Machine Learning Review

Francesco Saverio Zuppichini

November 4, 2017

The aim of this report is to summarise the topics threaded in my Machine Learning class at USI.

1 History of Machine Learning

Write something about the history of ML

2 Gradient Descend

The gradient descend is an iterative optimisation algorithm that follows the direction of the negative gradient in order to minimised an objective function. It can be effectively used as Learning Algorithm because it reduces the error function, Equation 3, and adjusts the weights properly. Equation **??** shows the generic update rule.

$$w_{k+1} = w_k - \eta \nabla E(w_k) \tag{1}$$

Where η is the step size, also called **learning rate** in Machine Learning. This parameter influences the behaviour of gradient descent, a small number can lead to local minimum, while a bigger learning rate could "over-shoot" and decreasing the converge ration. Later in this project you will see how a wrong η can strongly change the output of a Neural Network.

For this reasons, numerous improvements have been proposed to avoid local minima and increase its convergence ration, some of them are: Conjugate Gradient and Momentum.

3 Perceptron

4 Definition

The **Perceptron** is binary **linear classifier** algorithm used in **supervised learning**. It can be seen as the most basic form of Neural Network. Equation 2 defines the its output.

$$f(x) \begin{cases} 1 & \text{if w} \cdot x + b \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Given a training set $D = \{(x_1, t_1, ..., (x_n, t_i)\}$, $x_i \in X$ and $t_i \in Y$ denotes the input vector and the target vector respectively. We express $y = f(x)$ as the output of the algorithm, w the weight and b

the bias. At each iteration the error is calculated using the Mean Square Error, defined in equation 3

$$E(w) = \frac{1}{N}\sum_{i=1}^{N}(\underbrace{y(x_i, w_i)}_{\text{predicted}} - \underbrace{t_i}_{\text{actual}})^2) \tag{3}$$

The algorithm uses stochastic **Gradient Descent** in order to update the weight at each iteration using the formula defined in Equation 1.

$$\frac{\partial E}{\partial w_k} = y - t \tag{4}$$

5 Neural Network

A **Neural Network** is a universal function approximation. It is a nested composite functions like $f(g((h...)))$. In its simplest representation, an FeedForward Neural Network, it is composed by a **input layer**, an **hidden layer** and an **output layer**. The size of the hidden layer is usually refers as the **deepth** of the network.

5.1 Forward pass

In order to get the prediction out of our network we need to calculate the compute the activation at each layer l. Equation 5 shows the activation a of layer l for the j-th neuron on that layer.

$$a_j^l = \sigma(\sum_k w_{jk}^l a_k^{l-1} + b_j^l) \tag{5}$$

Where w_{jk}^l is the connection from neuron k in the $l-1$ layer to j, a^{l-1} is the activation of the previous layer and b_j^l is the bias of j-th neuron in the l layer. With this in mind, we can rewrite 6 in a efficient vectorised form

$$a^l = \sigma(W^l a^{l-1} + b^l) \tag{6}$$

5.2 Delta rules

In a Neural Network the weights are iteratively changed in order to decrease the cost function, called E. We want to find out how much they should be updated, in order to do so we need the output error at each layer. Equation **??** defines δ_j^l as the output error of neuron j in layer l

$$\delta_j^l = \frac{\partial E}{\partial z_j^l} \tag{7}$$

Strictly speaking, δ_j^l, is how much the error function changes by changing the weighted input on that layer. Applying the chain rule, Equation 7 becomes:

$$\delta_j^l = \frac{\partial E}{\partial a_j^l}\frac{\partial a_j^l}{\partial z_j^l} \tag{8}$$

By knowing that $a_j^l = \sigma(z_j^l)$, Equation 8 can be expressed as

$$\delta_j^l = \frac{\partial E}{\partial a_j^l}\sigma'(z_j^l) \tag{9}$$

Also, delta at layer l can be expressed by using the next $l+1$-th delta. Equation 10 shows the new rule.

