Introduction to Machine Learning for Everyone

Objectives

After watching this video, you will be able to:

- Define the term machine learning
- Explain how machine learning works
- Describe use cases of machine learning
- Differentiate between AI, machine, and deep learning
- List the different categories and branches of machine learning

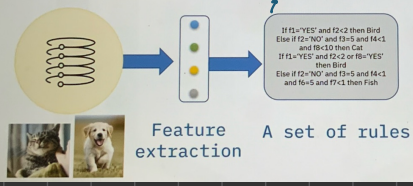
What is machine learning?

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed."

Arthur Samuel

How does machine learning work?

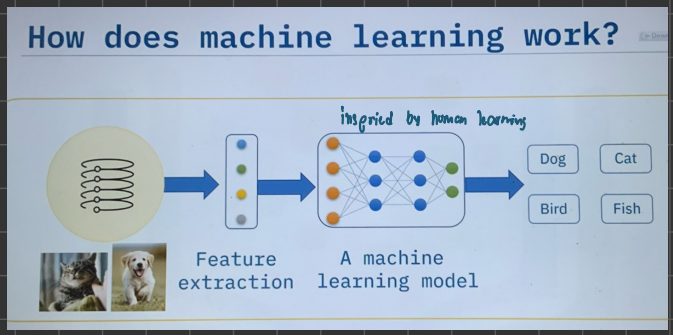
This method was shifted. The rules were numerous not generalized enough to detect out of sample cases



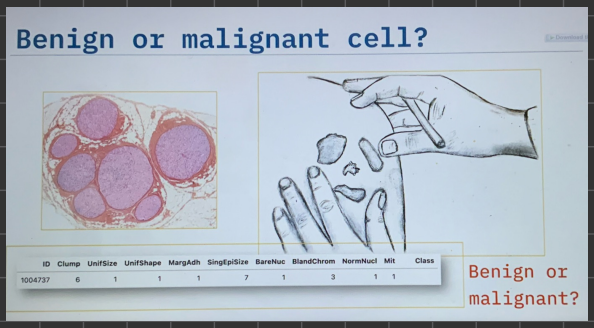
Feature extraction

A set of rules

How does machine learning work?



Benign or malignant cell?



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1004737	6	1	1	1	7	1	3	1	1	

Benign or malignant?

Machine learning helps to predict

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1002348	5	1	1	1	2	1	3	1	1	Benign
1013267	5	4	4	5	2	10	3	2	1	Benign
1023484	3	10	1	1	7	2	3	1	1	Malignant
1018098	5	2	8	1	3	4	3	7	1	Benign
1034349	6	1	1	3	2	1	3	1	1	Benign
1028561	5	4	10	3	2	10	3	1	1	Malignant
1033738	4	2	1	8	2	10	2	1	5	Benign
1023843	5	2	1	1	2	1	1	2	1	Malignant

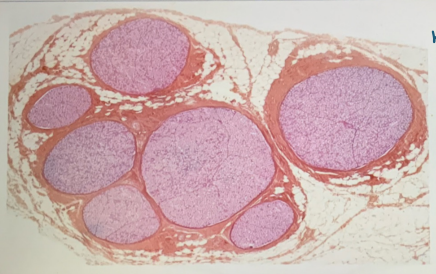
Modeling

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1004737	6	1	1	1	7	1	3	1	1	Benign

Prediction

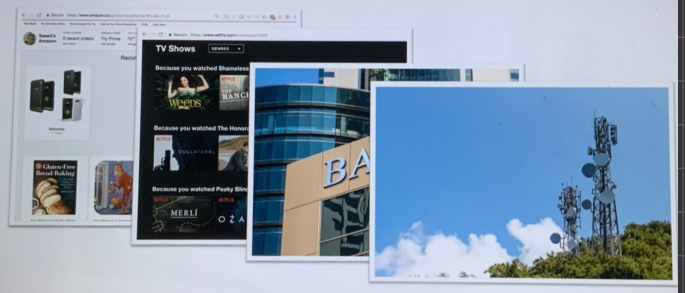
Accuracy = 89%

Machine learning helps to predict



help doctor to classify faster

Examples of machine learning



AI, machine, and deep learning

tries to make computers intelligent enough to mimic humans' cognitive functions

AI components:

- Computer vision
- Language processing
- Creativity
- Summarization

branch of AI, cover statistical part of AI. It teaches com. by many examples, their using that experience to solve the problem

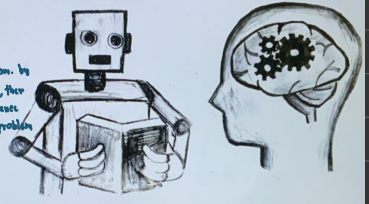
Machine learning:

- Classification
- Clustering
- Neural network

learn and make intelligent decisions independently.

