ML/AI Introduction + Hands-on Exercise

Umair Mohammad, Florida International University

OpenCI Award: OCI 21-38296

Scope

- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

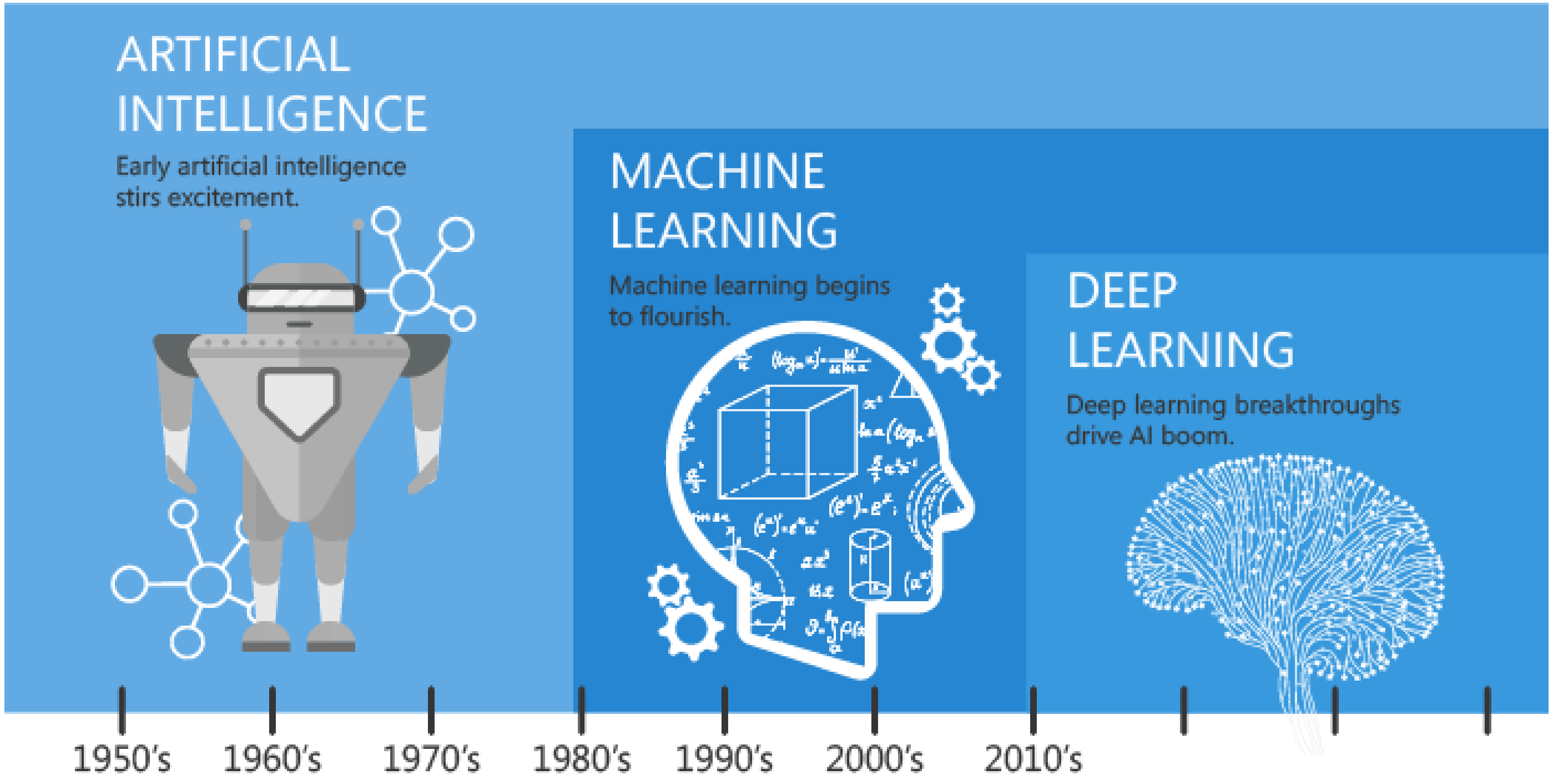
- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

Short Intro

- SaeedLab at FIU School of Computing
- Two Major thrusts
- Machine learning model development for
  - Proteomics
  - Neurological Disorders
- Diagnosing ASD, AD
- Detecting and Predicting Epileptic Seizures

- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

AI vs ML vs DL

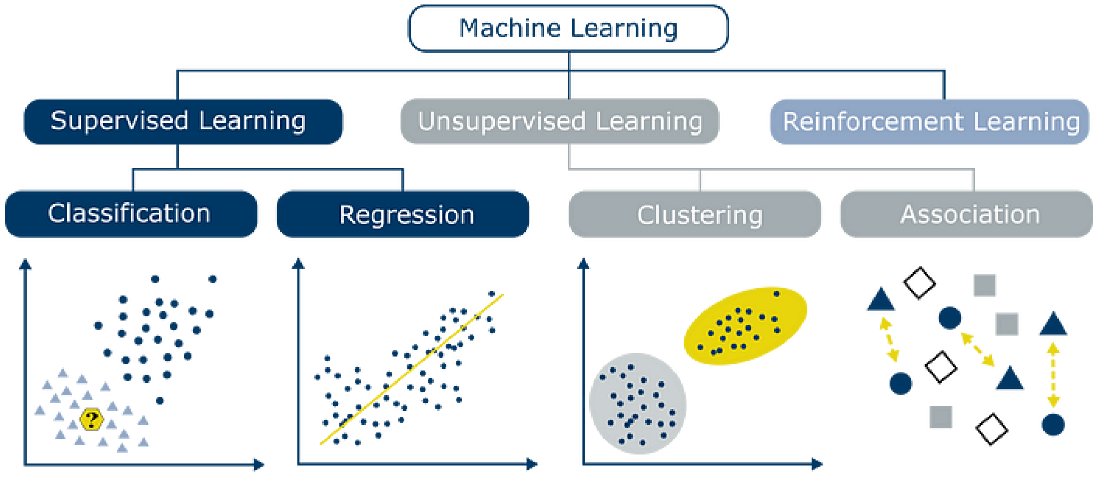

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

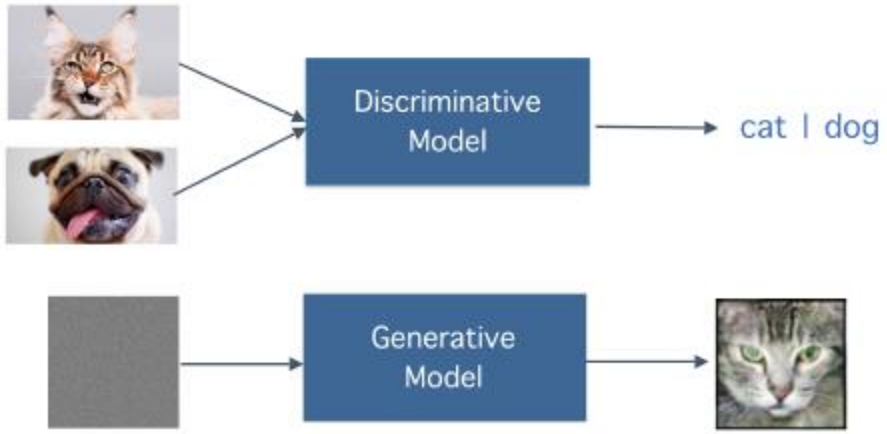
Types of ML Tasks

- Prediction/Forecasting
- Classification/Recognition
- Object Detection
- Image Segmentation
- Recommendation
- Speech-to-text, etc.
- Grouping
- PDF estimation
- Dimensionality Reduction
- Similarity Matching
- Link Prediction
- Anomaly Detection
- Querying
- Synthesis

Learning types

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi/Self-Supervised
- {In/Con/Trans}-Duction
- Summary:
  - Classification Phase
  - Generative Phase