$$\delta^l = (W^{l+1}\delta^{l+1}) * a^l \tag{10}$$

5.3 Back Propagation

The **Back Propagation** algorithm defines an efficient and interactive method to calculate the gradient at each layer. We want to compute $\frac{\partial E}{\partial w^l_{jk}}$. We can applying the delta rule:

$$\frac{\partial E}{\partial w^l_{jk}} = \frac{\partial E}{\partial z^l_j}\frac{\partial z^l_j}{\partial w^l_{jk}} = \frac{\partial E}{\partial a^l_j}\frac{\partial a^l_j}{\partial z^l_j} \tag{11}$$

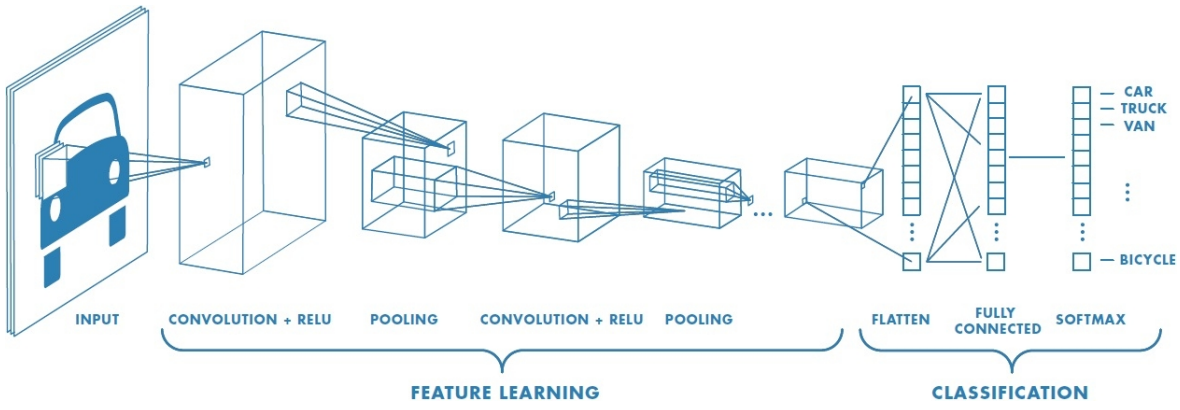
After some calculation, Equation 12 shows how to calculate the gradient for the weight w of the l-th layer for the j-th neuron.

$$\frac{\partial E}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j \tag{12}$$

While δ^l can be calculated used the $l+1$-th layer's information, this is why is called *back* propagation since we used the last layer information in order to, each time, find out the next layer's delta until we reach the input layer.

6 Convolutional Neural Network

Convolutional neural networks are neural network that uses a **convolutional** operation at place in at least one layer. They are used for processing grid-like data, such as time-series, a 1D grid and images, 2D grid. Figure **??** shows a the generic architecture for a CNN.



6.1 Notation

We express an Image as a function $f : Z^2 \to R^c$, where c is the number of channel, in a color image there are three (rgb). A Windows is a subset of the image domain, $W \in Z^2$, that corresponds to a rectangle inside the image.

6.2 Convolutional Layer

As in a hidden layer, a Convolutional layer is formed by n neurons. The difference is that they are not necessarily connect to the activations of the neurons in the previous layer, but only in a particular window. The architecture is composed as followg:

- A input $[w * h * c]$ holds the an image
- A convolutional layer computes the output of its neurons that are connected to some window in the input, each computes the dot product between its weight and the region they are connected to. Each neuron apply a filter, usually of size $3x3$ or $5x5$.
- A rectifier function, RELU $max(0, x)$, is applied.
- A max-pool layer perform a downsampling operation.
- A fully connected layer computes the class scores

7 Recurrent Neural Network

7.1 Definition

A Recurrent Neural Networks can remember past decision by taking as input not only the current input but also the last time state. For this reason it is said that a RNN has **memory**. Figure 1 shows a classic representation. Usually, a RNN is represented unfolded to highlight the time dependencies. Due to its ability to remember it mostly used in text and speech recognition.

