Introduction to Machine Learning

WiMLDS Accra and AIMS Mathematics Bootcamp

AIMS Ghana

African Institute for Mathematical Sciences
Ghana

June 2, 2024

Overview

Introduction

AIMS Ghana

Definition

Machine Learning

Machine Learning(ML) is a sub field of Artificial Intelligence(AI). It helps us learn from data to make prediction or decision through:

- Algorithm development
- Statistical development

Benefits of Learning Algorithms

Computers learn from data and continually enhance their performance in specific tasks eg computer vision and natural language.

- **Automation:** ML algorithms helps reduce the need for manual intervention.
- **Insights:** Uncovering valuable patterns and correlations within large datasets.
- **Cost savings:** Optimizing resource allocation, minimizing errors leading to reduced expenses.

Types of Machine Learning

- **Supervised:** Algorithms are trained on labeled datasets..
- **Unsupervised:** Algorithms are trained on unlabeled datasets.
- **Reinforcement Learning:** Algorithms learn to make sequential decisions by interacting with an environment and receiving feedback.

Types of Machine Learning

Most learning algorithms in the industry are supervised and unsupervised. Additionally, there is: **Semi-supervised Learning:** Learning from labeled and unlabeled data, leveraging both to improve learning outcomes.

Applications of Machine Learning

- **Computer vision:** Include face recognition and object detection.
- **Recommendation engines:** Algorithms that provide personalized recommendations based on user preferences.
- **Speech recognition:** Automatic transcription of spoken language into text.

Applications of Machine Learning

- **Signal processing:** Analyzing and extracting information from signals.
- **Natural Language Processing (NLP):** Involves tasks such as sentiment analysis and machine translation.

Common Steps in a Machine Learning Processing

Some important steps before performing any machine learning algorithm.

Data Preprocessing

- Cleaning: Handle missing values.
- Normalization: Scale the features to a similar range to prevent certain features from dominating others.
- Encoding: Convert categorical variables into numerical.

Common Steps in a Machine Learning Processing

Exploration Data Analysis

Data visualization using statistical tools in order to get a best data understanding.

Train-Test Split

Split the dataset into two subsets: a training set (80%) and a testing set (20%).

Common Steps in a Machine Learning Processing

Model Selection

Choose an appropriate machine learning algorithm based on the problem type (classification, regression...) and the characteristics of the dataset.

Model Training

Train the selected model on the training data.

Common Steps in a Machine Learning Processing

Model Evaluation

Using appropriate evaluation metrics evaluate the model's performance.

Validation Model

Validate the final model on the testing set to ensure that it generalizes well to unseen data.

Common Steps in a Machine Learning Processing

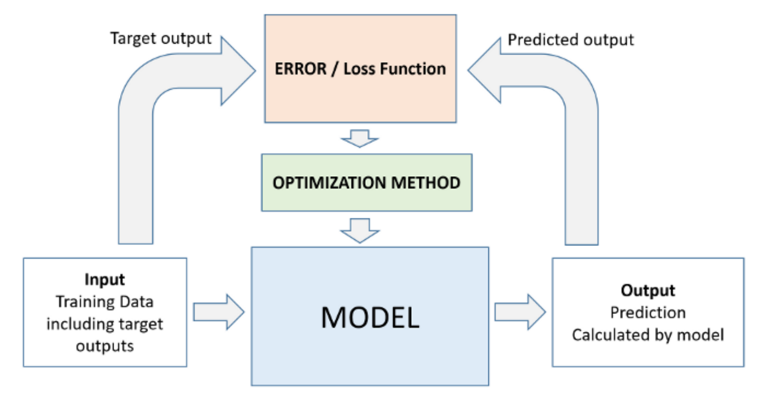

Figure: Common Steps in a Machine Learning Processing

Linear Regression

AIMS Ghana

Definition

The **Linear Regression** is a supervised learning algorithm in machine learning used for predictive analysis.
Let us consider an unknown function f with some noise ϵ:

$$y = f(\mathbf{x}) + \epsilon$$

Where y is a dependent variable called target,
$\mathbf{x}$ are the independent variables called features of y.

Goal

The goal of the linear regression is to find the best-fitting straight line that describes the relationship between the variables, allowing us to make predictions about the target variable based on the values of the features.

