Data Analysis 3

Neural Networks

Jan Platoš

November 18, 2018

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Table of Contents

Artificial Neural Networks

Artificial Neural Networks

- Artificial Neural Networks (ANN) are models of the human nervous system.

Human nervous system

- The system is composed of cells, called neurons.
- The neurons are connected to other neurons using synapses.
- The strength of the synapses is affected by learning (external stimuli).

Artificial Neural Networks

- The system is composed of nodes, called neurons.
- These neurons are units of computation that:
  - Receive the inputs from other neurons.
  - Processes these inputs (computes).
  - Set its output.
- The computation process is affected by the input weights and activation function.
- The weights are analogous to the strength of the synapse.
- The weights are affected by the learning process.

Artificial Neural Networks

- The neural networks ability to learn is based on the architecture of the network.
    - Single-layer neural network.
    - Multi-layer neural network.
    - Recurrent neural networks.
    - Kohonen Maps (Self Organized Maps).
    - Convolution networks.
    - Deep neural networks.
    - ...
    -
- The learning is done by presenting the test instances to the network and correction of the output according to the expected output by weight adjusting.

Single-layer Neural Network: The Perceptron

- The basic architecture of neural network.
- The structure has two layers.
    - The input layer has one node for each input attribute.
    - The input node only transmit the input value to the output node.
    - The connection between input and output nodes are weighted.
    - The output layer consist of one output neuron.
    - The output neuron computes the output value.
- The class labels are from the set of $\{-1, +1\}$.

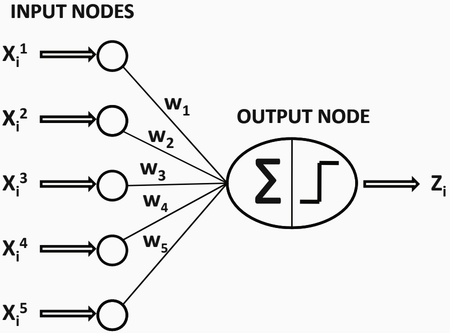

Figure 1: The Perceptron

- The weighted inputs are transformed into output value.
- The value in drawn from the set $\{-1, +1\}$.
- The value may be interpreted as the perceptron prediction of the class variable.
- The weights $W = \{w_1, \ldots, w_d\}$ are modified when the predicted output does not match expected value.

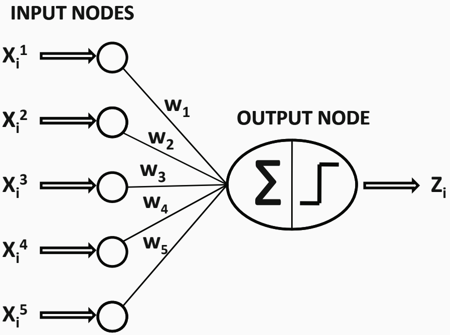

Figure 2: The Perceptron

- The function learned by the perceptron is referred as *activation function*.
- The function is usually signed linear function (e.g. weighted sum).
- The $W = \{w_1, \ldots, w_d\}$ are the weights for the connections of d different inputs to the output neuron.
- The d is also the dimensionality of the data.
- The b is the bias associated with the activation function.
- The output $z_i \in \{-1, +1\}$ is for the data record $\overline{X_i} = (x_i^1, \ldots, x_i^d)$ computed as follows:

$$z_i = sign\left\{\sum_{j=1}^{d} w_j x_i^j + b\right\} = sign\left\{\overline{W} \cdot \overline{X_i} + b\right\}$$

$$z_i = sign\left\{\sum_{j=1}^{d} w_j x_i^j + b\right\} = sign\left\{\overline{W} \cdot \overline{X_i} + b\right\}$$

- The difference between the prediction of the class value z_i and the real class value y_i is $(y_i - z_i) \in \{-2, 0, 2\}$.
- The result is 0 when the prediction and reality is the same.
- The weight vector $\overline{W}$ and bias b need to be updated, based on the error $(y_i - z_i)$.
- The learning process is iterative.
- The weight update rule for i-th input point $\overline{X_i}$ in t-th iteration is as follows:

$$\overline{W}^{t+1} = \overline{W}^{t} + \eta(y_i - z_i)\overline{X_i}$$

- The η is the learning rate that regulate the learning speed of the network.
- Each cycle per input points in the learning phase is referred as an *epoch*.

