CODING WEEK
MACHINE LEARNING

INSTRUCTORS: PRANAV GARG, ROSHAN SHAJI

CODING CLUB
IIT GUWAHATI

INTRODUCTION

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Some basic tasks that you can do with machine learning include regression, classification, clustering and ranking.

Watch this video to get a brief overview of what machine learning is; and its types.

REGRESSION

You can watch this playlist to understand the mathematical concepts associated with **linear regression**.

The overall idea of regression is to examine two things:

1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
2) In particular, which variables are significant predictors of the outcome variable, and in what way do they –indicated by the magnitude and sign of the beta estimates– impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.

The simplest form of the regression equation with one dependent and one independent variable is defined by the following formula:

$$\boldsymbol{y = c + bx}$$

where

y = estimated dependent variable score

c = constant

b = regression coefficient

x = score on the independent variable.

Some commonly used regression-based algorithms are:

1. Lasso Regression
2. Elastic Net Regression
3. Kernel Ridge Regression
4. Gradient Boosting Regression
5. XG Boost
6. Light GBM

CLASSIFICATION

You can watch this playlist to understand the mathematical concepts associated with **logistic regression**, one of the classification techniques here.

Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. A binary outcome is one where there are only two possible scenarios—either the event happens (1) or it does not happen (0). Independent variables are those variables or factors which may influence the outcome (or dependent variable). You know you're dealing with binary data when the output or dependent variable is dichotomous or categorical in nature; in other words, if it fits into one of two categories (such as "yes" or "no", "pass" or "fail", and so on).

Some common algorithms for classification problems are:

1. Logistic Regression
2. K - Neighbors Classifier
3. Random Forest
4. Stochastic Gradient Descent
5. Multi-layer Perceptron
6. Naïve Bayes

WRITING THE CODE

In the world of machine learning, the most used language is **Python**, and we would also be using the same. Advanced concepts would not be needed; you just need to know the basics and can have a look at the basic syntax here.

To load, save and analyze the data provided, you would need the help of the **pandas** library. Here is a basic tutorial to pandas, and this is a cheat sheet to some functions of the library.

None of the above-mentioned algorithms need to be hard-coded by you and can be imported directly from the library **scikit-learn**. This tutorial would give you an idea about the library. You can also have a look at the scikit-learn cheat sheet here.

CONCLUSION

Here are a few more resources that might help you in solving the coding week problem.

- Overfit Vs. Underfit
- Bias Vs. Variance
- ML Micro-course