Activation functions cannot be linear because neural networks with a linear activation function are effective only one layer deep, regardless of how complex their architecture are. Input to networks are usually linear transformation (input \* weight), but real world and problems are non-linear. To make the incoming data nonlinear, we use nonlinear mapping called activation function. An activation function is a decision making function that determines the presence of particular neural feature. It is mapped between 0 and 1, where zero mean the feature is not there, while one means the feature is present. Unfortunately, the small changes occurring in the weights cannot be reflected in the activation value because it can only take either 0 or 1. Therefore, nonlinear functions must be continuous and differentiable between this range. A neural network must be able to take any input from -infinity to +infinite, but it should be able to map it to an output that ranges between {0,1} or between {-1,1} in some cases - thus the need for activation function. Non-linearity is needed in activation functions because its aim in a neural network is to produce a nonlinear decision boundary via non-linear combinations of the weight and inputs.

https://stackoverflow.com/questions/9782071/why-must-a-nonlinear-activation-function-be-used-in-a-backpropagation-neural-net