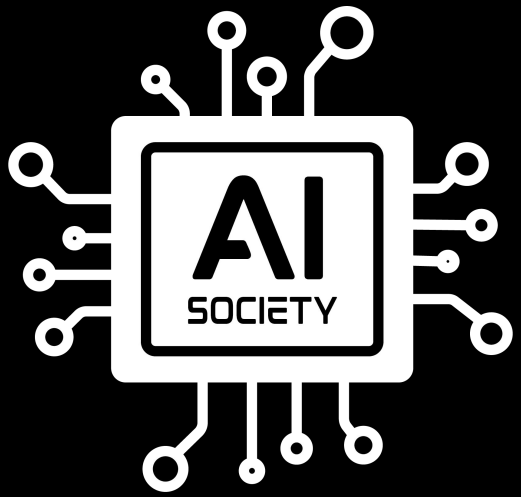


Machine Learning Lab

Week #4 - 2 Oct ‘24

The AI Society at Arizona State University

Objective

By the end of this workshop, you will understand

What/How to:

- Introduction to Machine Learning
- Supervised vs Unsupervised Machine learning
- What is classification
- Classification Algorithms
- Performance Metrics for classification task

Quick Recap : Feature Generation from text data

Quick Recap

Data Cleaning

- Removal of Unnecessary text from the data.
- Highly subjective towards the objective and nature of data.
- To reduce the number of features.

Tokenization

- Converting the words in tokens for latter use as features.
- In-built in tf-idf function in scit-learn library in python.

Feature generation

- Finding the importance of individual words in the context of the document.
- Simplest way to do it is using tf-idf
- Other common algorithm is n-gram model.

TF-IDF

Term Frequency (TF) : Measures how often a term appears in a document.

Inverse Document Frequency (IDF) : Weights the term based on its rarity across all documents.

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t)$$

Helps prioritize important, unique words in documents while downplaying common terms.

Term Frequency

$$\mathrm{TF}(t, d) = \frac{\text{Frequency of } t \text{ in } d}{\text{Total words in } d}$$

Example :

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

Source Images : https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d

Inverse Document Frequency

$$\mathrm{IDF}(t) = \log\left(\frac{\text{Total Documents}}{1 + \text{Number of Documents containing } t}\right)$$

Example :

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

Source Images : https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d

Advantages and Limitations of TF-IDF

Advantages

- Weighs words based on importance, discounts common words.
- Improves Search and Retrieval Relevance.

Limitations

- Context Ignorance
- Synonym Problem
- Cannot Handle Complex Relationships

Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy. –by IBM.

Supervised vs Unsupervised

- Uses labeled training data to teach an algorithm how to map inputs to outputs.
- Train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer.
- Examples include linear regression, logistic regression, and decision trees.

- Uses unlabeled data to teach an algorithm to identify patterns and structure in the data.
- Models work independently to discover patterns and insights without any guidance.
- Examples include k-means clustering, hierarchical clustering, and principal component analysis (PCA)

Machine Learning Life Cycle

Data Gathering

Data Preprocessing

Modeling

Testing and deployment

Classification

Classification is a type of supervised learning where the goal is to predict categorical labels (classes) for input data.

Classification Task

Definition: Classification is a type of supervised learning where the goal is to predict categorical labels (classes) for input data.

Target could be single class or multiple classes.

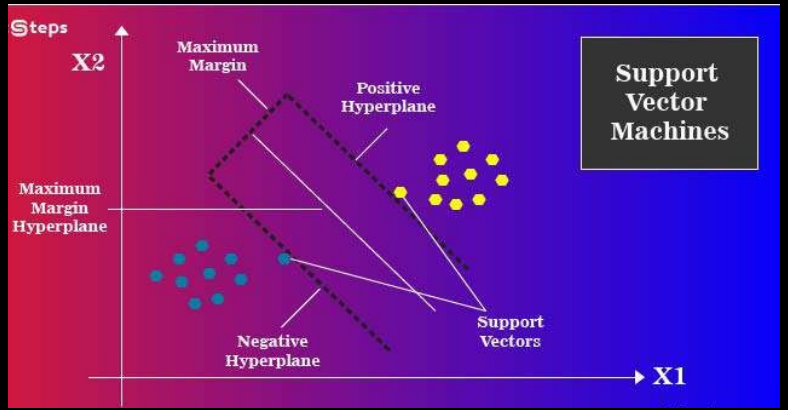
Examples:

- Email Spam Detection (Spam vs. Not Spam)
- Medical Diagnosis (Disease vs. No Disease)
- Image Classification (Cat vs. Dog vs. Rabbit)

Support Vector Machine (SVM)

A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

- Binary or Multiclass classification
- Linear or Nonlinear
- Using kernel tricks or functions



Other Classification Models

Binary

- Logistic Regression
- Support Vector machine (SVM)
- Decision Tree

Multi Class

- Random Forest
- Naive Bayes
- KNN

One-hot Encoding

One-hot encoding is a technique used to convert categorical variables into a numerical format that can be fed into machine learning algorithms.

One-hot Encoding

Why?

- Many machine learning algorithms work better with numerical inputs.
- Prevents the algorithm from assuming an ordinal relationship between categories.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

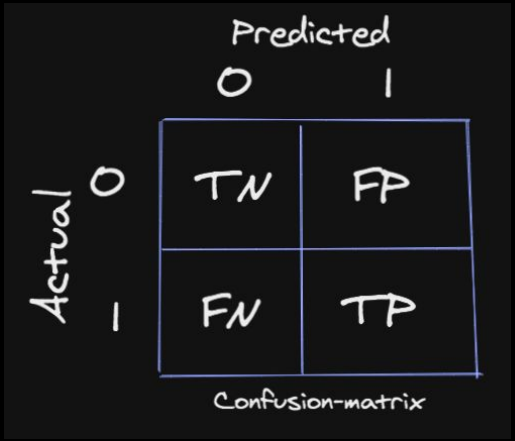
One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Performance Metrics for Classification

Performance Metrics help measure the quality of predictions made by the classification model

Confusion Matrix

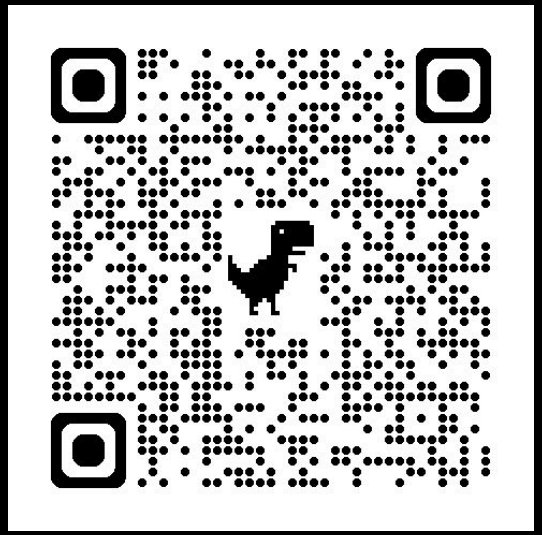


$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Test your knowledge!



Group Picture!