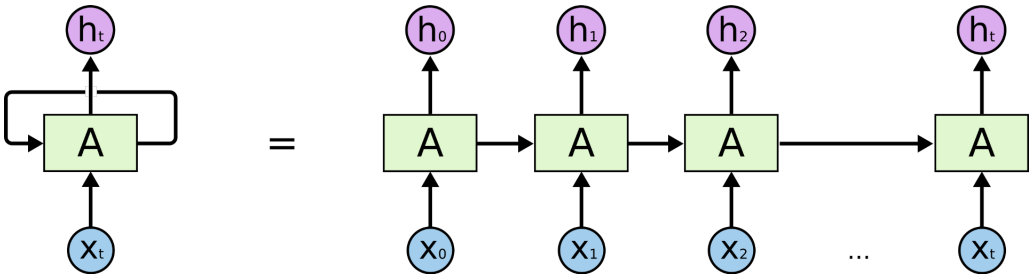


Figure 1: Fold and Unfold representation of a RNN

Very similar to a feedforward network, Equation 13 shows the weighted output at layer j. The first term on the right-hand is just the feedforward's weighted output, while the second term is the time-dependent term. Matrix ω is a hidden-state-to-hidden-state matrix. Basically, we are adding previous informations to our new state at time t.

$$z[t]^l = (W^l a[t]^{l-1} + b_j^l) + (\omega^l a[t-1]^{l-1}) \tag{13}$$

Equation 14 shows the activation of layer j at time t.

$$a[t]^l = \sigma(z[t]^l) \tag{14}$$

While Equation 15 shows the updating rule for the weight w^l

$$\begin{aligned} \frac{\partial E}{\partial w[T]l_{jk}} &= \sum_{t=0}^{T} \frac{\partial E}{\partial z[t]_j^l} \\ \frac{\partial E}{\partial w[T]l_{jk}} &= \sum_{t=0}^{T} a[t]_k^{l-1} \delta[t]_j^l \end{aligned} \tag{15}$$

7.2 Loss

Usually in a RNN the loss used function is the **cross-entropy loss**. For example, suppose we have N samples with each sample labeled by $n, ...N$. The loss function is then given by:

$$E = -\frac{1}{N}\sum_{n=1}^{N} t_n \log y_n \tag{16}$$

Where t is the true label and y our prediction

7.3 Vanishing Gradient Problem

Since the network layers and time steps are related to each other through multiplication, the gradient becomes smaller and smaller at each t. Figure 2 shows the effect of applying *sigmoid* function over time. The data is flatted more and more at each step and therefore the slope will $\rightarrow 0$.

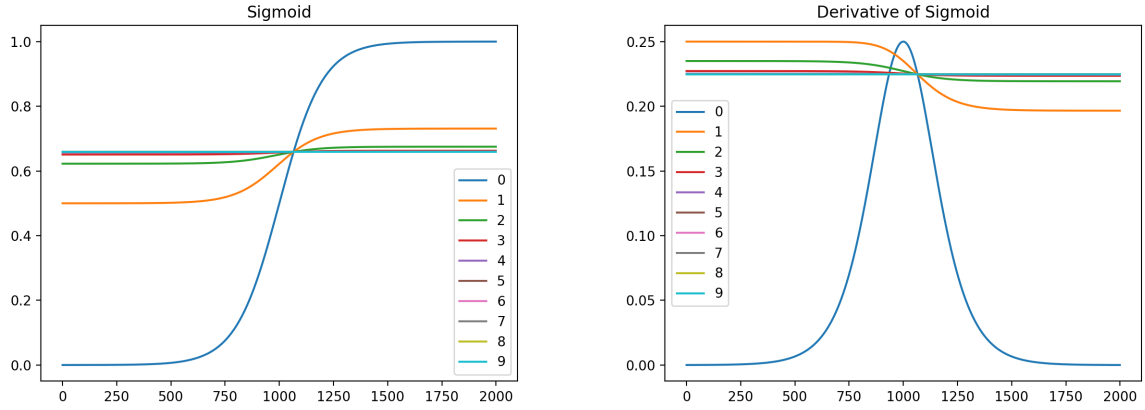


Figure 2: Sigmoid and its derivative over time

Since gradient expresses the change in all weights with respected to the error, without nothing it we cannot adjust properly the network.