$$\overline{W}^{t+1} = \overline{W}^{t} + \eta(y_i - z_i)\overline{X_i}$$

- The incremental term $(y_i - z_i)\overline{X_i}$ is the approximation of the negative of the gradient of the least=squares prediction error $(y_i - z_i)^2 = (y_i - sign(\overline{W} \cdot \overline{X_i} - b))^2$
- The update is performed on a tuple-by-tuple basis not a global over whole dataset.
- The perceptron may be considered a modified version of a gradient descent method that minimizes the squared error of prediction.
- The size of the η affect the speed of the convergence and the quality of the solution.
    - The higher value of η means faster convergence, but suboptimal solution may be found.
    - Lower values of η results in higher-quality solutions with slow convergence.
- In practice, η is decreased systematically with increasing number of epochs performed.
- Higher values at the beginning allows bigger jumps in weight space and lower values later allows precise setting of the weights.

Multi-layer Neural Network

- The perceptron, with inly one computational neuron produces only a linear model.
- Multi-layer perceptron adds a hidden layer beside the input and output layer.
- The hidden layer itself may consist of different type of topologies (e.g. several layers).
- The output of nodes in one layer feed the inputs of the nodes in the next layer - this behavior is called *feed-forward network*.
- The nodes in one layer are fully connected to the neurons in the previous layer.

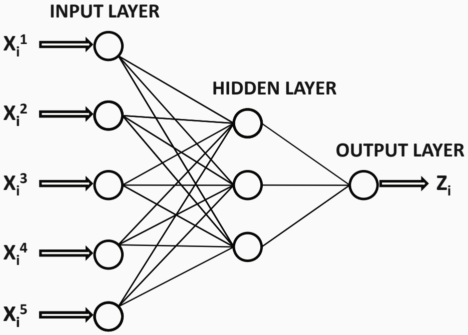

Figure 3: Multi-layer neural network

Artificial Neural Networks - Multi-layer Neural Network

- The topology of the multi-layer feed-forward network is determined automatically.
- The perceptron may be considered as a single-layer feed-forward neural network.
- The number of layers and the number of nodes in each layer have to be determined manually.
- Standard multi-layer network uses only one hidden layer, i.e. this is considered as a two-layer feed forward neural network.
- The activation function is not limited to linear signed weighted sum, other functions such as logistic, sigmoid or hyperbolic tangents are allowed.

Sigmoid/Logistic function	$\sigma(x) = \frac{1}{1+e^{-x}}$
TanH	$\mathbf{tanh}(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$
ReLU (Rectified linear unit)	$f(x) = \begin{cases} 0 & for x \leq 0 \\ x & for x \geq 0 \end{cases}$
Sinc	$f(x) = \begin{cases} 1 & for x = 0 \\ \frac{sin(x)}{x} & for x \neq 0 \end{cases}$
Gaussian	$f(x) = e^{x^2}$
Softmax	$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$

Learning algorithm

- The learning phase is more complicated than the one in perceptron.
- The biggest problem is the get the error in the hidden layer, because the direct class label is not defined on this level.
- Some kind of *feedback* is required from the nodes in the forward layer to the nodes in earlier layers about the *expected* outputs and corresponding errors.
- This principle is realized in the *back-propagation* algorithm.

Back-propagation algorithm

- *Forward phase:*
    - The input is fed into input neurons.
    - The computed values are propagated using the current weights to the next layers.
    - The final predicted output is compared with the class label and the error is determined.
- *Backward phase:*
    - The main goal is to learn weights in the backward direction by providing the error estimation from later layers to the earlier layers.
    - The estimation in the hidden layer is computed as a function of the error estimate and weight is the layers ahead.
    - The error is estimated again using the gradient method.
    - The process is complicated by the using of non-linear functions n the inner nodes.