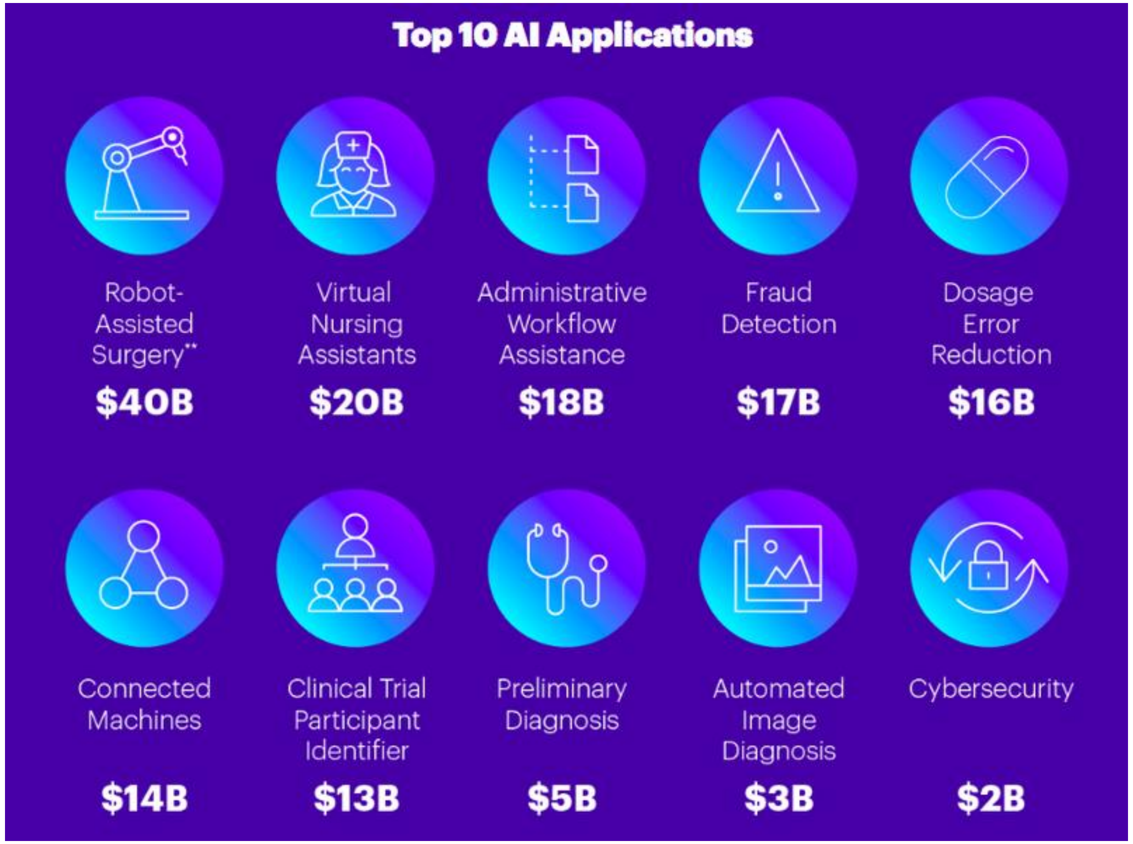




- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

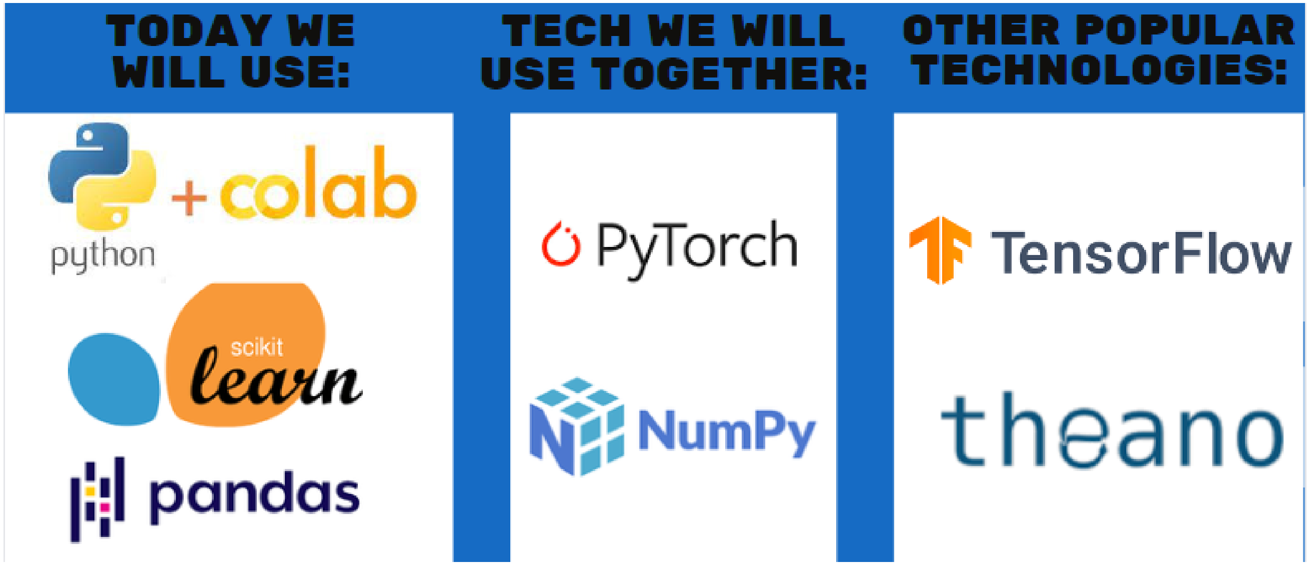
Applications

- Autonomous Cars
- Unmanned Aerial Vehicles (UAVs)
- Medical Diagnostics/Detection
  - Arrythmia, Seizure, Cancer, etc.
- Amazon/Social Media Recommender Systems
- Content Generation (Ethical?)
- Bank/Credit Card Fraud Detection
- LLM’s e.g. ChatGPT and Bard
  - AI-Whisperer’s?



Why Now? Tools and Tech

- AI-Winter 1960’s
  - Algorithmic Limits
  - Hardware Limits
- Resurgence
  - 21^{st} century up-to 2010
  - New algorithms
  - Open-source models
- New “Programming”
- Applicable to any area!



- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

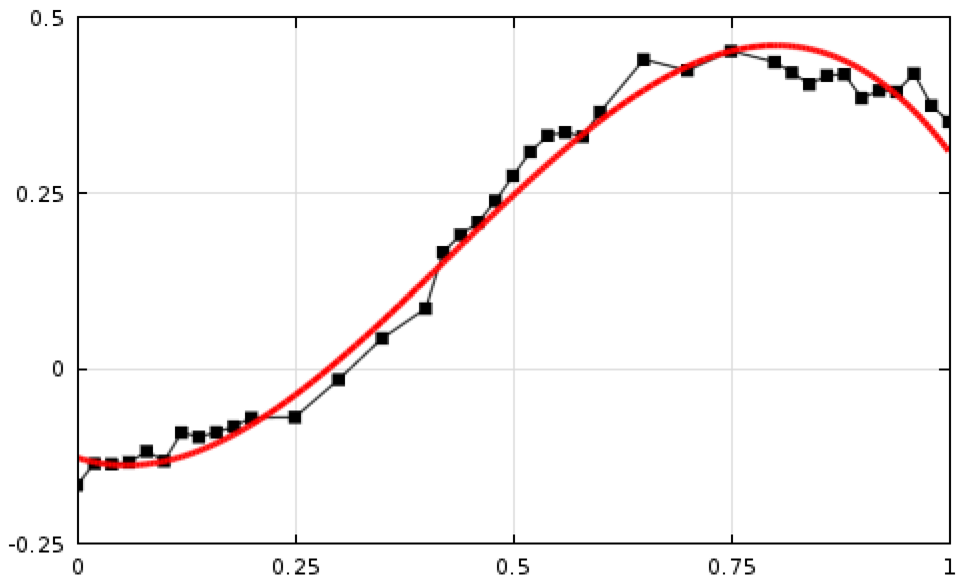
Other tools and technologies

- General-purpose Graphical Processing Units (GPGPUs)
  - Nvidia
- High Performance Computing Clusters (HPC's)
  - Expanse
- Cloud Computing
  - Microsoft Azure, Amazon Web Services (AWS), Google Cloud, etc.
- Open-source and reproducible models
- Python-based libraries; Weights and Biases

- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

Linear Regression

- All of you have probably done ML in middle school!
- Predict/Forecast a scalar-valued target
- Curve-Fitting simplest example
- Equation of a straight line



Gradient Descent

Model: y is a linear function of x:

$$y = \mathbf{w}^\top \mathbf{x} + b$$

y is the prediction

$\mathbf{w}$ is the weight vector

b is the bias

$\mathbf{w}$ and b together are the parameters

$$w_j \leftarrow w_j - \alpha \frac{\partial \mathcal{J}}{\partial w_j}$$

$$= w_j - \frac{\alpha}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})\, x_j^{(i)}$$

Loss function: squared error

$$\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$$

$y - t$ is the residual, and we want to make this small in magnitude

The $\frac{1}{2}$ factor is just to make the calculations convenient.

Cost function: loss function averaged over all training examples

$$\mathcal{J}(w, b) = \frac{1}{2N} \sum_{i=1}^{N} \left(y^{(i)} - t^{(i)} \right)^2$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \left(\mathbf{w}^\top \mathbf{x}^{(i)} + b - t^{(i)} \right)^2$$

$$\mathbf{y} = \mathbf{X}\mathbf{w} + b\mathbf{1}$$

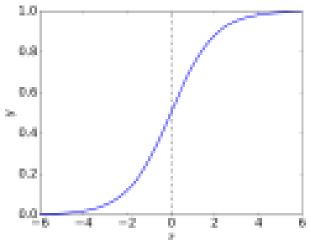
$$\mathcal{J} = \frac{1}{2N} \|\mathbf{y} - \mathbf{t}\|^2$$

Logistic Regression

- Similar to Linear regression
- A sigmoid activation is used
  - Cross-entropy loss
- Used for binary classification
- Outputs are probabilities
- Labels are discrete
- Continuous needed to be differentiable

The logistic function is a kind of sigmoidal, or S-shaped, function:

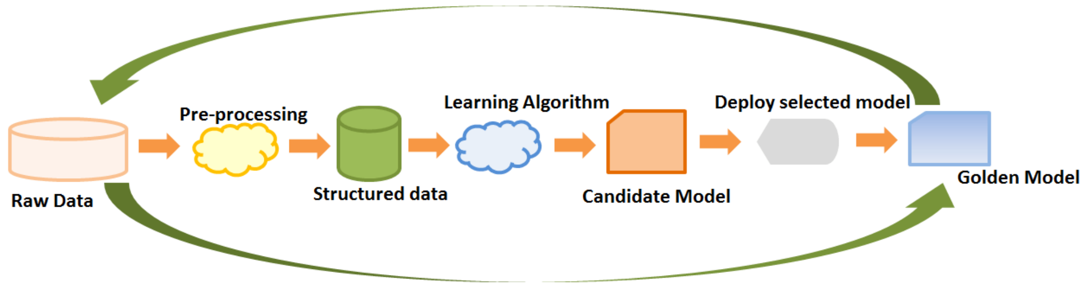
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

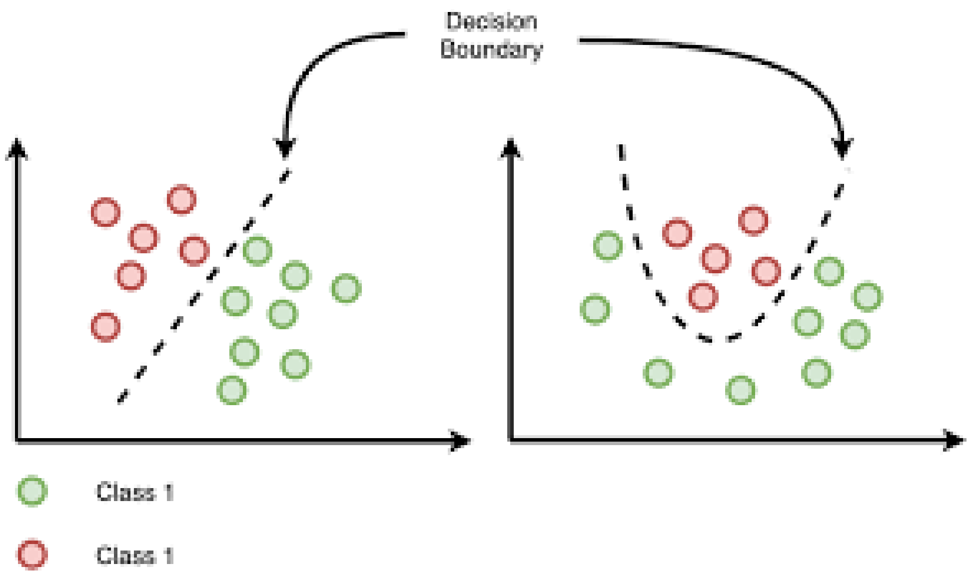
The ML Process

- Should I?
- Data Wrangling
- Preprocessing/Visualization
- Feature Extraction/Selection
- Model Selection
- Optimization
- Evaluation and Improvement



