AN EMPIRICAL STUDY OF DATA SCIENCE WITH RESPECT TO MATHEMATICAL TOOLS

Dr. Sheeja Ravi
Assistant Professor, Sies Nerul College of Arts Science and Commerce
sheejak@sies.edu.in

ABSTRACT

This paper deals with various mathematical tools used for prediction, forecasting, time series, regression, classification algorithms. The simple regression models are extended to advanced polynomial models that more can be used for top dimensional and multi parameterised data sets. Classification will used for categorical data. The objects are organized around linear classification that any cast-off in mathematical logic and neural networks. The distance and nearest neighbour ideas are accustomed develop pattern and algorithm. Gradient departed concepts facilitate find the stripped-down purpose so we have a tendency to get an optimum solution. Graph theory models are utilized in graph structured data wherever every vertex is acting as proxy and therefore the movement of information is obtained from the importance of the vertex. Most of the likelihood techniques are used in uncertainty, predictions, Bayesian analysis and randomization algorithms in data science.

Keywords: *Gradient, Regression, Classification, Bayesian Analysis.*

1. INTRODUCTION

Data science is a union of quite a lot of fields that use statistics. Data analytics and machine learning to analyze data and extract knowledge.It is about collecting, analyzing, and making decisions. Data science is about finding patterns in data and making predictions about the future through analysis. Companies use them to perform better decisions, predictive analysis, and pattern discoveries.

Data Science blends the application of computer science with mathematical models. Data science initially used only a single source, later it got extended to multiple data sources. Data Science with big data benefits extracts values from different cloud domains. Mathematical models enable us to quickly do calculations and predictions. Regression helps in improving performance. Data Science deals with software like Scikit, Matlab, R Studio, SAS, BigML, Weka, and Python and Mathematical tools like Regression, Classification, Time series, statistics, derivatives.

OBJECTIVE

To study the tools for mathematical which is useful for technologies like data science, machine learning.

Overview of Libraries and mathematical tools used for Data Science Algorithms

1.1 Numpy

It is a mathematical library to create an array. It stores images and has applications in linear algebra. Fourier transformation connect vector from special domain to frequency domains that is the characteristics domain. Image is stored in 2-dimensional array. Numpy is efficient for nearest neighbour search. Matrix can be represented in the form of array using numpy.

1.2 Regression

Forecasting, predictions concepts of regression for forecasting. Machine models searches given data to predictions means given data Eg. Share market predictions given actual data some data is dependent and some independent, consumers mentality is independent variable. Consider two variable data, We use linear regression

$$Y=ax+b \qquad (1)$$

Where a and b are unknown. We assume all the points close to the line Equation (1), so approximately,

$$y_i = ax_i + b \qquad (2)$$

So $\sum_{i=1}^{m} y_i = a\sum_{i=1}^{m} x_i + mb$

Or $\bar{y} = a\bar{x} + b$

Where $\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}, \bar{y} = \frac{\sum_{i=1}^{m} y_i}{m}$

Hence, $(\bar{x}, \bar{y}\ )$ lie on the line.

So, $\bar{y} = a\bar{x} + b$

Or $(y - \bar{y}) = a(x - \bar{x})$ (3)

Hence for each set of data we have

$(y_i - \bar{y}) = a(x_i - \bar{x})$ (4)

We define the vectors

X= $(x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x} \dots x_m - \bar{x})$

Y=$(y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y} \dots y_m - \bar{y})$

Using equation(4) for each component

Y= aX

So, a=$\frac{X.Y}{x.x} = \frac{X.Y}{\|x\|^2}$, b=$\frac{X.Y}{y.y}$=$\frac{X.Y}{\|y\|^2}$ (5)

$\bar{y} = a\bar{x} + b$ (6)

So, b= $\bar{y} -\ a\bar{x}$

From equation (5) and (6) we find equation (1) which can be used to predict y for x.

When it is n dimensional linear regression, then the vectors are reduced to matrix form which can be used for prediction (Cootes, T.F & Taylor, C.J.,2004).

1.3 Classification

Scikit is a versatile tool for science and engineering. If one or more datasets are represented as two dimensional array, then we can make use of scikit-learning information skills. They can be understood as a list of multi-dimensional observations (Duda, R.O & Hart, P.E. ,1973). Open cv help to run classification, segmentation and detection. libraries built model. If we give new object it will segregate to predefined classes. New sample will be matched with the class.

1.4 Linear Algebra

Principle Component Analysis (PCA) is a dimension reduction technique which give most dominant feature from the data. An original principal component is linear combination of original variables. It is extracted in such a way that the first principal component explains maximum variance dataset. Second explains the remaining variances in the dataset and is uncorrelated to first principal component. Third tries to explain the component that is not explained by first two. Hence each additional dimension we add to this technique it takes less variance in the model.

Here we use normalization techniques. We normalize the data in [0,1] by using the equation given below:

$$z = \frac{x - \mu}{\sigma}$$

To understand how the variables are varying from mean with respect to each other we compute covariance matrix. That is we see if there is any relationship between variables. To find the correlation between the data set we find covariance matrix. This covariance matrix is symmetric.

Further from the covariance matrix we find the eigen values and if any eigen values have smaller magnitude, we neglect them.

