DOMS | Page No
Date / /

FDL

* ANN MLP

→ If the dataset is only linearly separable, then only you can use SLP.

→ Hidden layers always use RELU / family of Relu, because it helps in achieving non-linearity.

* HOW TO DESIGN A NEURAL NETWORK.

→ MLP is used solve only supervised learning.

→ In output layer of regression, there should be only 1 node, whereas for classification, it depends on the no of classes.

→ For output layer of regression, linear activation function is used so that the output from last hidden layer is not changed.

→ LeakyReLU was invented to avoid dead neuron problem.

→ Leaky ReLU is max(0.1x, x)

→ Parametric ReLU is: max(yx, x) ; y is variable.

If y=0.1 then it is leaky.

→ For binary classification, output layer has sigmoid activation function.

→ If in classification, the output labels are in negative values then we use tanh activation.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

→ For multi-class classification, no of neurons in output layer is no of classes. even tho each sample can only belong to 1 class at a time, because there is no activation function which can produce the class ids as output.

→ Softmax is used to predict the probabilities.

$$\text{Softmax}(x) = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}}$$

→ For multi-label classification, output layer uses sigmoid/tanh activation function. It is just a concept.

★ LOSS FUNCTIONS.

→ If label encoding is done in multi-class categorical classification, then loss is sparse-categorical-cross entrophy

→ If one-hot encoding is done, then the loss used is categorical cross entropy.

→ For binary classification, loss used is binary cross entropy.

→ $L_{BCE} = -\frac{1}{n} \sum_{i=1}^{n} y_i (\log \hat{y}_i) + (1 - y_i)(\log (1 - \hat{y}_i))$

Q)

	x_1	x_2	y	$\hat{y}$
S1	0	0	1	0.8
S2	0	1	0	0.2
S3	1	0	0	0.6
S4	1	1	0	0.9

$$BCE = -\frac{1}{4}\left[(1 \times \log(0.8)) + (1-1)(\log(1-0.8)) + (0 \times \log(0.2)) + (0-0)(\log(1-0.2))\right]$$

$$BCE = 0.367$$

→ $L_{CCE} = -\sum_{i=1}^{n} t_i \log(s_i)$ T_i = ground truth
S_i = softmax probability

→ For regression, loss used is mse / mae / rmse / mape / R2

→ why only mse, mae, etc are used & other loss functions discovered is all graph which can converge to the global minima.

* DROPOUT

→ Applied to avoid overfitting by not depending on certain features or patterns.

→ Dropout can be used in input & hidden layers but never in output layers.

→ Safe zone of dropout is 10-50%.

→ When dropout is applied in training phase, then the weights are scaled down by multiplying it by the dropout % before testing to avoid overfitting etc.