Midsem Report: Machine Learning

Summer of Science 2022

Ashwin Abraham

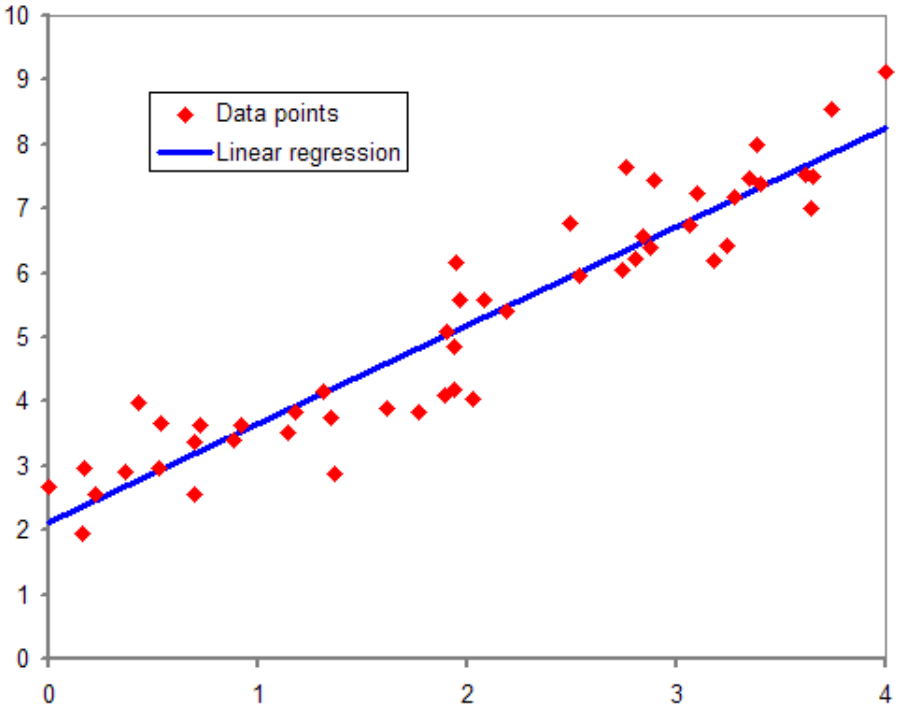
19th June, 2022

Contents

Chapter 1

Linear and Logistic Regression

Regression is the statistical process to find the relationship between a dependent variable and an independent variable that minimizes some cost function, given a set of datapoints.



The above graph shows an example of Linear Regression with one dependent and one independent variable, where a set of datapoints is given,

and we try to find the line that best represents the relationship between the dependent variable and the independent variable.

The notion of what we mean by "best representation" is formalized by the cost function. This is a function of the dataset given to us and the approximate relationship we use to model the relationship between the dependent and independent variables. The best representation of this relation is the one that has the least cost (minimizes the value of the cost function).

1.1 Linear Regression

In linear regression, we impose the constraint that approximate relationship between the dependent and independent variables should be a linear function. The cost function usually chosen for regression problems is the sum of squares function (leading the method itself to be called the method of least squares).

If there are N datapoints of the form (x_i, y_i) and the relation function is f, then the cost of f is then defined as:

$$C(f) = \sum_{i=1}^{n} ||y_i - f(x_i)||^2 \tag{1.1}$$

Here, if there are more than 1 dependent/independent variable, x and y become vectors.

Assuming f is linear and there are only 1 dependent and 1 independent variable, we get $f(x) = Ax + B$. Now we must find constants A and B such that $C(f) = C(A, B)$ is minimized. Now,

$$C(A, B) = \sum_{i=1}^{n} (y_i - Ax_i - B)^2 \tag{1.2}$$

Imposing the conditions that $\frac{\partial C}{\partial A} = \frac{\partial C}{\partial B} = 0$, we get

$$\sum_{i=1}^{n} (Ax_i + B - y_i)x_i = 0 \tag{1.3}$$

and

$$\sum_{i=1}^{n} (Ax_i + B - y_i) = 0 \tag{1.4}$$

Simplifying, we get a system of Linear equations in A and B:

$$A(\sum_{i=1}^{n} x_i^2) + B(\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} x_i y_i \tag{1.5}$$

$$A(\sum_{i=1}^{n} x_i) + Bn = \sum_{i=1}^{n} y_i \tag{1.6}$$

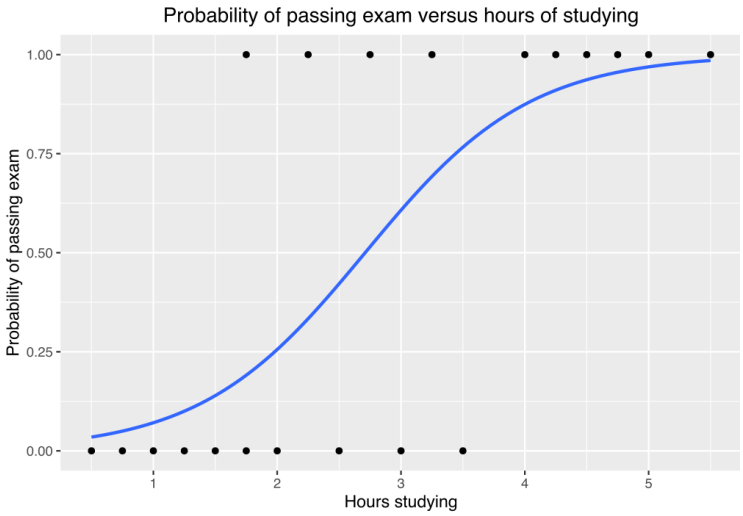
Note, that since in the RMS-AM inequality $(n(\sum_{i=1}^{n} x_i^2) \geq (\sum_{i=1}^{n} x_i)^2)$, equality occurs only when all x_i are equal (which can never happen with a proper data set), this equation will always have a unique solution, given by:

$$A = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{1.7}$$

$$B = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i y_i)(\sum x_i)}{n(\sum x_i^2) - (\sum x_i)^2} \tag{1.8}$$

1.2 Logistic Regression

In a logistic regression on the other hand, the dependent variable is constrained to be only either 0 or 1. In this case, the best approximation to the relation between the dependent and independent variables will also be the Probability Function of the dependent variable (this is a function of the independent variables that gives the probability that the dependent variable will be 1 for a particular choice of independent variables).



The above graph shows an example of a Probability function obtained by Logistic Regression.

The cost function for Logistic Regression is given by:

$$C(p) = -\sum_{i=1}^{n} [y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))] \tag{1.9}$$

and we model the relation between the dependent and independent variables by a logistic function, given by:

$$p(x) = \frac{1}{1 + \exp(-(Ax + B))} \tag{1.10}$$

Now, expressing C in terms of A and B and then imposing the condition that $\frac{\partial C}{\partial A} = \frac{\partial C}{\partial B} = 0$, we get the values of A and B. Note that the equations involved may not always have exact solutions and may have to be solved numerically.

