GeoNeurale

MACHINE LEARNING ALGORITHMS APPLICATIONS IN SEISMIC AND PETROPHYSICAL ANALYSIS

Part 1 - LINEAR HYPOTHESES

Applications of machine learning algorithms are increasingly used in seismic and petrophysical interpretation both for analog prediction and for classification.
In “supervised mode”, a training set of input variables (parametrized by some features) and target values is used to model / train a hypothesis function useful to predict future target values.
Each time that we are confronted with an optimization problem we have to go through 3 main phases:

1. Definition of a hypothesis function for the best approximation of the training set
2. Calculation of a cost/objective function (error between the regression approximation and the training set) as difference between each element of the training set ordinate and the corresponding hypothesis ordinate (at same abscissa of the considered element).
This corresponds to half of the sum of square errors normalized for the number of samples in the training set.
3. Calculation of the gradient descent of the objective function to define the parameters that will minimize this function.

If the hypothesis of a linear regression is optimized to forecast an unknowns which belong to the class of the training set with maximum probability we still can have problems in forecasting events which are not strictly linear.
In this case higher order polynomials have to be used to better approximate the optimal hypothesis.
However linear regression is well used in many problems that describe clear linear relationships between parameters. This could refer to many examples like Archie properties forecast in clean sands or pure intercrystalline porosity in carbonate for petrophysics and seismic inversion for flat simple tectonic and stable elastic formation properties.
One log can be predicted from three or more different logs through linear or non-linear regression.
With multiattributes analysis we can forecast the distribution of petrophysical properties in the seismic volume through seismic attributes.
In new applications like PreSDM, tomographic inversion, FWI and velocity analysis (SO-CIG), multivariate regression is often used to flatten the CIG or minimize the difference between the model and other seismic or petrophysical constraints.

To adapt the nomenclature and notation from machine learning theory to geophysical theory the name “training set element” will be translated into “input attribute” for geophysical applications.

The theory of machine learning applications starts from simple linear regression, where the equation of a line is described by its equation. Here **x** will be an input attribute, **M** (angular coefficient) and **q** (intercept). **M** and **q** are the parameters and **y** is the target attribute (Eq. 1).

1. $$\mathbf{y = M\,x + q}$$

These terms in machine learning are described with different notations to underline that this line is introduced as a hypothesis which is a function of a training set attributes $\mathbf{X^{(i)}}$ of **m+1** elements with **0 <i < m** and that the intercept **y** and inclination of this line is a function (parametrized) by coefficients θ **j**. Parametrized means that coefficients $\theta\mathbf{j}$ have the power to progressively change the position of the line during the progress of the optimization process until a final position is reached where this line will be able to forecast future target attributes with the minimum error.

In machine learning notation the regression line equation is written as:

2. $$\underset{\text{LIN}}{h_\theta}(x) = x_o\,\theta_o + x_1\,\theta_1 \qquad (\text{ where } X_o = 1\ )$$

This equation is called “linear hypothesis” used to model and approximate or forecast the process described by the training set **x1** composed by **m** elements:
“linear hypothesis of the line described by the training set $\mathbf{X^{(i)}}$ and parametrized by parameters θ_o and θ_1 ”. In this notation of **x** , $^{(i)}$ is not an exponent but the index of the elements of vector **X** describing the input attributes.

MULTIVARIATE REGRESSION

Input variables that belong to one single class of attributes can be also called “feature”.
Suppose we wanted to predict porosity from resistivity. In this case resistivity is the feature and porosity the target. Another example could be predicting reflected wave amplitude from elastic impedance, then elastic impedance is the single feature and amplitude the target property.
As we know these examples are not realistic, we need more that one feature to predict seismic and petrophysical target attributes. We need to extrapolate the regression operators from the2 to the 3 to the n-dimensional space. For n features the extrapolated form of the linear regression. is eq. 3 .

3. $$\underset{\text{LIN}}{h_\theta}(x) = x_o\,\theta_o + x_1\,\theta_1 + \ldots + x_n\,\theta_n \qquad (X_o = 1)$$

The general equation can be written in vectorial form:

4. $$h_\theta(x) = \theta^T X$$

In multivariate regression **X** is a vector of dimension $\mathbb{R}^{m+1}$ and θ is a vector of dimension $\mathbb{R}^{n+1}$ **.**
m being the number of of input attributes for each attribute class of the training set (feature) and n the number of features.

The standard notation for each sample of the training set is:

5. $\mathbf{X}^{(i)}{}_{j}$

Multiple features means that we have several input attributes available as data for the training set. If we could have for the rough calculation of **Sw** a set of 4 measurement type or 4 features **Rw**,**Rxo**,**Rmf**,**Rt**, one measurement every 1ft , then we would have for 1000 ft , m=1000 samples of each feature. For vectorization we have to add **Xo** therefore the vector dimension for each feature is 1001 → 0<i<1000. The number of feature would be also n = 4+1 (0<j<4), (consider **Xo** and θ_o). Consider another example.

Suppose that for the distribution of Porosity **Phi** as target attribute on the 3D seismic cube we have the following input attributes: **Vp**, **amplitude envelope**, **instantaneous phase**, **average frequency**, **acoustic impedance**.

Suppose that we have to analyze a formation of surface 1000x1000m and the bin dimension is 25x25m.

Then n = 5+1 , m = (1600+1) x fold. By neglecting the fold in case seismic attributes are averaged bin by bin then m= 1600+1.

Thus: X_o =1, θ_o =intercept , 0<i<m , 0<j<n .

OPTIMIZATION PROCESS

As previously discussed, the main problem is solving for θ and this will permit to write the best fit equation for the hypothesis.

For multivariate linear regression geophysical purposes two main approaches are used:

-Objective function minimization.

