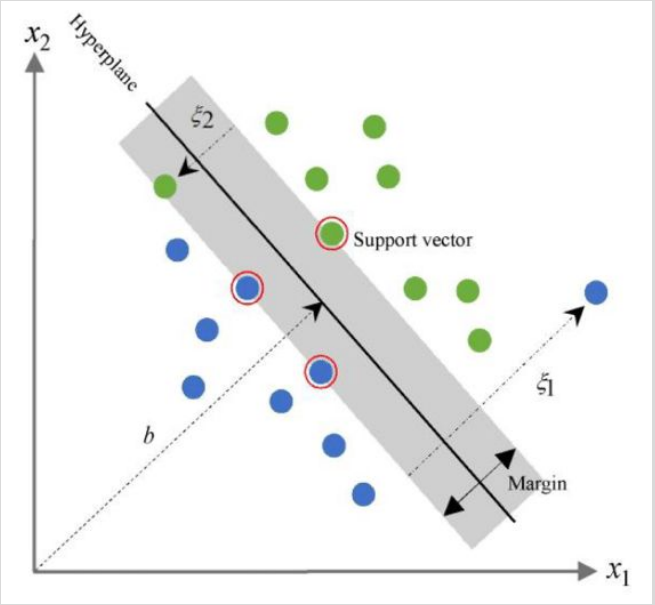
Source: Stanford CS 230 Lecture Note

Handwritten Digits Classification

Objective: classify pictures of handwritten digits (0-9) using classic ML techniques.

What you will learn:

- Simple practice of image processing
- Support Vector Machine (SVM)
- Convolution Neural Network (CNN)



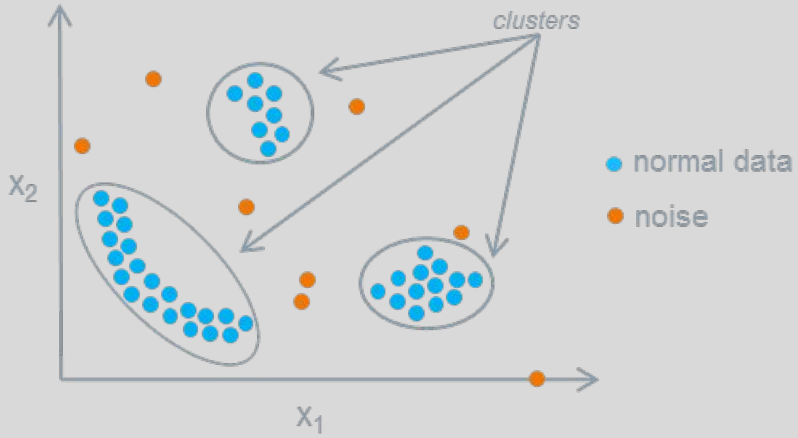
Source: ResearchGate - SVM

Anomaly Detection

Objective: find abnormal data points given a noisy dataset.

What you will learn:

- K-means Clustering
- Density-based clustering technique (specifically, DBSCAN)



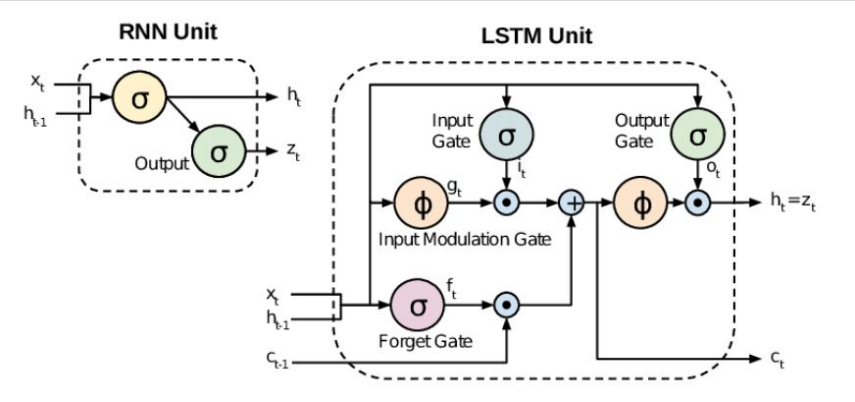
Source: Siemen's MindSphere Documentation

Appendix A.1: RNNs and LSTM

A RNN cell recurrently updates its hidden weights (a(t)) with each iteration of new input data (x(t)).

Traditional RNN has some known issues:

- High computation cost and long runtime
- Vanishing gradient after long sequence of updates