Artificial Neural Networks - Multi-layer Neural Network - Learning algorithm

Other learning algorithms:

- Gradient descent
- Stochastic Gradient Descent
  - Momentum
  - Averaging
  - AdaGrad
  - RMSProp
  - Adam
- Newton's method
- Conjugate gradient
- Quasi-Newton method
- Levenberg-Marquardt algorithm

Artificial Neural Networks - Multi-layer Neural Network

- The multi-layer neural network is more powerful than kernel SVM in its ability to capture arbitrary functions.
- It has ability not only to capture decision boundaries of arbitrary shapes, but also noncontiguous class distribution with different decision boundaries in different regions.
- With increasing number of nodes and layers, virtually any function may be approximated.
- **The neural networks are universal function approximators**.
- This generality brings several challenges that have to be dealt with:
    - The design of the topology presents many trade=off challenges for the analyst.
    - Higher number of nodes and layers provides greater generality but also the risk of over-fitting.
    - There is very little guidance provided from the data.
    - The neural network has poor interpretability associated with the classification process.
    - The learning process is very slow and sensitive to the noise.
    - Larger networks has very slow learning process.

Artificial Neural Networks

Recurrent Neural Network

- A class of artificial neural network where connections between units form a directed cycle.
- This structure allows the dynamic temporal behavior or memory.
- Such network are able do deal with sequences of inputs as sequences not isolated inputs.
- Several types if defined in the literature:
  - **Fully recurrent network** - basic architecture with recurrent connection in each level and time-varying activation function.
  - **Recursive neural network** - the network that applies the same weights recursively over a graph-like structure. Designed for representation of the structures like logical terms.
  - **Hopfield network** with proper learning method it is a robust content-addressable memory. Its variations it is the bidirectional associative memory.
  - **Echo state network** - a special RNN that has sparsely connected random hidden layer. Only the weights of the output neuron may be changed during training.
- Training is performed by gradient descent or global optimization techniques.

Self-organizing Map

- A special class of the neural network that is used for unsupervised learning.
- The structure of neurons is a plane with squared or hexagonal grid.
- Each node contains a *d*-dimensional vector - a value of the neuron.
- The training of the network is realized as a searching for the closest neuron in feature space (the winner) and moving the neuron value towards the input neuron.
- The neighborhood of the winner is modified too when the size of the move is inversely proportional to the distance from the winner.

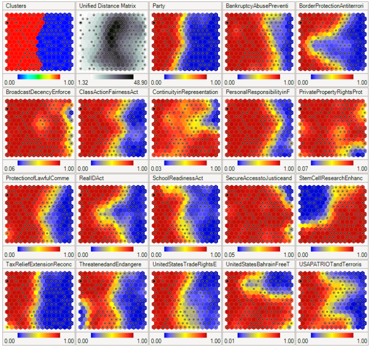

Figure 4: Self-organizing Map

Convolutional neural network

- A version of MLP that is inspired by the visual perception of animals.
- Instead of fully connected layers it deals with the image processing with different structure.
- The layers of a CNN have neurons arranged in 3 dimensions: width, height and depth.
- The neurons inside a layer are only connected to a small region of the layer before it, called a receptive field (filters).
- CNNs exploit spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers.
- Stacking many layers leads to non-linear ”filters” that become increasingly ”global”.
- This allows the network to first create good representations of small parts of the input, then assemble representations of larger areas from them.
- Each filter is replicated across the entire visual field with the same parametrization (weight vector and bias) and form a feature map.
- Features are detected regardless of their position in the visual field.

Artificial Neural Networks - Convolutional neural network

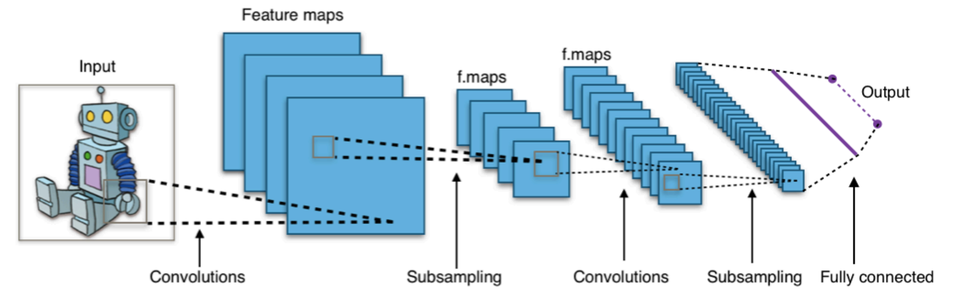

Figure 5: Typical CNN (source Wikipedia)

Questions?