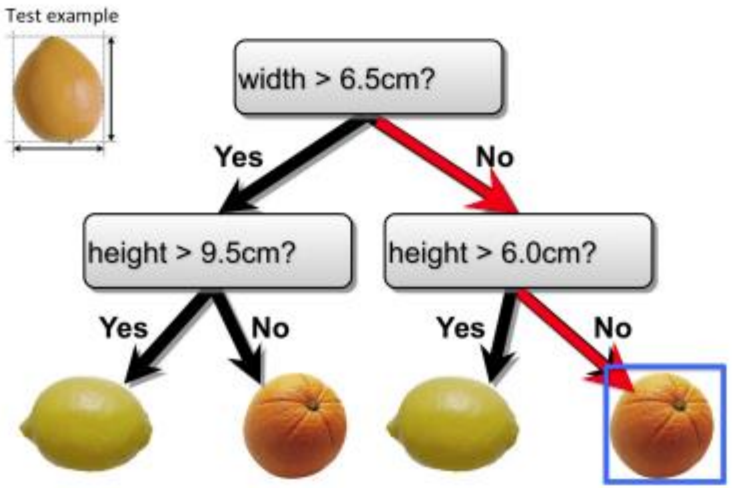
Linear Classification

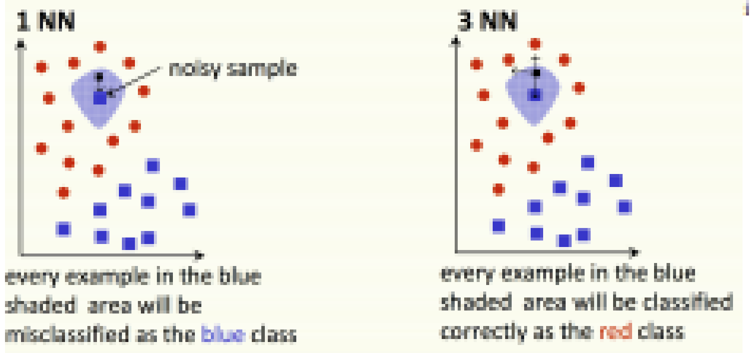
- Theory
  - Predict a discrete valued target
  - Linearly relate features
- Methods
  - Logistic Regression
  - Naïve Bayes
  - Support Vector Machines
  - Single-Layer Perceptron's



Other Algorithms

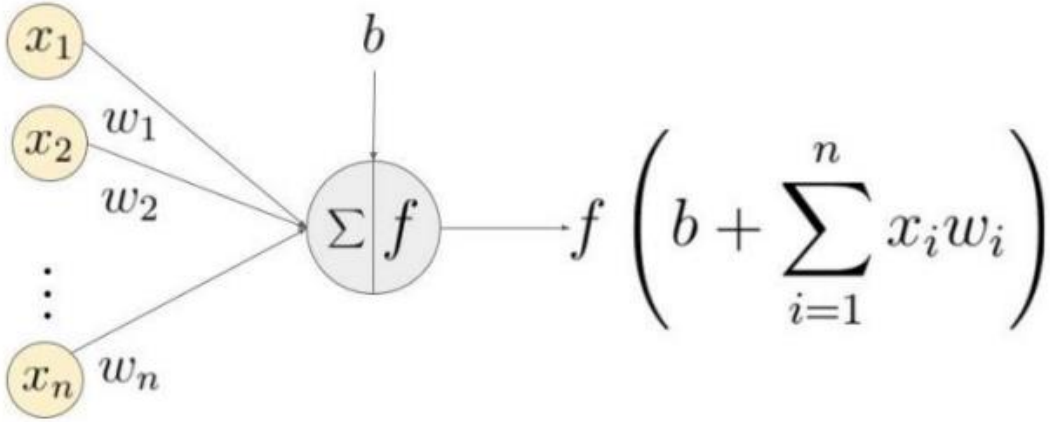
- Decision Trees
  - Binary trees with nodes, leaf's, etc.
  - Idea 2
- Random Forests
  - Bagging of decision trees
  - Decreases the variance, reduces over-fitting
- K-nearest Neighbor
  - Euclidean distance with k neighbors
  - Majority Voting
- Boosting and Bagging
- Ensemble Classifiers





Neural Networks (NNs)

- Multiple “hidden” layers
- Build upon single-layer
- Can do non-linear
- Feedforward
- Can approximate any function



- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

Backpropagation

How you would have done it in calculus class

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{2}(\sigma(wx+b)-t)^2 \\
\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial}{\partial w}\left[\frac{1}{2}(\sigma(wx+b)-t)^2\right] \\
&= \frac{1}{2}\frac{\partial}{\partial w}(\sigma(wx+b)-t)^2 \\
&= (\sigma(wx+b)-t)\frac{\partial}{\partial w}(\sigma(wx+b)-t) \\
&= (\sigma(wx+b)-t)\sigma'(wx+b)\frac{\partial}{\partial w}(wx+b) \\
&= (\sigma(wx+b)-t)\sigma'(wx+b)x
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b} &= \frac{\partial}{\partial b}\left[\frac{1}{2}(\sigma(wx+b)-t)^2\right] \\
&= \frac{1}{2}\frac{\partial}{\partial b}(\sigma(wx+b)-t)^2 \\
&= (\sigma(wx+b)-t)\frac{\partial}{\partial b}(\sigma(wx+b)-t) \\
&= (\sigma(wx+b)-t)\sigma'(wx+b)\frac{\partial}{\partial b}(wx+b) \\
&= (\sigma(wx+b)-t)\sigma'(wx+b)
\end{aligned}
$$