Mathematical Aspect

- Step 1
  Let us define the linear regression model.
  Mathematically the relationship between the target and the features is the following:

$$\hat{f}(\mathbf{X}, \theta) = \theta^T \mathbf{X} \tag{1}$$

  Where $\theta^T = (\theta_0\ \theta_1\ \theta_2\ \ldots\ \theta_n)$ is the vector learnable parameter.
  It is estimated from the training data using the gradient descent

Mathematical Aspect

$\mathbf{X}$ is the design matrix that contains the independent variables (features) of the dataset.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm}. \end{bmatrix} \tag{2}$$

Mathematical Aspect

- Step 2
  Let us write the loss function that is the Mean Square Error (MSE).

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x^{(i)}, \theta) - y^{(i)})^2$$

$$\mathcal{L}(\theta) = \frac{1}{n}||\mathbf{X}\theta - Y||_2^2$$

Mathematical Aspect

Note!

Using matrix form offers computational efficiency, simplicity, scalability, and flexibility.

Aim

Our aim is to minimise the loss function as far as possible in order to obtain the optimal parameter θ.

Mathematical Aspect

- Step 3
  Let us compute the gradient of the loss function with respect to each component of θ

$$\partial_\theta \mathcal{L}(\theta) = \frac{2}{n}\sum_{i=1}^{n} x^{(i)}\left(\hat{f}(x^{(i)}, \theta) - y^{(i)}\right)$$

$$\partial_\theta \mathcal{L}(\theta) = \frac{2}{n}\mathbf{X}^T(\mathbf{X}\theta - Y)$$

Mathematical Aspect

- Step 4
  Using the gradient descent let us write the linear regression's algorithm.
  Algorithm: Linear Regression via GD
  Input: labeled dataset $training\_set = (\mathbf{x_i}, y_i)$, learning rate $\eta > 0$
  Initialize: set $\theta := 0$ and counter $k = 0$

Mathematical Aspect

1. **Repeat:**
2. $k := k + 1$
3. $\theta^{(k)} = \theta^{(k-1)} - \eta \partial_\theta \mathcal{L}(\theta)$
4. **Until** convergence

Output: $\theta^{(k)}$ (Optimal parameter)

Model Evaluation

Regression metrics

- Mean Squared Error (MSE)
- Root Mean Squared Error(RMSE)
- Coefficient of determination (R-squared)

Note

Perform the model evaluation using the testing set.

Model Evaluation

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_{true,i} - y_{pred,i})^2$$

- Root Mean Squared Error(RMSE)

$$RMSE = \sqrt{MSE}$$

- Coefficient of determination (R-squared)

$$R^2 = 1 - \frac{SSR}{TSS}$$

Model Evaluation

where

$$SSR = \sum_{i=0}^{n} (y_{true,i} - y_{pred,i})^2$$

$$TSS = \sum_{i=0}^{n} (y_{true,i} - \bar{y})^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_{true,i}$$

Model Evaluation

- SSR: Sum of Squares of Residuals.
- TSS: Total Sum Squares.

Note

- The observable ($y_{true,i}$) is actually the dependent variable (target) from the testing set.
- The prediction ($y_{pred,i}$) comes from the built model using the testing set's features.

Logistic Regression

AIMS Ghana

Definition

What is Logistic Regression?

Logistic regression is a supervised machine learning algorithm that accomplishes classification tasks by predicting the probability of an outcome, event, or observation.

Definition

Logistic regression builds on linear regression and uses one or more independent variables to classify data into discrete classes. It is mostly used for binary classification, however, it can be used to classify three or more classes. It uses Logistic / Sigmoid function to calculate a probability value between 0 and 1, this value is then used for a classification. We fit an ”S” shaped logistic function (as seen in Figure 2 as compared) to fitting a regression line in linear regression.

Classification

Logistic Regression is a classification algorithm-these are used to predict categorical values. Examples include classifying an email as spam or not, whether a picture is of a specific animal or plant species. There are other algorithms used for classification including Support Vector Machines, K-Nearest Neighbours, Random Forest Classification etc..

Classification

Types of Classification

- **Binary Classification:** These are classifications where the categories or labels are of two classes. Eg. Classify whether a customer will buy product or not.
- **Multi-class classification:** Here the classification has more than two classes. Eg. Classifying genre of music.

Classification

Types of Classification

- **Multi-label classification** Here there are two or more class labels and each of these may be predicted for each example or instance. Eg. An image classification where we predict objects in the photo.