1.3 Use of Regression in Machine Learning

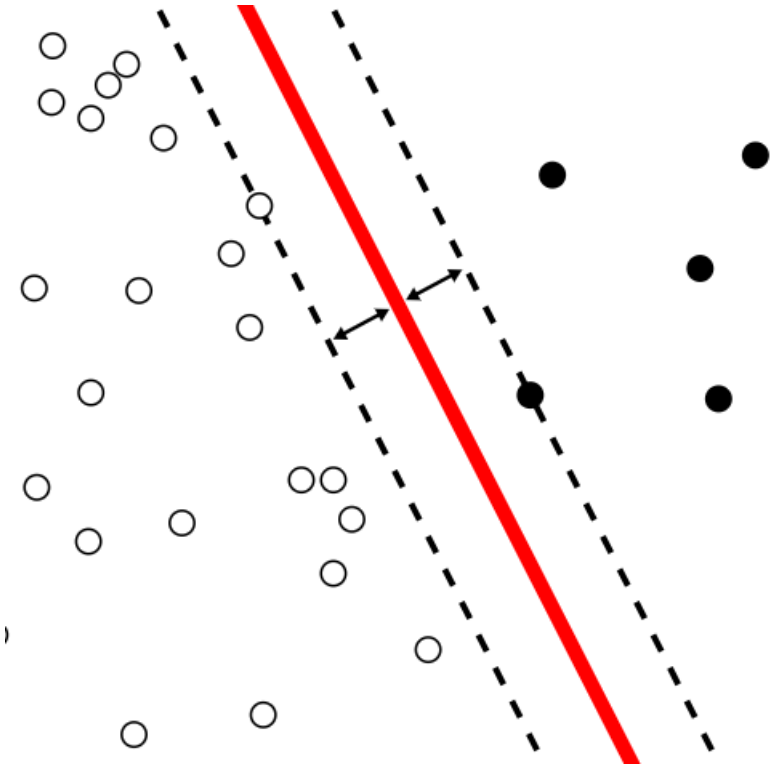
A system implementing Machine Learning techniques must be trained with a large amount of data and it must use the data it has been trained with in order to find ways of handling data that it has not come across before. This can be done effectively using regression techniques. The Machine can use regression analysis on the training data to get a relation between the data and the desired outcome. If the outcome is a binary decision, Logistic regression may be used, and if it is a continuous outcome, Linear regression may be used. The relation obtained is applied on the new data to get the outcome required.

Chapter 2

Support Vector Machines

In Machine Learning, we often need to classify data into multiple groups. If the data can be represented as an n-dimensional vector, then the dataset becomes a set of points in $\mathbb{R}^n$. The easiest way to classify data into different groups is to find a n-dimensional hyperplane dividing the $\mathbb{R}^n$ into two parts, one containing one set of points and another containing the remaining points.

To find the hyperplane, we use a set of points that have already been preclassified into two groups and try to find a hyperplane that separates the two groups into two halves of $\mathbb{R}^n$. If such a hyperplane exists, the dataset (along with the grouping) is said to be Linearly Separable. If a hyperplane exists, then (assuming that the number of points is finite), then an infinite number of such hyperplanes exist. We therefore, choose the hyperplane with the maximum margin, where the margin is defined as the max(minimum distance to a point on one side, minimum distance to a point on the other), as illustrated in the following figure:



For the training set of points $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ $\cdots$, $\boldsymbol{x_n}$, where $\boldsymbol{x_i} \in \mathbb{R}^m$, consider the function y defined as:

$$y(\boldsymbol{x}) = \begin{cases} 1 & , \boldsymbol{x} \in \text{Group 1} \\ -1 & , \boldsymbol{x} \in \text{Group 2} \end{cases} \tag{2.1}$$

We can write the equation of the separating plane with the maximum margin as $\boldsymbol{w}^T\boldsymbol{x} = b$ where $\boldsymbol{w}$ is an arbitrary non-null vector in $\mathbb{R}^m$ and b is an arbitrary real number.

2.1 Linearly Separable Data

If the dataset is linearly separable, then there exist two parallel margin planes passing through the closest points to the separating plane. The seperating plane will also be parallel to this plane and will be exactly between the two planes. Therefore, we can have appropriate $\boldsymbol{a}$ and b such that the two margin planes are given by $\boldsymbol{a}^T\boldsymbol{x} - b = 1$ for the separating plane passing through the

closest point in Group 1 and $\boldsymbol{a}^T\boldsymbol{x} - b = -1$ for the separating plane passing through the closest point in Group 2.

The distance between the margins here is given by $d = \frac{2}{||\boldsymbol{a}||}$, which is the quantity to be maximized (i.e. $||\boldsymbol{a}||$ must be minimized). Imposing the condition that all the points in group 1 lie on the opposite side of $\boldsymbol{a}^T\boldsymbol{x} - b = 1$ as $\boldsymbol{a}^T\boldsymbol{x} - b$ and similiarly for the points in group 2 and $\boldsymbol{a}^T\boldsymbol{x} - b = -1$.