8 Long Short Term Memory

8.1 Definition

The Long Short Term Memory networks, or just **LSTM**, are a special type of RNN capable of learning long-term dependencies. They were introduced by Juergen Schmidhuber in order to solve the vanish gradient problem. They are composed by LSTM cell, Figure 3 shows a unrolled representation.

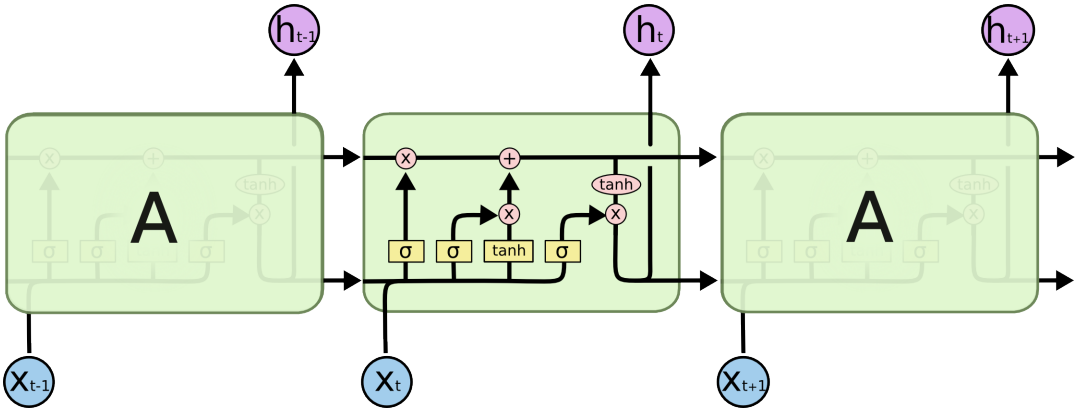
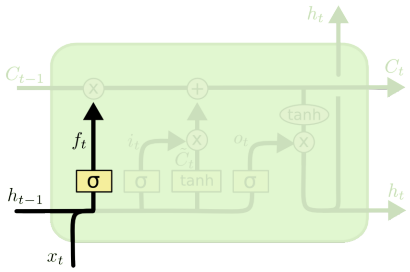


Figure 3: An unrolled LSTM

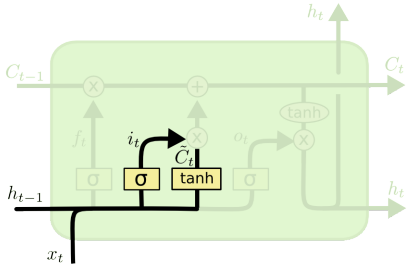
Each cell is composed by 3 gate: **forget** gate (f_t), **input** gate i_t and **output** gate (o_t). It takes as input the previous output h_{t-1} and the old cell state C_{t-1}, it outputs the next prediction and state, h_t and C_t. The cell computes four basic operations:

1. Forget Gate



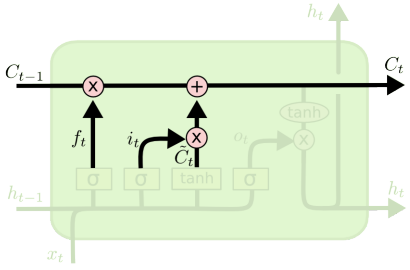
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

2. Input Gate



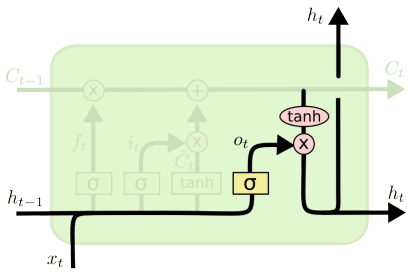
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

3. Update Cell State



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

4. Output Gate



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

9 Support Vector Machine

A Support Vector Machine maps the input space to a high dimensional space, called *feature space*, using a set of non linear function, called the *feature vector*, defined as $\phi(x)$. Then, it constructs an optimal hyperplane in order to separate the features discovered in the *feature set* in order to classify the points that belong to the positive or negative class. Equation 17 shows the decision surface.

$$w^T \phi(x) = 0 \tag{17}$$

Where $\phi(x) = \{\varphi_j(x)\}_j^\infty$, is the set of non-linear function, called *feature vector*. It tries to find the **support vectors**, a subset of the data-set, in order to maximise the margin between them using Quadratic Programming.