Revolution in ML:

- Deep learning → deeper level of automation



Machine learning categories

uses labeled data to train

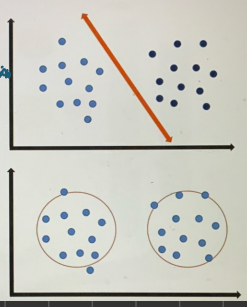
Supervised learning

- Regression: Predicting continuous values (Ex: estimating the CO_2 emission from a car's engine)
- Classification: Predicting the category of an observation (cat vs dog)

to uses unlabeled data and algorithms to detect patterns in the data

Unsupervised learning

- Clustering: Finding structure in the data

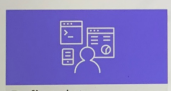


Branches of machine learning

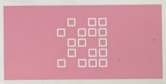
- Deep learning
- Natural language processing
- Computer vision
- Reinforcement learning

Role of Data Engineering in Machine Learning

What you will learn



Define data engineering in the domain of machine learning



List the various responsibilities of data engineers in machine learning

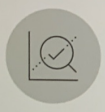
What is data engineering?

Plays a crucial role in machine learning

Responsible for designing, constructing, and overseeing the infrastructure and architecture

Includes gathering, storing, processing, and managing data

What is data engineering?



Allows for efficient access, analysis, and utilization

Ensures the data used for training and inference is dependable and accurate

Data collection

Data engineer's responsibility in ML projects:

Ensuring comprehensive, precise, and relevant data availability

Acquiring data from various sources

Data collection process

Step 1: Define the data needs

- Establish data requirements
- Specify the types of data required
- Example: Structured or unstructured data
- Specify the data sources
- Example: Internal or external sources

Step 2: Identify data sources

- Examples: databases, APIs, web scraping, or manual data entry
- Evaluate the data's quality, relevance, reliability, and accessibility
- Evaluate the cost of data

Data collection process

Step 3: Collect data

- Collect data from identified sources
- Use database query, web scraping, or manual data entry
- Ensure consistency and structure in collected data
- Facilitate analysis and decision-making

Data cleaning and preprocessing

Raw data undergo cleaning and preprocessing to:

- Ensure optimal performance of ML models
- Address errors and abnormalities

Data engineers:

- Prepare data for ML model training
- Ensure data accuracy, completeness, consistency, and analysis readiness

Data cleaning and preprocessing stages

Stage 1

Data exploration:

- Examines data to identify anomalies
- Examples: Missing values, inconsistencies, or outliers

Stage 2

Data cleaning:

- Corrects or removes errors or inconsistencies
- Examples: Duplicates, typos, missing values, and outliers

Stage 3

Data transformation:

- Converts data into a format suitable for analysis
- Examples: Normalization, scaling, and feature selection

Data cleaning and preprocessing stages

Stage 4

Data integration:

- Combines data from various sources into a single data set
- Techniques: Merging, joining, or concatenating data

Stage 5

Data formatting:

- Formats data appropriately for analysis
- Examples: Converting text or dates into numerical formats

Data storage and management

Data engineer's responsibility in ML projects:

Designing and implementing data storage solutions

Handling volume and speed of data required

Ensuring data is stored in suitable formats

Data storage and management stages

Stage 1

Data storage:
- Identifies the appropriate type of storage based on characteristics
- Selects and configures storage systems to handle scale and speed

Stage 2

Data organization:
- Organizes data logically and effectively
- Examples: Data schemas, partitioning strategies, and indexing mechanisms

Stage 3

Data security:
- Protects sensitive data from unauthorized access or manipulation
- Examples: Access controls, encryption techniques, and backup strategies

Data storage and management stages

Stage 4

Data retrieval:
- Implements mechanisms for querying and accessing stored data
- Optimizes data retrieval performance
- Examples: Indexing, caching, and data distribution

Stage 5

Data backup and recovery:
- Creates regular data backup processes
- Implements recovery mechanisms to restore data

Data transformation and feature extraction

Data engineer's responsibility in ML projects:

Transform and extract features from the original dataset

Enhance usefulness for analysis

Steps

1 **Feature selection:** Identifies relevant features for ML

2 **Feature scaling:** Scales features to a common scale

3 **Feature engineering:** Apply techniques to derive new features

Steps

4 **Dimensionality reduction:** Reduces the number of features

5 **Encoding categorical variables:** Converts categorical features into numerical representations

6 **Data imputation:** Fills in missing values using various techniques

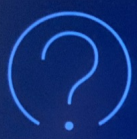
Machine Learning Model Lifecycle

Objectives

After watching this video, you will be able to:

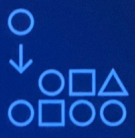
- List the processes within the lifecycle of a machine learning product

Process Lifecycle



Problem Definition

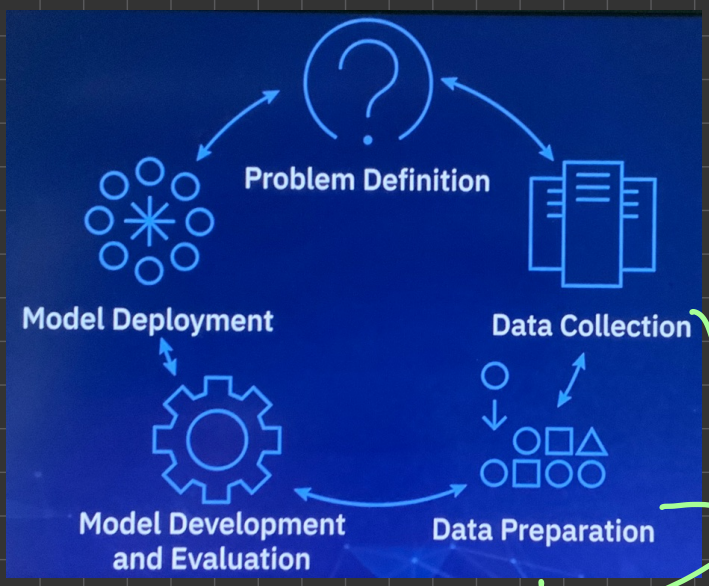
Data Collection



Data Preparation

Model Development and Evaluation

Model Deployment

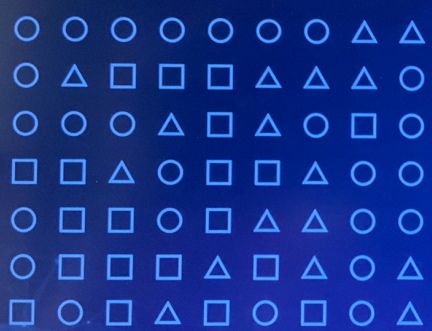


lifecycle is iterative, means we tend to go back and forth between these processes

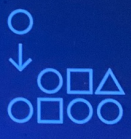
called

ETL

Extract, Transform, and Load



Data Collection



Data Preparation

→ then it can be accessible to the machine learning

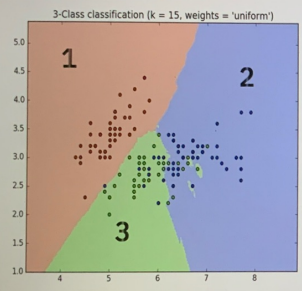
Supervised vs Unsupervised Learning

Objectives

After watching this video, you will be able to:

- Define supervised and unsupervised learning
- List examples of supervised and unsupervised machine learning use cases
- Define clustering
- Differentiate between supervised and unsupervised learning

What is supervised learning?



We "teach the model." Then, with that knowledge, it can predict unknown or future instances.

Teaching the model with labeled data

recorded components from patients historical data

labeled datasets

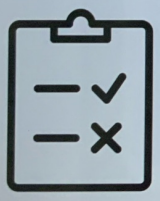
ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	1	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

most used data type that is numerical

These are attributes : also referred to as an observation

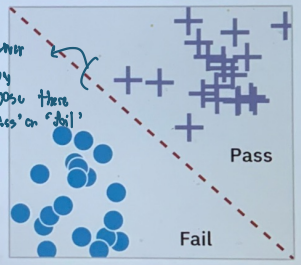
What is classification?

Classification is the process of predicting discrete class labels or categories



"Will I pass or fail my biology exam?"

model can only answer to one category by choosing there 'pass' or 'fail'



What is regression?

relationship between a dependent and an independent variable

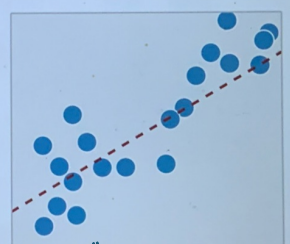
Regression is the process of predicting continuous values

continuous values

not categorical value

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values



by the dataset you can calculate a regression line to predict it

What is unsupervised learning?