This condition may be expressed as:

$$y(\boldsymbol{x_i})(\boldsymbol{a}^T\boldsymbol{x_i} - b) \geq 1, \forall i \in \{1 \cdots n\} \tag{2.2}$$

If we know the two closest points to the separating plane (from opposite sides) (say $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$), we can reduce these constrained minimization problem to minimizing $||\boldsymbol{a}||$ based on the constraints:

$$y(\boldsymbol{x})(\boldsymbol{a}^T\boldsymbol{x} - b) = 1, \boldsymbol{x} \in \{\boldsymbol{x_a}, \boldsymbol{x_b}\} \tag{2.3}$$

$\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are known as support vectors (hence the name Support Vector Machine). $\boldsymbol{a}$ and b obtained are clearly dependent only on the Support Vectors.

2.2 Non-Separable Data

For data that is not linearly separable, a separating plane doesn't exist that clearly divides space into regions containing only points of the same group. However, it is still possible for there to exist a plane that roughly divides space into regions containing points mostly of the same types. We can still use this plane to make predictions about points not in our data.

Instead of minimizing $||\boldsymbol{a}||$, we instead minimize the quantity:

$$k||\boldsymbol{a}||^2 + \frac{1}{n}\sum_{i=1}^{n}[H(\boldsymbol{a}^T\boldsymbol{x_i} - b, y(\boldsymbol{x_i}))] \tag{2.4}$$

for parameters k and then choose the most appropriate k.

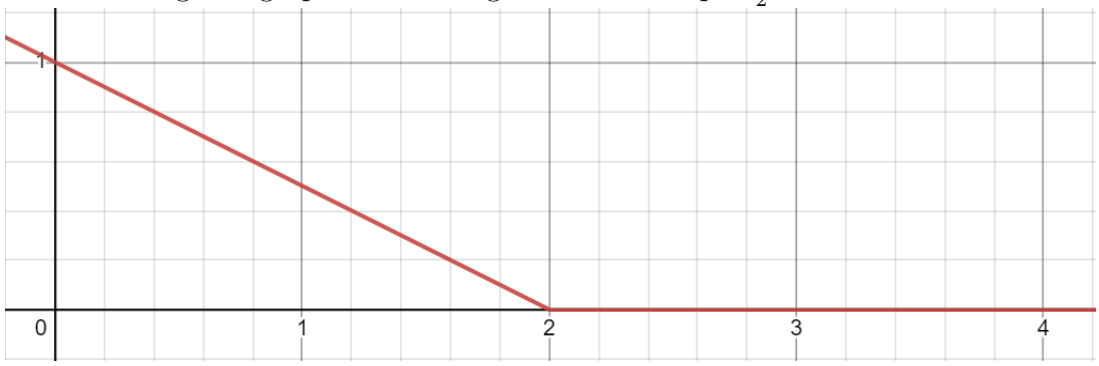
Here H represents the hinge function, defined as:

$$H(p, q) = max(0, 1 - pq) \tag{2.5}$$

Clearly, if $y(\boldsymbol{x_i})(\boldsymbol{a}^T\boldsymbol{x_i} - b) \geq 1$ then $H = 0$, i.e., if the points are on the correct side of the plane, there is no increase in the quantity to be minimized. If

they are on the wrong side, then there is an increase, which is **proportional** to the degree in which the points are in the wrong side.

The following is a graph of the Hinge function for $q = \frac{1}{2}$:



Chapter 3

Updated Plan of Action

Timeline

Dates	Topics to be covered
16th June - 20th June	Principal Component Analysis, K-means clustering and Decision Trees (with Bagging and Boosting)
21st June - 25th June	Neural Networks and Convolutional Neural Networks
27th June - 1st July	(Endsems)
2nd July - 4th July	Recurrent Neural Networks
5th July - 7th July	Autoencoders
8th July - 11th July	Recommender Systems
12th July - 15th July	Final Report

References

1. A basic ML course with minimal maths to get started with
2. A five-course deep learning specialization by Coursera
3. This is an advanced course on Machine Learning by Cornell University with a good amount of focus on maths

4. Tensorflow tutorial page has numerous implemented examples
5. Follow this lecture series for learning Support Vector Machine (SVMs) and other topics which you find to be good in this series