9.1 Linear separable data

In the simple linear case, where the data is linearly separable, Equation 18 can be re-express by Equation 18, notice that we do not need the non-linear map anymore:

$$w^T x + b = 0 \tag{18}$$

Equation **??** shows the optimal hyperplane for the linear case

$$\begin{array}{lll} w_0^T x + b_0 \geq +1 & \text{for} & d_i = +1 \\ w_0^T x + b_0 \leq -1 & \text{for} & d_i = -1 \end{array} \tag{19}$$

In this case, the **support vectors** is a small subset of points (x_i, d_i) of the dataset for which Equation 19 holds. Figure 4 shows a graphic representation.

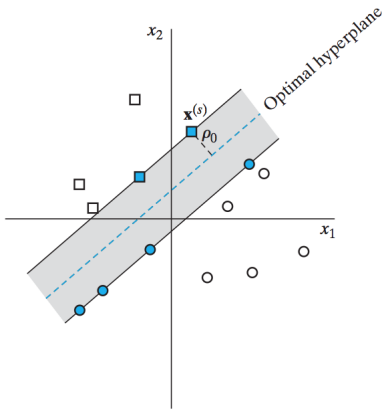


Figure 4: Support Vectors

Therefore we want to find the maximum margin possible, defined in Equation 21.

$$g(x) = wx + b = \pm 1 \tag{20}$$

$$r = \frac{g(x^{(2)})}{||w_o||}$$

$$r = \begin{cases} \frac{1}{||w_o||} & \text{if} \quad d^{(s)} = 1 \\ -\frac{1}{||w_o||} & \text{if} \quad d^{(s)} = -1 \end{cases} \tag{21}$$

$$r = \frac{2}{||w_o||}$$

Where $g(x)$ is the discriminant function, Equation 20, $x^(s)$ are the support vectors and w_0 is the weight. Therefore, maximising the margin is equal to minimising w_0, so we can state our object function and the constrains, defined in Equation 22. This is call **hard-margin**.

$$\begin{aligned} \min \quad & \Phi(w) = \tfrac{1}{2} w^T w \\ \text{subject to} \quad & d_i(w_i^T w_i + b_i) \geq 1 \qquad \text{for} \quad 1 = 1, 2, ..., N \end{aligned} \tag{22}$$

This basic linear classifier can still be used to classify non-linear separable data by adding a set of non negative variables, *slacks*, $\{\epsilon\}_i^N$ in order to measure the deviation of the data from the ideal condition of pattern separability. Equation 23 shows the updated equation, this method is called **soft margin**

$$\begin{aligned} \min \quad & \Phi(w) = \tfrac{1}{2} w^T w + c \textstyle\sum_{k=0}^{R} \epsilon_k \\ \text{subject to} \quad & d_i(w_i^T w_i + b_i) \geq 1 - \epsilon_i \qquad \epsilon_i > 0 \quad \text{for} \quad \text{all} \quad i \end{aligned} \tag{23}$$

We skip the dual problems for simplicity, they are very simular to the non linear case only there is no *kernel* and for the case of *soft margin* the Lagrange multipliers must be also less than C. Thus we can express the optimal weight using this new information as follow:

$$w_o = \sum_{i=1}^{(s)} \lambda_{0,j} d_i x_i^T x^{(s)} \tag{24}$$

There $\lambda_{0,j}$ is a Lagrange multiplier.