- work independently to discover patterns and structures in the data that may not be visible to the human eye
- uses more difficult algorithms because we know little to no info. about the data or outcomes that are to be expected

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering

The model works on its own to discover information.

ALL OF THIS DATA IS UNLABELED

What is clustering?

Clustering is the grouping of data points or objects that are somehow similar by the characteristics of the data.

- Discovering structure
- Summarization
- Anomaly detection

Supervised vs. Unsupervised Learning

Supervised Learning	Unsupervised Learning
• Classification: Classifies labeled data	• Clustering: Finds patterns and groupings from unlabeled data
• Regression: Predicts trends using previously labeled data	• Has fewer evaluation methods than supervised learning
• Has more evaluation methods than unsupervised learning	• Less controlled environment
• Controlled environment	

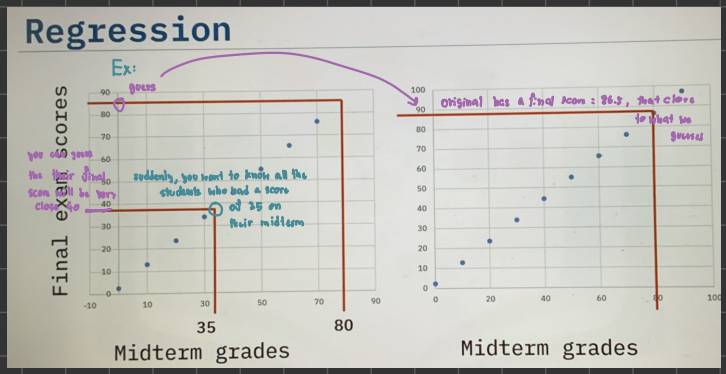
Regression

Objectives

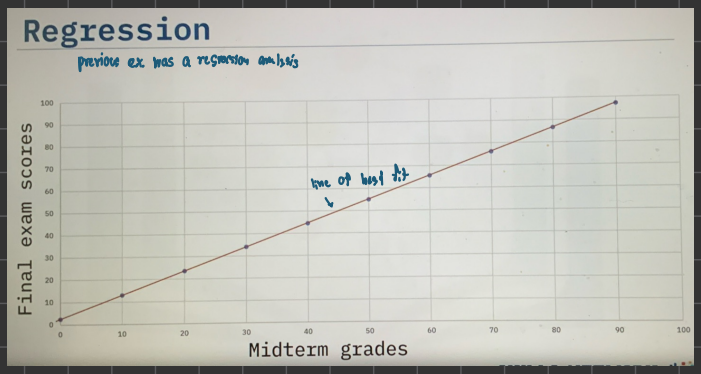
After watching this video, you will be able to:

- Define key concepts in regression
- List some common regression algorithms
- Interpret the results of regression
- Differentiate between classification and regression
- Determine whether classification or regression is suitable for your problem type

Regression



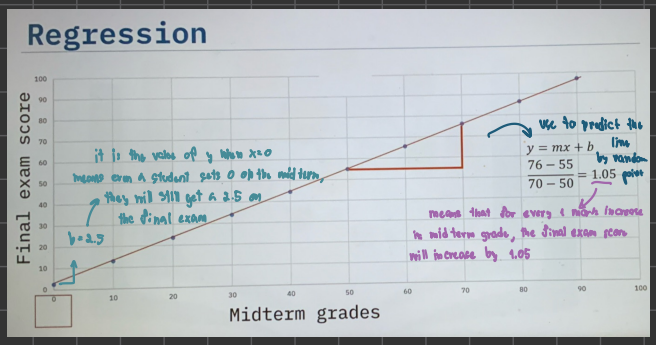
Regression



Regression

- Regression is the relationship between a dependent and independent variable
- The dependent variable is a continuous variable
- Examples include predicting housing prices in Manitoba, predicting scores on a final exam, predicting the weather, and so on