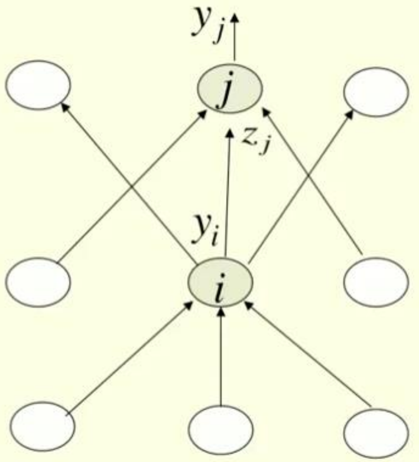
Computing the loss:

$$
\begin{aligned}
z &= wx+b \\
y &= \sigma(z) \\
\mathcal{L} &= \frac{1}{2}(y-t)^2
\end{aligned}
$$

Computing the derivatives:

$$
\begin{aligned}
\frac{d\mathcal{L}}{dy} &= y-t \\
\frac{d\mathcal{L}}{dz} &= \frac{d\mathcal{L}}{dy}\sigma'(z) \\
\frac{\partial \mathcal{L}}{\partial w} &= \frac{d\mathcal{L}}{dz}x \\
\frac{\partial \mathcal{L}}{\partial b} &= \frac{d\mathcal{L}}{dz}
\end{aligned}
$$

Backpropagating dE/dy



$$\frac{\partial E}{\partial z_j} = \frac{dy_j}{dz_j}\frac{\partial E}{\partial y_j} = y_j(1-y_j)\frac{\partial E}{\partial y_j}$$

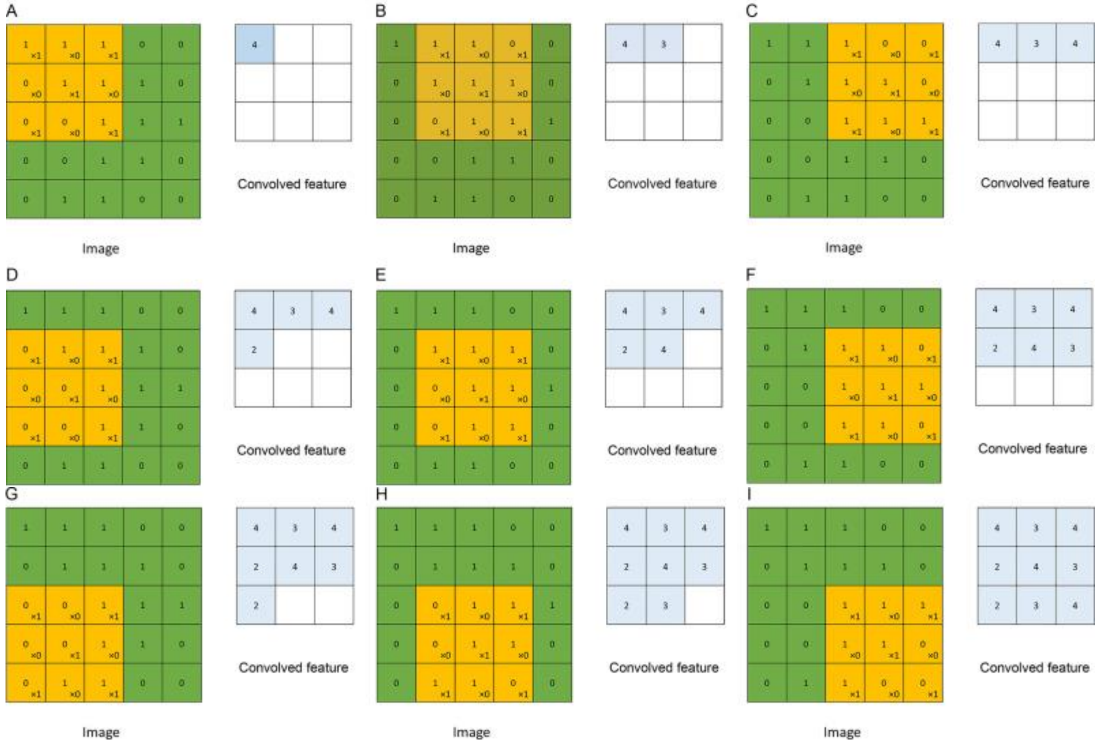
$$\frac{\partial E}{\partial y_i} = \sum_j \frac{dz_j}{dy_i}\frac{\partial E}{\partial z_j} = \sum_j w_{ij}\frac{\partial E}{\partial z_j}$$

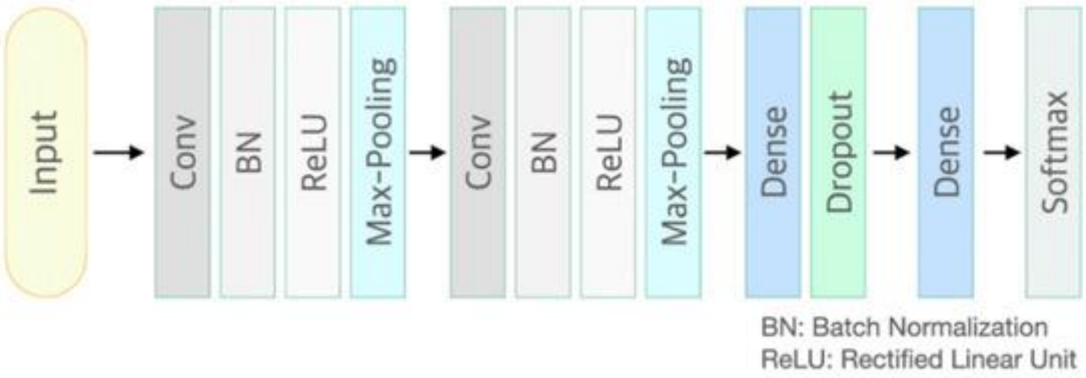
$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial z_j}{\partial w_{ij}}\frac{\partial E}{\partial z_j} = y_i\frac{\partial E}{\partial z_j}$$