9.2 Non linear separable data

As we stated before, a SVM maps from a linear input set to a non-linear *feature set* using the *feature map* to infinite dimensional space, called feature space, then it perform a linear map to the *output state* . Figure 5 shows a graphical representation of a generic SVM

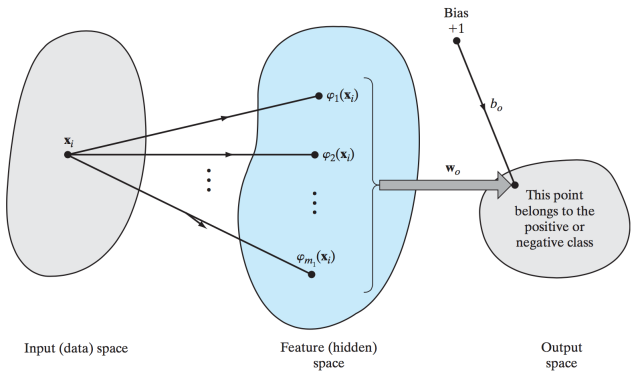


Figure 5: Mapping in a SVM

Specifically, the *feature map* is made by multiplying the inputs point with the *feature vector*, a set of non linear function that transform , defined as follow:

$$\phi(x) = \{\varphi_j(x)\}_{j=1}^{\infty} \tag{25}$$

Keeping in mind the Equation 19, the optimal decise surface, we can re-state the optimal weight Equation 24 as follow:

$$w = \sum_{i=1}^{N^{(s)}} a_i d_i \phi(x_i) \tag{26}$$

Hence, substituting Equation 19, we obtain

$$\sum_{i=1}^{N^{(s)}} a_i d_i \phi(x_i)^T \phi(x) = 0 \tag{27}$$

The term $\phi(x_i)^T \phi(x)$ represent a *inner product*, let us denote it as the scalar:

$$k(x, x_i) = \phi(x_i)^T \phi(x) = \sum_{j=1}^{\infty} \varphi_j(x_i)\varphi_j(x) \tag{28}$$

The function $k(x, x_i)$ is called the **kernel**. According to this, we can re-express the optimal decise surface, Equation 29, by:

$$\sum_{i=1}^{N^{(s)}} a_i d_i k(x_i, x) = 0 \tag{29}$$

Basically, the **kernel** computes the inner product of the images produced in the feature spaces under embedding ϕ of two data points in the input space. Equation 30 states the dual problem.

$$\begin{aligned} &\text{Maximixe } \lambda \quad \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j d_i d_j k(x_i, x_j) \\ &\text{Subject to} \\ &\sum_{i=1}^{N} \lambda_i d_i = 0 \\ &0 \leq \lambda_i \leq C \end{aligned} \tag{30}$$

The kernel must be decided *a priori* from the user.

10 Deep Learning

10.1 Supervised Learning

In **supervised** learning there is a fixed training set consists on a set of example patterns and their targets. The objective is to learn a map between the inputs and the targets. 'Supervised' comes from the likeness between a student that tries to answer a question, and a teacher that knows the solution.

Therefore, at each step the algorithm can know how much wrong is prediction was, using a cost function, and it can improve it by applying a learning algorithm such as gradient descent in order to reduce the error.

10.2 Reinforcement Learning

In **reinforcement** learning the *agent* takes *actions* in an *environment* in an to maximise a *reward*. Figure 6 shows the actions flow.

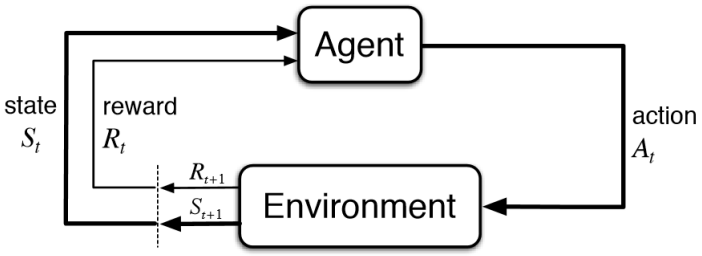


Figure 6: Reinforcement Learning Flow

10.3 Unsupervised Learning

10.4 Training Techniques

10.4.1 Mini-batch

10.5 Regularisation

10.5.1 L1 Regularisation

10.5.2 Dropout Regularisation

10.6 Activations Functions