CHAPTER: 14

MACHINE LEARNING ALGORITHM

[1]BHARGAVI GUDDATI
[1]B. Tech, CSE, Machine Learning Algorithm

Ch.Id:-ASU/GRF/EB/RPETHEAT/2022/Ch-14
DOI: https://doi.org/10.52458/9789391842888.2022.eb.grf.asu.ch-14

INTRODUCTION

We have been seeing the term Machine Learning as the buzzword because of it's high amount of data production by applications and increase in the development of efficient algorithms. ML allow computers to self learn from the training data. It is the subset of Artificial Intelligence (AI).

It is been used in every sector and we may already be using a device that utilizes machine learning. Examples of Machine Learning in use are as it can be used in prediction systems, face detection in images as Child growth monitor, medical diagnosis, banking software PayPal and GPay use ML, social media recommendation systems, predict harassment hotspots through ML driven heatmaps, using google earth engine predicting crop production, smart assistants, etc.

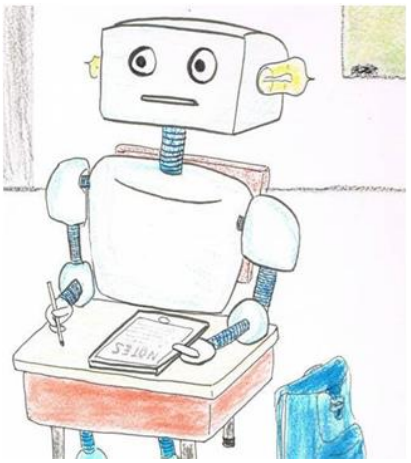
We have two types of machine learning:

Supervised Learning models makes the predictions based on the training data. The set of possible classes are known in advance.

Unsupervised Learning models are not trained with the correct answer so, it has to find the patterns on their own. The set of possible classes are unknown.

We have many classification models some are

1. Statistical-Based Methods
    - Regression
    - Naïve Bayes Classifier
2. Distance-Based Classification
    - K-Nearest Neighbours(KNN)
3. Decision Tree-Based Classification
    - ID3,CART
5. Classification using Machine Learning (SVM)
6. Classification using Neural Network (ANN)



7. Random Forest

Naïve Bayes Classifier:

Here, Naïve is nothing but that each feature is independent of other features but it's not true in real life and Bayes means it depends on the principle of Bayes Theorem for conditional probability and tries to find the conditional probability of target variable given the probabilities of features. It is supervised learning algorithm. It is mostly used in Text Classification as Spam Filtering in Emails, Sentiment Analysis, Recommender Systems etc.

Let us Assume we has a bunch of emails using Naive Bayes classifier we want to classify as *spam or not spam*. We used sklearn CountVectorizer to convert the email text into a matrix of numbers and we use sklearn MultinomialNB classifier to train our model for spam filtering. The model score with this approach was 98.56%.

In Python:

Spam mails Dataset: https://www.kaggle.com/datasets

In [4]:
```
#Import libraries
import pandas as pd
```

In [5]:
```
# The next step is to read the Spam dataset and see their information
df = pd.read_csv("spam.csv")
df.head()
```

Out[5]:

	Category	Message
0	ham	Go until jurong point, crazy.. Available only
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

In [6]: **df.groupby('Category').describe()**

Out[6]:

				Message
	count	**unique**	**top**	**freq**
ham	4825	4516	Sorry, I'll call later	30
spam	747	641	Please call our customer service representativ...	4

In [7]: ***# we are coverting category and message text into numbers because ML understands only numbers.***

df['spam']=df['Category'].apply(lambda x: 1 if x=='spam' else 0)

df.head()

Out[7]:

	Category	Message	spam
0	ham	Go until jurong point, crazy.. Available only ...	0
1	ham	Ok lar... Joking wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	ham	U dun say so early hor... U c already then say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	0

In [8]: ***# imported train test split method from sklearn#Training and Testing data***

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df.Message,df.spam)

In [9]: ***# Still message column is text so, need to convert into numbers.# By using CountVectorizer technique.***

 # From sklearn documentation.

from sklearn.feature_extraction.text import CountVectorizerv = CountVectorizer()

X_train_count = v.fit_transform(X_train.values)
X_train_count.toarray()[:]

Out[9]:

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ..., [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [10]: ***# here we use sklearn MultinomialNB classifier to train our spam filter model***

from sklearn.naive_bayes import MultinomialNBmodel = MultinomialNB() model.fit(X_train_count,y_train)

Out[10]: MultinomialNB()

In [11]: ***# Test***

emails = [

'Hey mohan, can we get together to watch footbal game tomorrow?',

'Upto 20% discount on parking, exclusive offer just for you. Dont miss this reward!']

emails_count = v.transform(emails)model.predict(emails_count)

Out[11]: array([0, 1], dtype=int64)

In [12]:
```
# Accuracy
X_test_count = v.transform(X_test)
model.score(X_test_count, y_test)
```

Out[12]: 0.9856424982053122