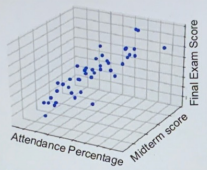
Regression



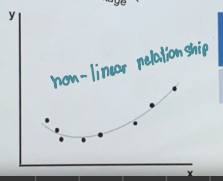
Regression

Input variables

- Midterm score
- Attendance percentage



to account for the nonlinear relationship



Polynomial regression

$$y = B_0 + B_1X_1 + B_2X_2^2$$

Regression

Regularized regression

- Regularized regression adds a penalty that constrains the coefficients to zero

Ridge: Shrinks the coefficients by the same factor but doesn't eliminate any of the coefficients

Lasso: Shrinks the data values toward the mean and will make the model sparser

ElasticNet: Combination of Ridge and Lasso

Regression algorithms

Random forest

Random forest is a group of decision trees combined into a single model

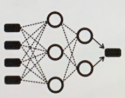
Support vector regression

SVR creates a line or a hyperplane that separates the data into classes

Gradient boosting

Gradient boosting makes predictions by using a group of weak models like decision trees

Neural networks



Neural networks function loosely like the neurons in the human brain to make predictions

Classification vs. regression

Classification	Regression
Values are mapped to a predefined class	Values are mapped to a continuous variable
Values are not ordered	Values are ordered
Performance is measured by accuracy	Performance is measured by the error

Classification

Objectives

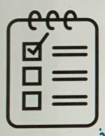
After watching this video, you will be able to:

- Define key concepts of classification
- List the types of classification
- List some common classification algorithms
- Evaluate the accuracy and results of a classification model
- Differentiate between classification and regression
- Determine whether classification or regression is suitable for your problem type

Classification and regression

Classification and regression are two of the prediction problems in machine learning and data science

Classification



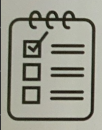
"Will I pass or fail my biology exam?" (just / can't do both)

Regression

"What will my biology exam score be?"

Classification

It is the process of predicting class based on some given inputs



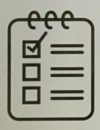
"Will I pass or fail my biology exam?"

Input Variables (independent variables - can be one variable)

- Average score on previous biology tests
- Percent of classes attended
- Number of hours studied

Classification

It is the process of predicting class based on some given inputs



"Will I pass or fail my biology exam?"

helps to answer the question

Input Variables

- Average score on previous biology tests
- Percent of classes attended
- Number of hours studied

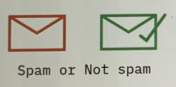
dependent on input variables

Pass

Classification

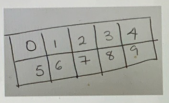
P F

Pass or Fail



Spam or Not spam

Binary classification: Predicts two classes



Multiclass classification: Predicts three or more classes

Classification terminologies

Classifier

Is an algorithm to solve classification problem

Feature

Is an independent variable used as input in the model

Evaluation

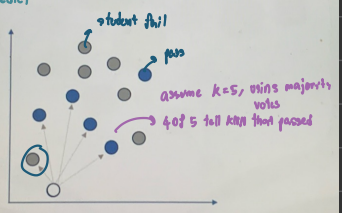
Is the means of validating how well of its performance

Classification algorithms

Lazy learner: Waits to have test data before "training" the data set (doesn't have training phase) (is taking longer to predict)

K-nearest neighbors

KNN: classifies by finding the most common classes in the K-nearest then find the closest match

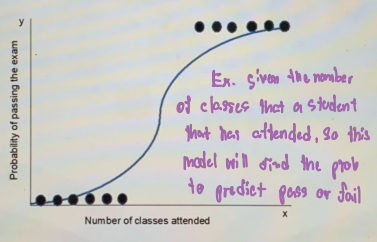


Classification algorithms

Eager learner: Spends a lot of time training and generalizing the model (so it spends less time to predict)

Logistic regression

use probability of a class

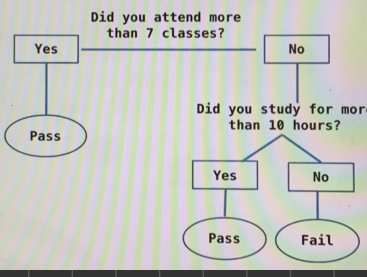


Ex. Given the number of classes that a student that has attended, so this model will find the prob to predict pass or fail

Classification algorithms

Decision trees

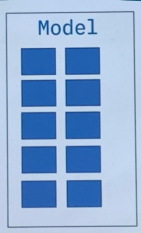
Uses an if-then algorithm



Evaluating Machine Learning Models

Train/test split

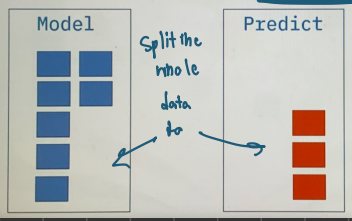
To train a machine learning model, you don't want to feed all the data from the data set