The Foundation - Linear Regression

Given the independent input features X and the dependent variable Y.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm}. \end{bmatrix} \qquad Y = \begin{cases} 0 & \text{if class 1 (dog)} \\ 1 & \text{if class 2 (cat)} \end{cases}$$

The Foundation - Linear Regression

We use linear regression to obtain θ^T the vector learnable parameter that is estimated from the training data using the gradient descent and $\hat{f}(\mathbf{X}, \theta)$ in Equation 1 using X the design matrix in Equation (2) as well. Here

$$\hat{f}(\mathbf{X}, \theta) = \theta^T \mathbf{X}$$

All these have been discussed in the linear regression part of the course.

The Sigmoid Function

The sigmoid function maps the predicted continuous values to a value between 0 and 1 and is given by.

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

Substituting our regression equation gives

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

We then choose a certain threshold, e.g. 0.5 to make our classification.

The Sigmoid Function

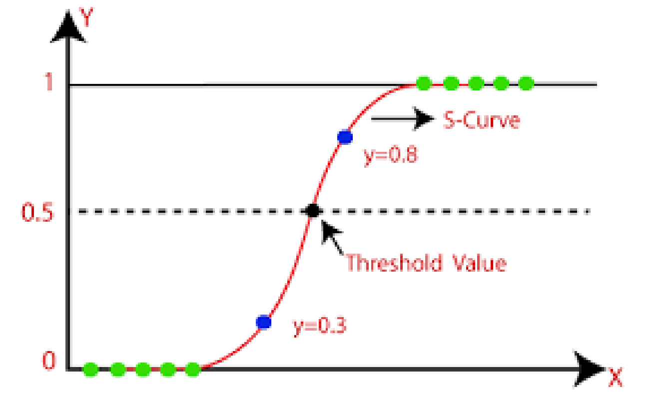

Figure: Graph of Sigmoid Function

Evaluating the Model

Since the Logistic Regression Model, we can use classification metrics like accuracy, precision, recall and F1 Score.
We shall gain an understanding in the above by looking at the confusion matrix.

Confusion Matrix

It summarizes the classification performance by showing:

- **True Positive (TP)**: The number of positive instances that are correctly classified as positive.

Evaluating the Model

Confusion Matrix

- **False Positive (FP)**: The number of negative instances that are wrongly classified as positive. Also known as TYPE I error.
- **True Negative (TN)**: The number of negative instances that are correctly classified as negative.
- **False Negative (FN)**: The number of positive instances that are wrongly classified as negative. Also known as TYPE II error.

Confusion Matrix

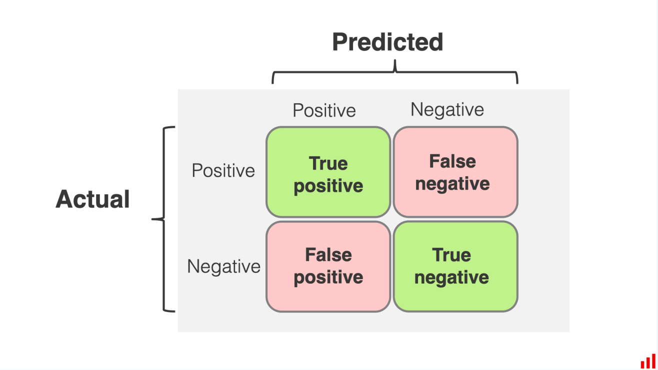

Figure: Confusion matrix for Binary Classification

Evaluation Metrics

We now look at the following logistic regression metrics:

Accuracy: It tells the proportion of correctly classified instances.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total}$$

Precision: Precision shows the accuracy of positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positive}$$

Evaluation Metrics

Recall (Sensitivity or True Positive Rate: Recall measures the proportion of correctly predicted positive instances among all actual positive instances.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 Score: F1 score is the harmonic mean of precision and recall

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

Applications

- Health Applications. It can be used to classify whether a patient has a disease or not based on input features, eg. heart attacks.
- Education Sector. Classification of the acceptance of student based on input features.

Applications

- Identifying spam in emails.
- Finance. Loan application to determine whether a loan applicant should be granted a loan or not.
- Classifications tasks in general.

Conclusion

- Machine Learning is a sub-field of AI used to make predictions based on data.
- It has numerous applications especially in the real world.
- They are different types of ML, and we looked mainly at Supervised models.

Conclusion

- Linear Regression is a supervised learning algorithm for predicting continuous data.
- Logistic Regression is a supervised learning algorithm for classification.

References

References

Issa Karambal, PhD, January 25, 2024.
Machine Learning Fundamentals

GeeksforGeeks, January, 30, 2024.
Logistic Regression in Machine Learning URL

Thank you for your attention

AIMS Ghana

African Institute for Mathematical Sciences
Ghana

June 2, 2024