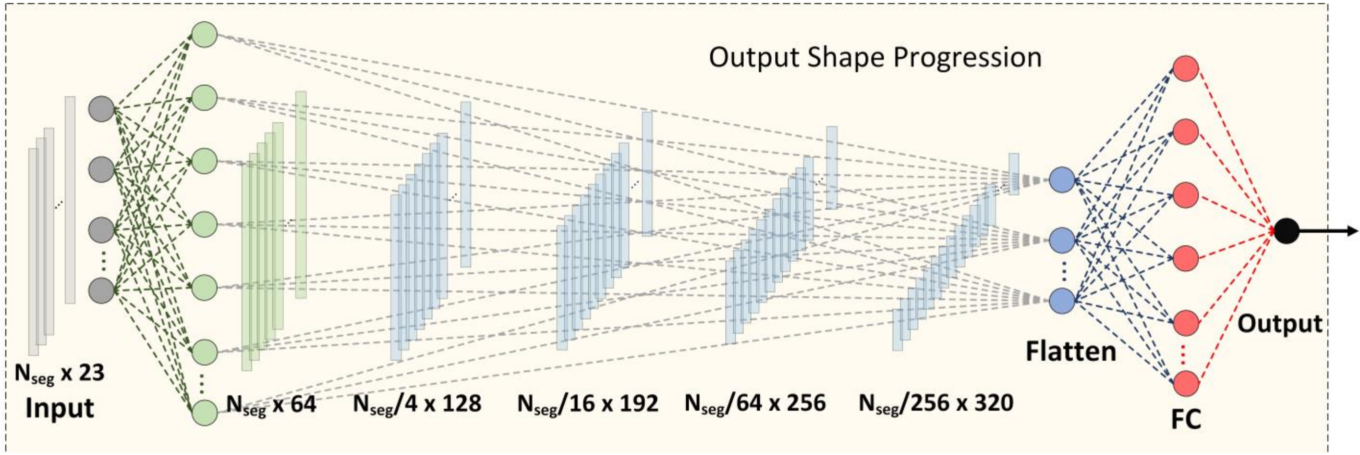
Deep NN's

- Multiple hidden layers
  - Complex Backpropagation using the chain-rule
- Convolutional
  - Discrete convolution – a DSP concept
  - Element-wise multiplication, summing and aggregation
  - Very powerful tool – represents filtering – Edge Detection
- Convolutional Neural Network (CNN)
  - Yann LeCunn and MNIST (The postal system started the AI revolution)

Convolution Neural Networks

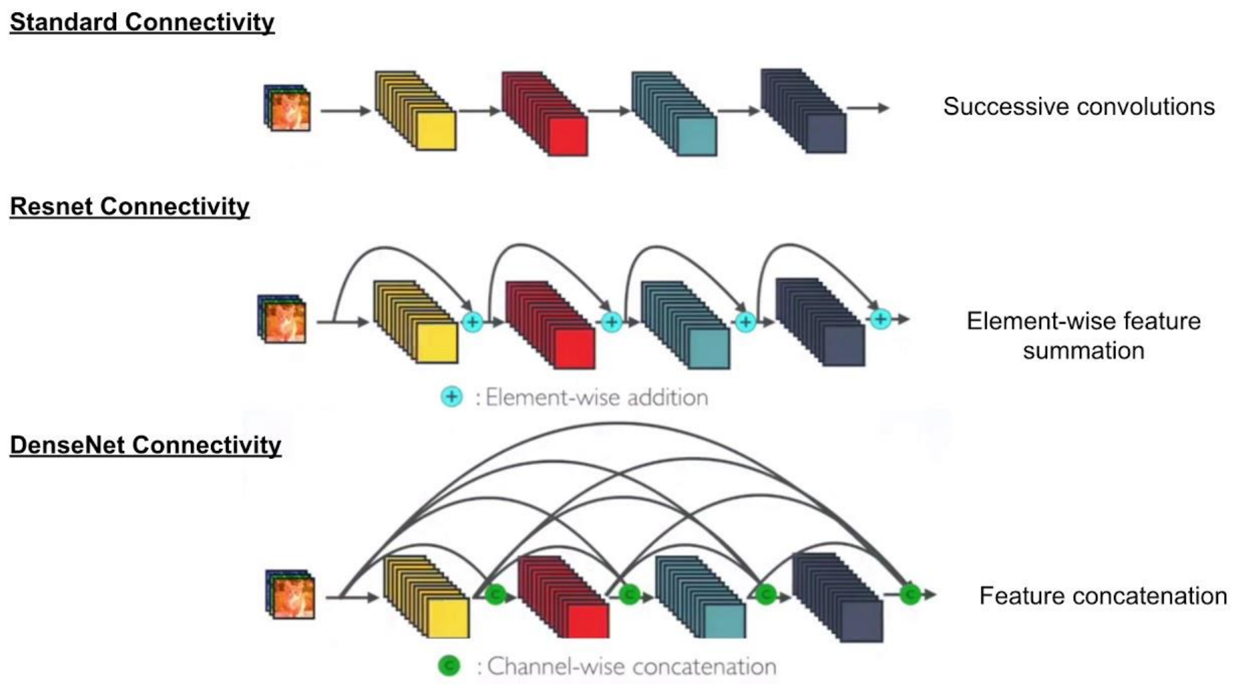






ResNets And DenseNets

- Residual NNs
- Dense NNs
- Feature Skip connections
- ‘Chain’ of convolutions
- May improve CNN performance
- Others
  - AlexNet, LeNet, Unet, etc.