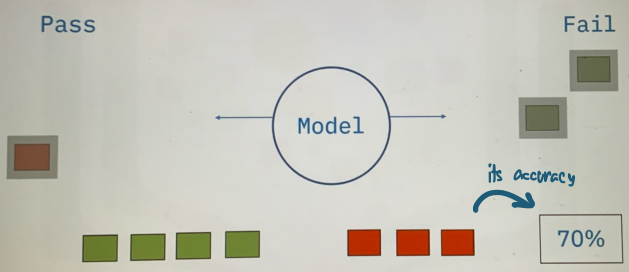
Train/test split

Training set: To teach the model with lots of examples

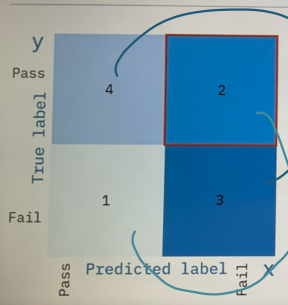
Test set: Will act as the new data to evaluate how well the model performs



Classification accuracy

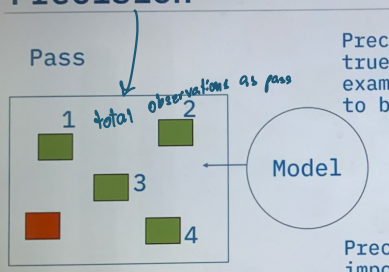


Confusion matrix



- **True positive:** You predicted pass and it is "true"
- **True negative:** You predicted fail and it is "true"
- **False positive:** You predicted pass, but it is "fail" predict incorrect
- **False negative:** You predicted fail, but it is "pass"

Precision



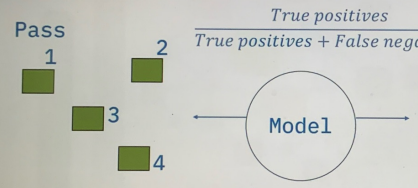
Precision is the fraction of true positives among all the examples that were predicted to be positives

$$\frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

the more FP, the more cost incur

Precision may be more important than accuracy in a movie recommendation engine because it may cost more to promote to a user, more FP mean the user doesn't interest in it

Recall



$$\frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Recall is the fraction of true positives among all the examples that were actually positive

Recall may be more important than accuracy in the medical field

4/6 = 66.7%

real true the 'pass' class

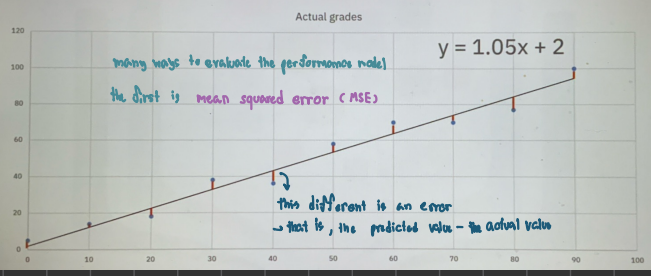
all 'pass' class

important to account for FP, especially when it comes to patient health

F1-score

- In some cases, like a patient misdiagnosis, precision and recall can be equally important
- F1-score is the harmonic mean of precision and recall
- $\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
- F1-score does a great job at balancing both the precision and recall

Evaluating regression models



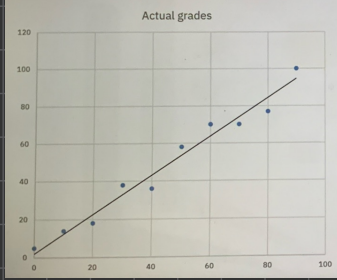
Evaluating regression models

- Mean squared error is the average of squared differences between the prediction and the true output

$$\sum_{i=1}^{n} \frac{(y_{predicted} - y_{actual\ value})^2}{n}$$

- Aim is to minimize MSE, the lower the MSE, the closer you are to the actual predictions
- Root mean squared error is the square root of MSE and has the same unit as your target variable making it easier to interpret
- Mean absolute error is the average of the absolute values of the errors MAE

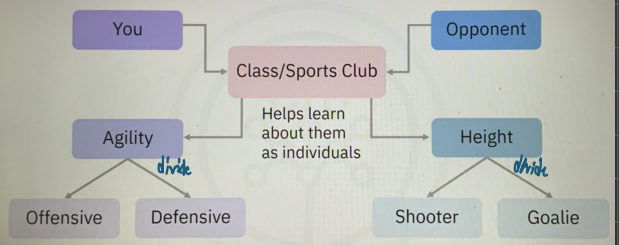
R-squared



- R-squared is the amount of variance in the dependent variable that can be explained by the independent variable
- It is also called the coefficient of determination and measures the goodness of fit of the model
- The values range from 0 to 1

Clustering

What is clustering like? Ex: form sports team in a school



Both approaches provided interesting details about the players

What is clustering?