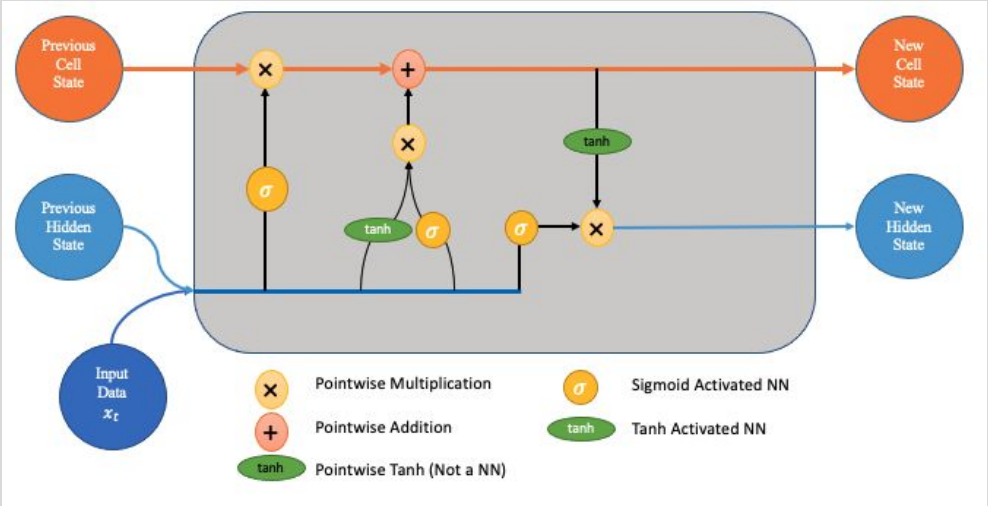
Appendix A.2: RNNs and LSTM

LSTM is a RNN unit designed to fix those issues. It uses three gates:

- Forget Gate
- Keep Gate (Input Gate)
- Output Gate

The gating mechanism allows a chain of LSTM cells to:

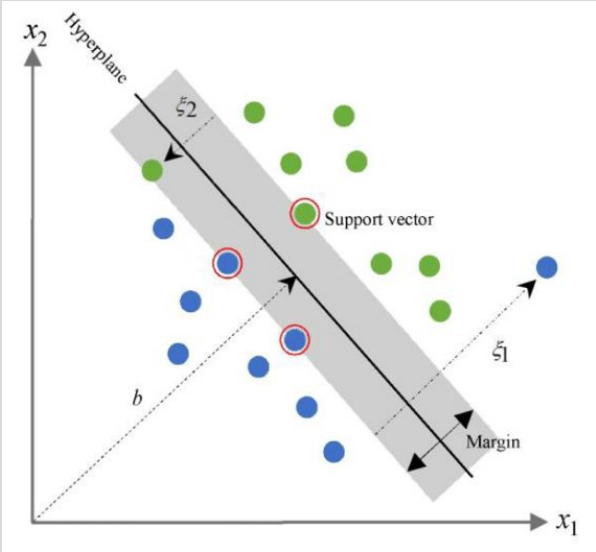
- Skip input at specific time step
- More expressive with longer sequences, less likely to suffer from vanishing gradient



Appendix B: Support Vector Machine

Support Vector Machine aims to maximize the margin between the closest data points, with an allowed error rate.

- The points circled in red are the ‘support vectors’ that define the margin size.
- The green data point is on the incorrect side of the hyperplane, but it is within the allowed error rate.
    - This helps the model generalize.



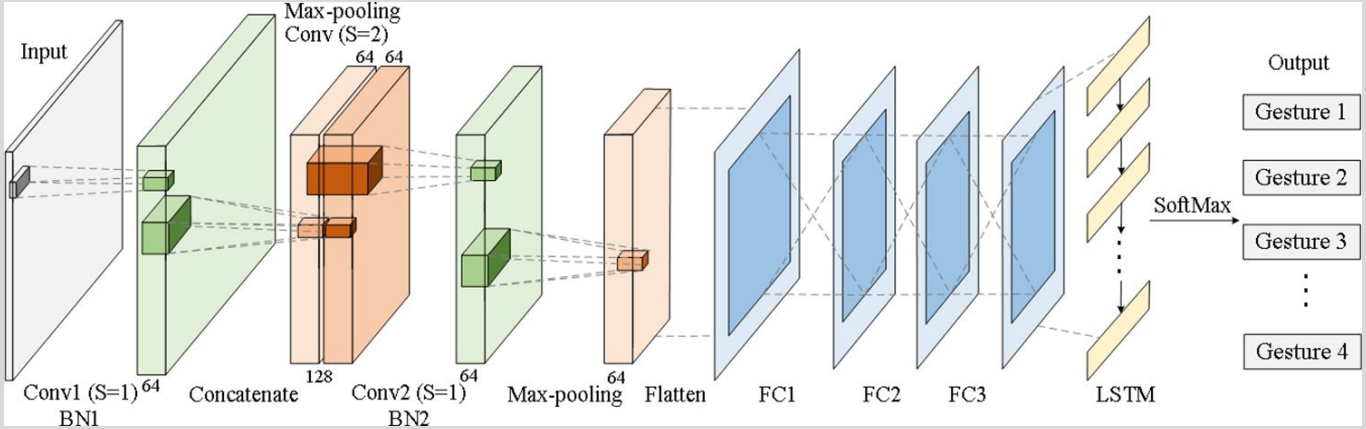
Appendix C: CNNs

Convolution Neural Network has three essential components:

1. Convolutional layers
2. Pooling layers
3. Non-linear activation

Other components:

1. Dropout layer
2. Normalization Transformation
3. Additional model (LSTM, MLP, etc)



Appendix D: DBSCAN

Uses a density-based approach to search for cluster

- With a given radius of each point, and iteratively move from point-to-point

This is different from kmeans, because:

- No fixed number of cluster
- More geometrically flexible