-Normal equation linear algebra solution.

Objective function minimization is the operator which allow solution where larger number of parameters are present and will be the subject of the following discussion.

As mentioned before, the optimization process of solving for the best regression line which forecasts the optimal results for a specific process is reached with two steps:

1. Calculation of an “Objective” or “Cost Function”
2. Search of the minimum value of the objective function which expresses the minimum difference between training set and hypothesis.

There are many optimization methods for searching the minimum. Here we will consider the gradient descent.

The Obiective function $J(\theta)$ quantifies the difference between the training set and the hypothesis.

6.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right]^2$$

This is a surface on the 3D space θ_o , θ_1 , $J(\theta)$,
However this is a hypersurface in the **n** dimensional space θ_o , .. , θ_n , $J(\theta)$ with (0 <j < n).

The gradient descent method is applied to find a local minimum in the objective function. This can be a variable issue, as there can be many local minima and the final result can depend from my starting point θ_o , .. , θ_n where I start searching the minimum. Physical methods related to a specific process have to be applied to reach appropriate absolute minima.

In FWI this gradient descent method is integrated by reverse time migration of residuals which is analogue to the back propagation procedure of errors in Neural Networks (NN).

The operator for the gradient descent is implemented by contemporary update of directional derivatives of the objective function.
The gradient is by definition the maximal directional derivative and the gradient here is the vector sum of the directional derivatives along each component of the features θ_o to θ_n .

By applying the partial derivative of the function $J(\theta)$ we obtain eq. 7:

7.

$$\frac{d}{d\theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right] x^{(i)}_j$$

Updating θ_j new with gradient descent means find new points in the hypersurface with abscissae θ_o to θ_n with decreasing values of $J(\theta)$.

The partial derivatives have to be multiplied by a factor α called "learn rate" and this factor must be subtracted from the previous value for θ_j .

The choice of α is critical for the convergence of the minimization process. There are minimization procedures where α is automatically updated. However as the partial derivative decreases constantly and tend to zero through the minima also the product for proper choice of α will tend to zero.

By software implementation care must be taken that new θ_j components must be simultaneously updated with the assignment operator (:=).

The update is performed through programming loops eq. 8 :

8.

$$\theta_{j\,\text{NEW}} := \theta_{j\,\text{OLD}} - \alpha \left(\frac{1}{m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right] x^{(i)}_j \right)$$

There are cases where a proper choice of input attributes are not modeled through linear hypotheses. In this case we still can use the operator of the regression process but introduce a polynomial regression setting $x_1, x_2, x_3, \ldots, x_m$ equal to exponential factors.
The regression line can be modeled by positioning of $x_1 = X$, $x_2 = X^2$, $x_3 = X^3$, $x_m = X^m$.
This algorithm is basic for many applications in machine learning and also development platform for non-linear operators, logistic regression and NN.

Fig 1. shows a 3D objective function $J(\theta)$, as function of the parameters θ_o and θ_1 with the directions through the optimal minimum calculated from gradient descent algorithmus which depends from the starting point. A slightly change of the starting point can lead to different local minima. This is why the dynamical parameters of the physical system has to be integrated on the gradient descent calculation to reach the absolute minimum.

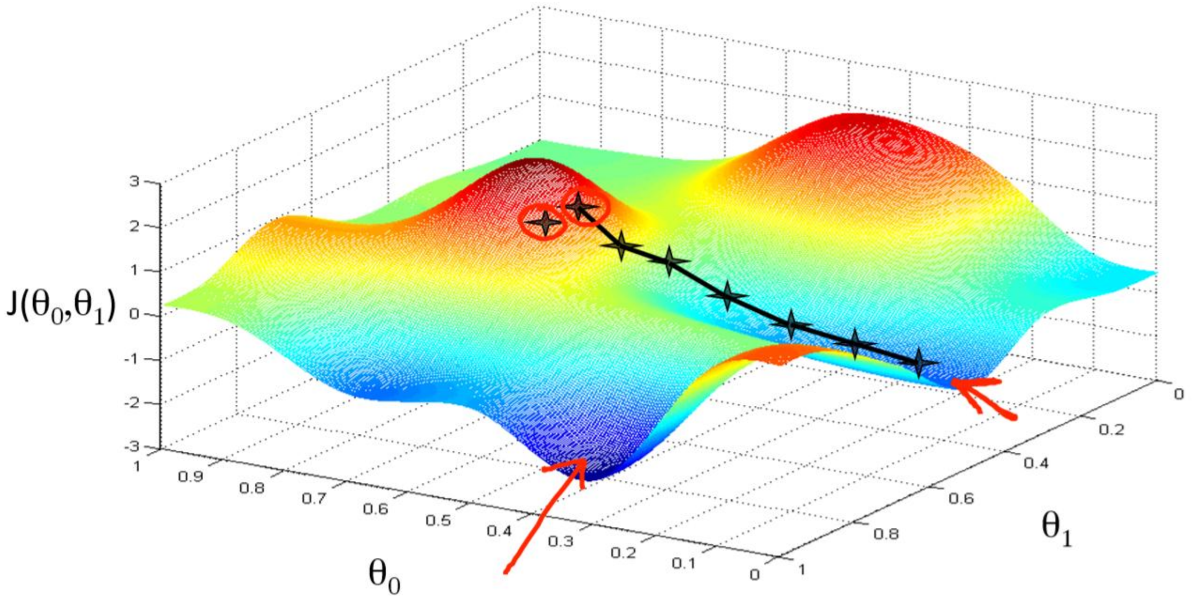

Fig. 1

A. Piasentin
GeoNeurale Research

GeoNeurale
Am Nymphenbad 8
81245 München
T 089 8969 1118
F 089 8969 1117
www.GeoNeurale.com