- Machine learning groups examples
- Clustering groups unlabeled examples

Identifies patterns or connections

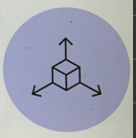
Uses unsupervised machine learning

Identifies groups based on similarity or distance

Application of clustering

Some common clustering applications:

- Customer segmentation
- Image segmentation
- Anomaly detection
- Document clustering
- Recommendation systems



Applications of clustering

Customer segmentation

- Categorizes customers based on purchase history and demographics
- Helps gather data and personalize service offerings

Image segmentation

- Allows image division based on color and content
- Helps with image analysis and computer vision applications

Anomaly detection

- Identifies data points that are unusual or anomalous
- Helps with the detection of fraud and cybersecurity

Application of clustering continued

Document clustering

- Groups documents that are similar in content and keywords
- Assists in retrieving information and grouping news articles

Recommendation systems

- Groups similar items or products based on customer behavior
- Helps create recommendation systems

Clustering types and popular algorithms

Partitioning clustering

- Dataset is partitioned or clustered into k partitions or clusters
- k is a user-specified parameter
- Includes:
  - k-means
  - k-medoids

Clustering types and popular algorithms

Hierarchical Clustering

- Dataset is divided into clusters in a hierarchy
- Each hierarchy level represents a different level of granularity
- Includes:
  - Agglomerative clustering
  - Divisive clustering

Clustering types and popular algorithms

Density-based clustering

- Defines clusters as dense regions of data points
- Separated by lower-density regions
- Includes:
  - DBSCAN
  - OPTICS

K-means: How are they used?

- Used in machine learning and data mining
- Unsupervised learning algorithm
- Divides a dataset into k clusters
- Simple and efficient
- Used to solve clustering problems

K-means: The steps to operate the algorithm

Step 1

Select k initial cluster centroids randomly from the dataset

Step 2

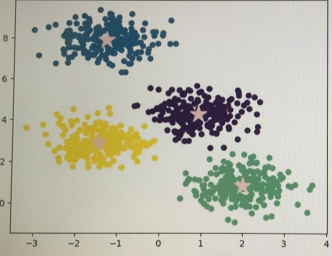
Based on the Euclidean distance, assign each data point to the cluster whose centroid is closest to it

K-means: The steps to operate the algorithm

Step 3

Recalculate the cluster centroids using the mean of the data points assigned to each cluster

K-means: The steps to operate the algorithm



Step 4

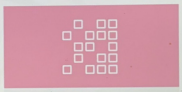
Using the updated centroids, reassign each data point to the cluster whose centroid is closest to it

Repeat steps 3 and 4 until the cluster assignments stay the same

Generative AI Overview and Use Cases

What you will learn

Define Generative AI and describe its significance



Explain different use cases of Generative AI

Artificial Intelligence versus Generative AI

Artificial Intelligence (AI)	Generative AI (GenAI)
Augmented Intelligence that helps experts scale their capabilities • Recognizing speech • Playing game • Making decisions	A type of AI technique that can generate new and unique data • Images • Music • Text • Entire virtual worlds

Generative AI

Uses deep learning techniques

Relies on large datasets to generate new data

Generative AI and LLM

Large Language Model (LLM) is a type of AI that uses deep learning techniques to process and generate natural language.

