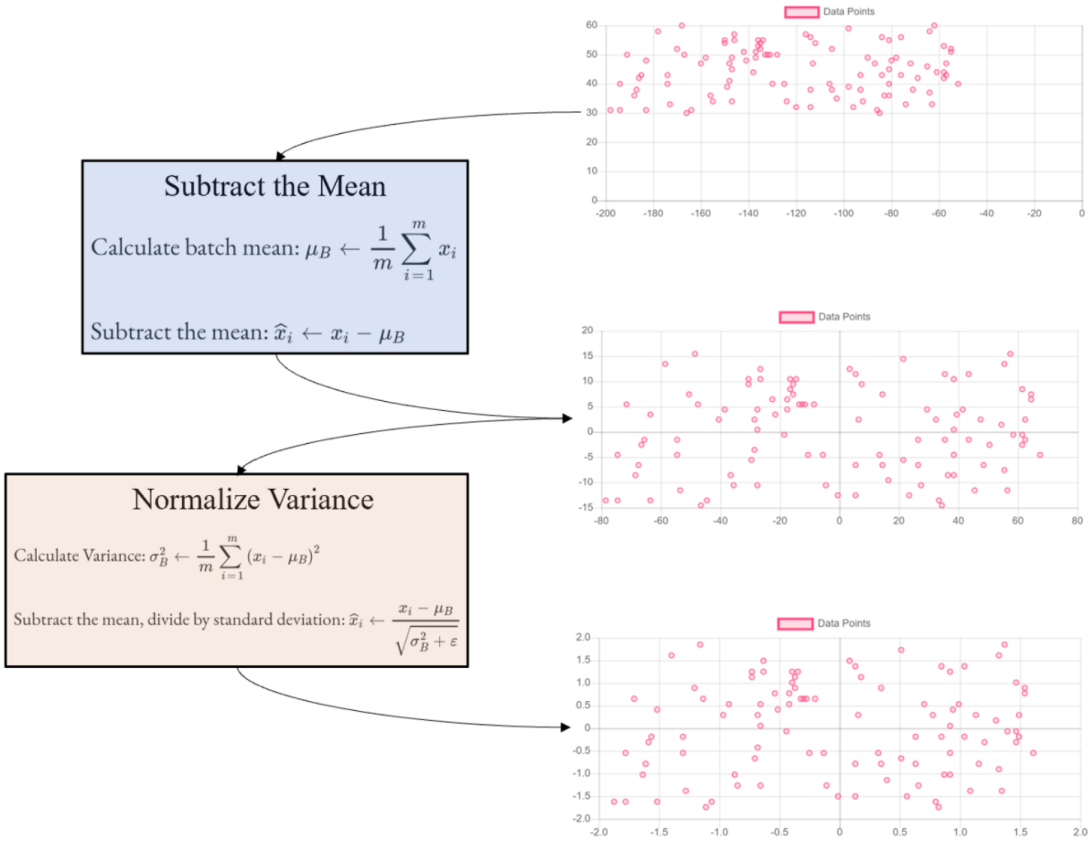
Dataset Cleaning / Pre-processing

Dataset cleaning/ pre-processing is a vital sub-task for deep learning models' training. The rationale behind this step is:

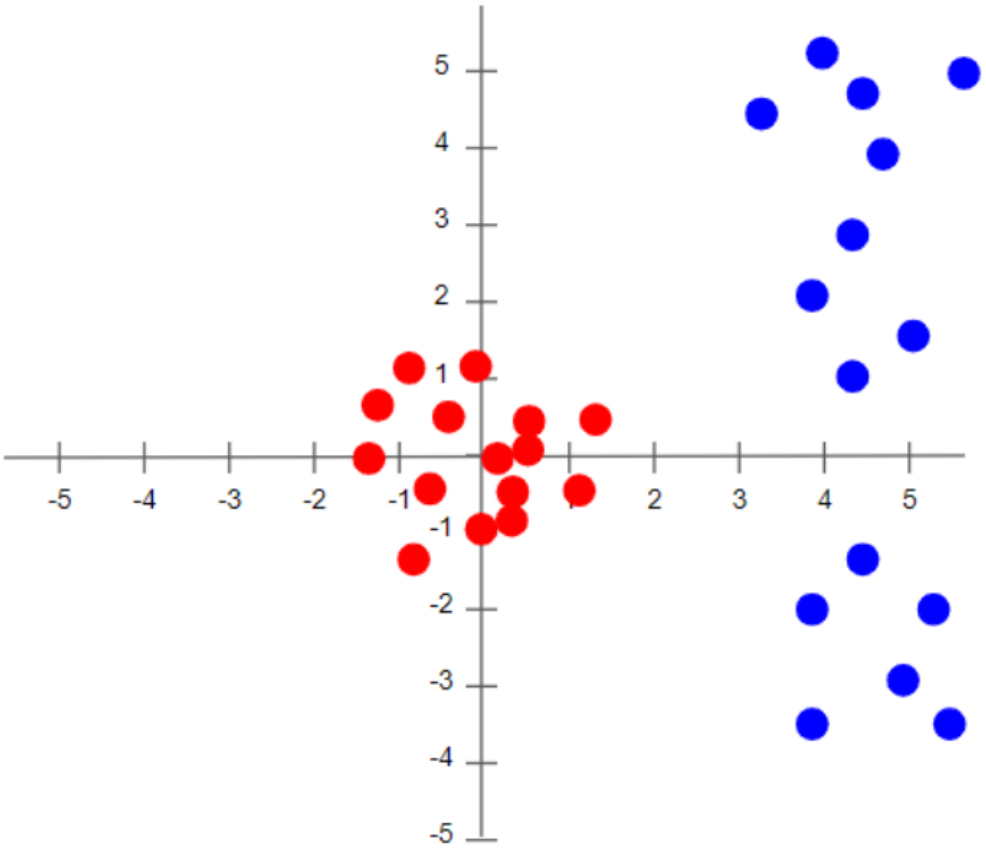
1. The input of the network needs to be in a non-chaotic distribution. Since neural networks essentially model the distribution around the input, it is crucial that the inputs (both training and testing are in the same domain).
2. The outputs that we expect can not be from any distribution. This is due to the non-linear activations available at the hands. A bounded nonlinearity (sigmoid, tanh) makes the modeling much easier than the un-bounded distribution (ReLU, Leaky-ReLU).

Detailed Explanation on the choice of pre-processing methods:

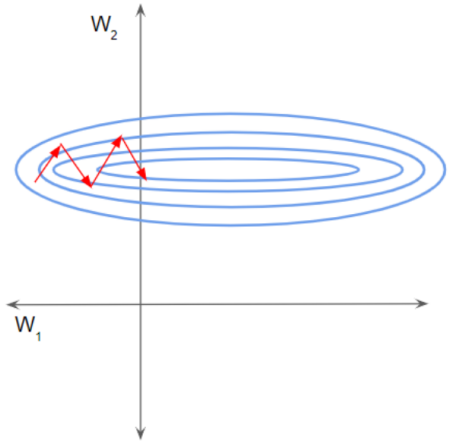
1. Input Data (SAR):
    - For SAR images, the inputs, we chose to go with zero mean and 1 Standard-deviation distribution (Standard gaussian). This is done to ensure the concept of whitening is implanted into the training.



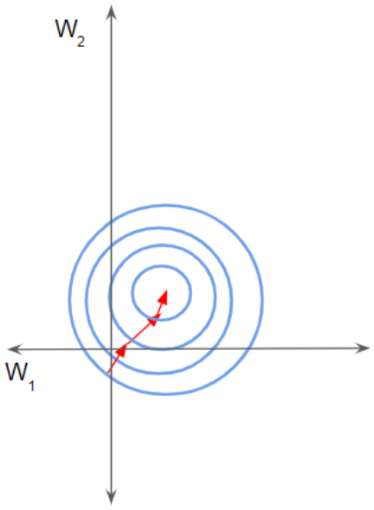
- This normalization makes it easier for the model to train in a stable fashion without the gradients swinging aggressively.



- If the inputs are not normalized, the space would look something like the image below, as can be seen, the weights would have to oscillate alot to reach the minima.



- Normalizing the inputs, makes the space zero centered and uniform, making the gradient descent smoother and faster.



- For these reasons, we have gone with the approach taking a sample mean and standard deviation from the training set, implying it to be the mean and std of the distribution. Using these values we have normalized the training set and will continue the same pre-processing during testing.

2. Output Data (Optical)
    - For the output data, we have to consider the output non-linear activation of the neural network. This is critical as the network will fail if the expected output space and actual output space do not match.
    - We have carefully examined all the activations and decided to go with a bounded nonlinear function, TanH activation for the following reasons:
        - The nonlinearity plays a major role (S-shaped curve makes it easier since Batch normalization can be used after activation, making it faster to train)
        - The bounded-ness of the function makes the prediction easier as the output space is bounded (always between -1 and 1)
        - The -ve and +ve values bring in the reward-punishment duality into the loss function if the function is designed carefully.
    - For these reasons, we have taken a standard min-max normalization with scale between -1 and 1

Agenda of meeting to be scheduled on November 2021:

1. Visual results of the pre-processing
2. Discussion on possible draw-backs and hitches
3. Testing the efficiency on a dummy network
4. Discussions on possible neural network architecture design
5. Deciding on a preliminary model for training.

Meeting Minutes from xx-11-2021:

a.