Machine Learning

Minor Project Report

Ayush Choudhary(2K11/CO/032),
Aviral Takkar(2K11/CO/031),
Amanjeet Singh Bhatia(2K11/CO/020,
Ayush Ravi Rai(2K11/CO/033)

Introduction

Machine learning is a branch of artificial intelligence, concerns the construction and study of systems that can learn from data also known as 'experience'. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders.

The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory.

There is a wide variety of machine learning tasks and successful applications. Optical character recognition, in which printed characters are recognized automatically based on previous examples, is a classic example of machine learning.

Objectives and Scope

In our project, , we plan to study various machine learning algorithms, such as classification, regression, etc and apply them to build a basic machine learning application.

We will be creating an adblocker for websites,ads which are static iimages will be considred and not the flash ones.We plan to obtain data about different images/pictures which are ads and nonads from the internetand use that data set to our machine learning aglorithm to tell that a new image not from that data set is a ad or not.We then obatin different images from a random website and apply our machine learning algorithm to predict whether images ad or not and if ad we will remove them or show something else from our side.The work of obtaining the images from a website is done with the help of setting up a proxy sever using python in ubuntu/mac system.

Machine learning algorithms can be organized into a taxonomy based on the desired outcome of the algorithm or the type of input available during training the machine. We would be implementing the following class of algorithms.

Supervised learning algorithms are trained on **labelled** examples, i.e., input where the desired output is known. The supervised learning algorithm attempts to generalise a function or mapping from inputs to outputs which can then be used to speculatively generate an output for previously unseen inputs.

MAIN CONTENT

We have followed some basic steps in creating our adblocker
There are 2 main parts to it

- machine learning part
- proxy sever part/runnnig machine learning algo on website

MACHINE LEARNING PART

In this we have basically applied a machine learning algorithm on a data set of images containing various features to get results which can be used to detect that a ad is ad or not.

Classification : A classification problem is one in which result of the problem is a simple yes/no. This yes/no classification also called binary binomial or binary classification.

In our algorithm we apply a machine learning technique called Logistic Regression which is used in problems of classification as described above. Making an ad blocker is a classification problem as prediction is made about an image being an ad or nonad

Algorithm for logistic regression

we will use 3 things in our algorithm

- hypothesis
- cost function
- gradient descent

Hypothesis

hypothesis is a equation involving variables and constants and depdending upon these a result is found.

$$Y(\text{hypothesis}) = a + a1*x1 + a2*x2 + a3*x3 + \ldots\ldots\ldots\ldots$$

Y is the result found by evaluating above given some values.
x1,x2,x3 . . . . are the variables and in our case they are the features which take different values for different images.
Example for features in our case can be
height/width of the image
aspect ration of image
whether image contains name “ad” in its defination or not

a,a1,a2,a2 . . . . are constants which are initially taken 0 and are be to determined by our machine learning algorithm.

Though hypothesis equation may seem a bit difficult to implement it in computer but is can be done easily

We have 2 things in hypothesis

- variables(features)
- constants(parameters)

we will use matrix to represent them
So,
variable will be a column matrix i.e an matrix having only one column and multiple rows

$$X = \begin{pmatrix} X0 \\ X1 \\ X2 \\ X3 \\ . \\ . \end{pmatrix}$$

now above matrix is only for single image , for multiple image that above matrix will have columns equal to the number of images in data set.

Constant will be again a column matrix

$$THETA = \begin{pmatrix} A \\ A1 \\ A2 \\ A3 \\ . \\ . \end{pmatrix}$$

theta is the name given to above matrix for simplicity and will remain column matrix even there are many many images.

So final equation comes out to be

$$Y(hypothesis) = (THETA)'*X$$

where (theta)' is the transpose of the theta matrix.

Finally Y will itself be a matrix containing single column and rows equal to the number of images in data set and each row tell the value of hypothesis.

Cost Function

cost function can be defined as the summation of the difference between actual value of the result and value found out by the hypothesis given specific parameters that is THETA.

The mathematical formula is

J(THETA) = (1/m)*(

in above
m = number images in our data set
H(x) = value calculated from present parameters
y() = real value given to us

Gradient Descent

In machine learning algorithm our main goal is to find a set of parameters that fit the given data set nicely.So we have to minimize our cost function so that a particular set of parameters can be obtained.

In order to minimize the cost function we have to differentiate it and equate to 0 the result, but before that we give some modification in the formula

As this is logistic regression we have to have our value of hypothesis between 0 and 1.
so to do that we change our hypothesis.
New hypothesis is

$$Y = 1/(1 + e^{(-1*((THETA)'*X))})$$

It can be easily shown that above fucntions value will lie between 0 and 1
Due to change in the hypothesis the cost function also get a bit changed
new cost function

Gradient descent is a technique to minimize the cost function which is our main goal.
In each step of algorithm we find a new set of parameters which makes the result more closer to actual result .These parameters are found out with the help of the gradient descent technique whos formula is given below

THETA = THETA – alpha*differentiation(J(THETA))

where alpha is variation factor can be taken any value.

If we run gradient descent using above formula then a point will come that the function will converge to minimum difference and that is the point at which we get our value of THETA that we will be using in further process of predicting image for ad/nonad.

We have defined three things going to be used in machine learning algorithm.
Algorithm will run for 200-500 iteration. Once we have obatined out THETA we export it to a file names theta.txt for further use.

The code for machine learning part is attaached at the back of the report.

PROXY SERVER PART