Develop powerful new LLM algorithms or architectures

Design and incorporate LLM into a larger, more advanced AI system

Significance of Generative AI

Benefits of Generative AI technology include:

- Creativity and innovation
- Cost and time savings
- Personalization
- Scalability
- Robustness
- Exploration of new possibilities

Use cases for Generative AI

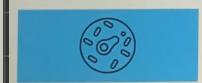
Healthcare and precision medicine

- Identify genetic mutations
- Provide personalized treatment options
- Generate medical images, simulate surgeries, and predict new drug properties
- Help doctors practice procedures and develop treatments

Use cases for Generative AI

Agriculture

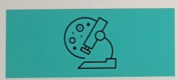
- Optimize crop yields
- Develop new more resistant crop varieties



Biotechnology

- Development of new drugs and therapies by:
  - Identifying potential drug targets
  - Simulating drug interactions
  - Forecasting drug efficacy

Use cases for Generative AI



Forensics

- Analyze DNA evidence
- Identify suspects

Environmental conservation

- Support the protection of endangered species
- Analyze genetic data
- Suggest breeding and conservation strategies

Use cases for Generative AI

Creative

- Produce unique digital art, music, and video content for:
  - Advertising and marketing campaigns
  - Films or video games

Gaming

- Create interactive game worlds
- Generate new levels, characters, and objects in real-time

Use cases for Generative AI

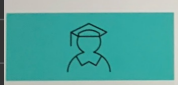
Fashion

- Design and produce virtual try-on experiences
- Recommend personalized fashion choices
- Analyze customer behavior and preferences

Robotics

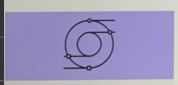
- Design new robot movements
- Adapt them to changing environments
- Enable them to perform complex tasks

Use cases for Generative AI



Education

- Create customized learning materials
- Develop interactive learning environments
- Adapt to the learning styles and learner's pace

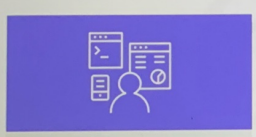


Data Augmentation

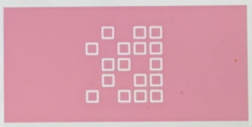
- Produce new training data for machine learning models
- Enhance learning data accuracy and performance

Generative AI Applications

What you will learn



List various applications of Generative AI



Explore the uses for each application

Generative AI applications

Helps create, generate, and simulate new content

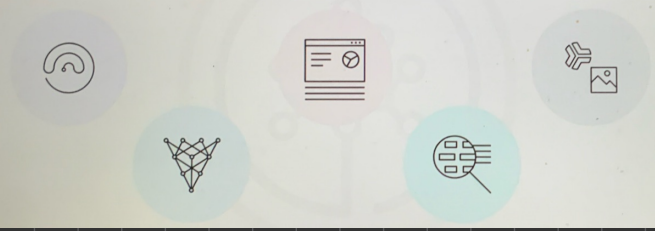
Leverages machine learning and deep learning techniques

Enhances applications capabilities and experiences

Acquires knowledge during training

Applications and tools of Generative AI

Generative AI is used in various fields due to its potential to create new and personalized content



Applications and tools of Generative AI

Generative Pre-trained Transformers (GPT)

LLM developed by OpenAI capable of human-like text

- Iterations – GPT-3.5, GPT-4
- Applications:
  - Chatbots – ChatGPT
  - Automated journalism
  - Creative writing

ChatGPT

Advanced language model by OpenAI

- Applications:
  - Trained on diverse internet text
  - Generates human-like responses
  - Offers various suggestion across various subjects

Applications and tools of Generative AI

Bard

Assists in producing high-quality writing

- Generates text using LaMDA
- Adjusts to the user's style and tone

IBM watsonx

An AI and data platform

- Applications:
  - Watsonx.ai - Model development
  - Watsonx.data - scalable analytics
  - Watsonx.governance - AI workflows
- Helps build, deploy and manage AI apps

Applications and tools of Generative AI

DeepDream

Generates surreal and psychedelic images from real-life images

- Produces one-off and visually stunning images in art and entertainment

StyleGAN

Produces high-quality images of unreal faces

- Applications:
  - Creating realistic video game avatars
  - Simulating human faces for medical research

Applications and tools of Generative AI

AlphaFold

Predicts protein structure

- Transform drug discovery
- Develop more effective disease treatments

Magenta

Creates music and art using generative AI

- Piano duet performed by a human and an AI-generated piano

Applications and tools of Generative AI

PaLM 2

An LLM trained on large datasets

- Applications:
  - Understands nuances
  - Generates coherent texts
  - Translates and answers questions

GitHub Copilot

Assists developers in writing code more efficiently

- Uses deep learning algorithms for:
  - Auto-completing code snippets
  - Functions based on the context of the code

Evolution and Ethics of Generative AI

A rapidly evolving space, expected to grow dramatically in the coming years.

Ethical concerns about Generative AI include:

Potential misuse of AI-generated content

Implications for intellectual property and copyright laws