2. APPLICATIONS

2.1 Edge Detection

Edge is the set of linked pixels that establish a boundary among two disjoint regions. The three varieties of edges are horizontal, vertical and diagonal edges. To preserve the structural properties and reduce the amount of data in an image edge detection is used. Edge detection operators are of two kinds Gradient and Gaussian (Hannah M. J.,1980).

2.1.1 Gradient

The gradient of any picture is attained by convoluting or twisting a filter over the image. The gradient in both x and y direction or axis is evaluated to find the orientation of the edge and find the resultant to find the edge (Shivansh Kaushal, 2022).

To detect filters we apply horizontal and vertical edge detecting filters also called as Robert filters. Here we use 2x2 filter and the gradient is approximated through discrete differentiation which is computed by summing the squares of the difference between diagonally adjacent pixels.

+1	0
0	-1

Gx

0	+1
-1	0

Gy

To find the image gradient using filters, take an 4x4 image

50	45	4	12
18	0	10	6
5	32	50	2
13	20	16	0

Now we will take both Gx and Gy filters and convolute them over the image

The gradient in x direction is

50 *1	45*0	4	12
18*0	0*-1	10	6
5	32	50	2
13	20	16	0

So, Gx = 50*1+45*0+18*0 - 0*1

= 50

The gradient in y direction

50 *0	45*1	4	12
18*-1	0*0	10	6
5	32	50	2
13	20	16	0

So, Gy= 50*0+45*1+18*-1+0*0

= 45-18

=27

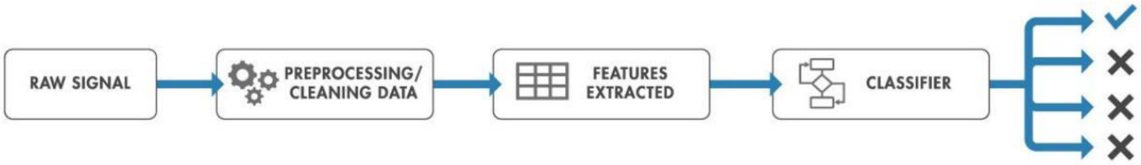
Gradient Magnitude = $\sqrt{50^2 + 27^2}$ = 42.08

Gradient orientation = tan $^{-1}$ (Gy/Gx) = 28.36

The main application of image gradient is in edge detection. This gradient helps in finding the boundary of an object. We can compute the gradient, its magnitude, and the direction of the gradient manually as shown above. Filters were used to find the magnitude and direction.

2.2 FEATURE DETECTION

It stores the raw data in numerical features, in the processed form while preserving in original data form. Create histogram for local gradient direction computed at selected scale. Image correlation can be calculated. Image contains some mathematical transformation of the template image (Moravec H. P. ,1977). Any two images with same patches are reduced to vector form and under Bayesian optimization techniques, the optimization of the image is obtained. Non linear least square techniques are applied to compute homograph. Machine learning algorithms are used for image classification. Machine learning helps in feature detection in digital image. The below figure shows the process



When a raw data is supplied directly it will give a poor output, hence time frequency transformation or Fourier transformation is applied to get better output (Medioni G.& Nevatia R.,1984).

3. RESULTS

Different mathematical tools plays an essential role in various technologies like data related analytics, Machine learning, Artificial intelligence, Deep learning etc. The algorithm built by these advanced technologies have mathematical work behind it. Mathematics helps to built logic behind every algorithm built. Probability concepts is used in decision making and finding solutions to complex problems while performing algorithms. Better Skills in mathematics plays an important role in finding solution to problems and give better insights about the problem. Machine learning is data driven. Mathematics helps to extract different hidden pattern from these data’s numerical calculation gives optimum solutions in algorithms (Chen Y.-W. & Jain L.C.,2014).

4. CONCLUSION

Mathematical tool is very important for data science. Data science cannot work without mathematical concepts and its tool. Software’s, which can use data and calculation make its easy. We have seen various mathematical tool used in different data related work, data science, image processing, feature detection, edge detection. These models give more accurate results compared to direct result and this further provides optimum solution to any field. There are lot more applications of mathematical models like neural network, fuzzy etc, but in this paper some of the models are discussed. In higher learning in the fields of machine and Deep learning, these concepts have a great importance.

REFERENCES

[1] Chen Y.-W. & Jain L.C. (eds.) (2014), “Subspace Methods for Pattern Recognition in Intelligent Environment, Studies in Computational Intelligence” 552, DOI: 10.1007/978-3-642-54851-2_1, Springer-Verlag Berlin Heidelberg.

[2] Cootes, T.F & Taylor, C.J. (2004): “Statistical Models of Appearance for Computer Vision”. Technical Report.

[3] Duda, R.O & Hart, P.E. (1973): “Pattern Classification and Scene Analysis”. Wiley, New York.

[4] Hannah M. J. (1980), "Bootstrap stereo," in Proc. Image Understanding Workshop, (College Park, Maryland), pp. 201-208.

[5] Medioni G.& Nevatia R. (1984), "Matching using linear features," IEEE Trans. Pattern Anal. Machine Intell., Vol. 6, pp. 675-685.

[6] Moravec H. P. (1977), "Towards automatic visual obstacle avoidance," in Proc. 5th Int. Joint Conf. Artificial Intell., (Cambridge, MA), p. 584.

[7] Shivansh Kaushal (2022), "Beginner's Guide To Image Gradient".