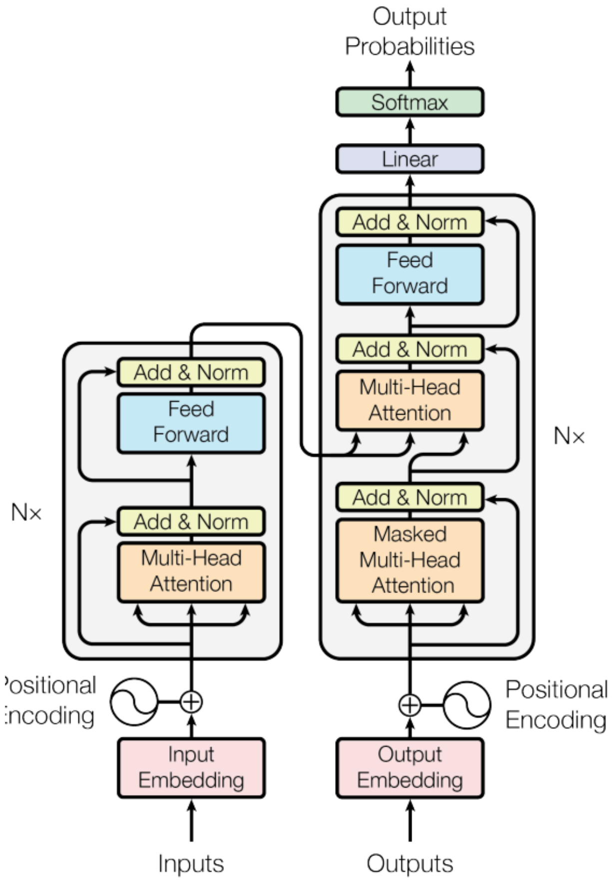
- Introduction
- AI, ML and DL
- How useful is ML/AI and why now?
- Tools and Technologies
- Regression
- Classical ML
- Deep Learning
- Generational Age

GANs, Diffusion and Tranformers

- General Adversarial Nets
  - Generator-Discriminator
  - Used to generate images such as DeepFake
- Diffusion
  - Stable vs non-stable
- Transformer Architecture
  - Similar Structure to an Autoencoder
  - Both encoder-decoder parts used for training
  - Only the Decoder part for inference

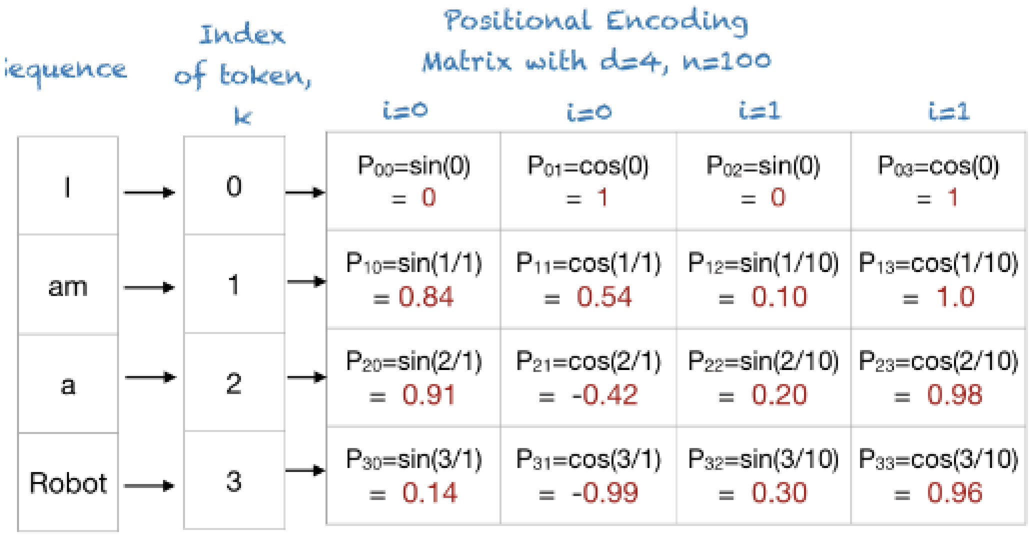
Self-Attention Layers

- Attention Mechanism
  - N Input and Outputs
  - Key, Query and Values
  - Softmax score
- Self-attention
  - Allows inputs to interact with each other
  - More attention to specific features
- Multi-headed self-attention
  - Multiple outputs
  - Parallel



LLMs and Chatbots

- General Purpose Transformer
  - GPT – used to make LLMs
  - Large dataset; large models
- Able to adapt to multiple tasks
- Can do “General-Purpose” AI
- Provide answers to real-world questions



Latest Inventions

- Llama 3 (Facebook – HuggingFace)
- Gemini (Google – Collab)
- GPT 4 (OpenAI – GPT 3.5 on GitHub among other tools)

Hands-on Activity