Geospatial data classification using Principal component analysis

Dataset='sst.mon.mean.nc'

The dataset contains monthly global sea surface temperature data. There is a region over pacific ocean called Nino 3.4 (longitude=[170:290] ,latitude= [84:96]). If the average sst ANOMALY(True value- long term mean) value for the region for a month is greater than .5 then it is called 'El-nino' if less than -.5 then 'La-nina'.

You need to perform classification task using PCA and NN classifier to classify the input data into 'El-nino' and 'La-nina'.

The steps to be performed are:

1)Dataset preparation and visualization

1.a)Load dataset using NetCDF4

1.b)Separate lon,lat,sst and time in different variables

1.c)Plot data on Basemap taking the mean Nino 3.4 region ([170:290] for longitude and [84:96] for latitude) over the years

1.d)Prepare sst anomaly (sst value - mean sst value) for the whole dataset as model input

1.e)Calculate sst anomaly for the nino 3.4 region

1.f)Calculate mean nino 3.4 sst anomaly for each year

1.g)Detrend mean nino 3.4 sst anomaly

1.h)Prepare labels such that, if mean nino 3.4 sst anomaly value >.5 then 'El-nino' , mean nino 3.4 sst anomaly value <.5 then 'La-nina', discard the rest.

1.i)Transform masked array into numpy array by replacing nan values by 0.

1.j)Split into training and testing set with test set fraction as .33

2)PCA

2.a)Define and fit PCA

2.b)Transform training and testing set using PCA

2.c)Plot mean PCA map

2.d)Plot explained variance and variance ratio on graph

3)Classification

3.a)Create a device variable which will be used to shift model and data to GPU if available

3.b)Define pytorch NN classifier which reflects highest testing accuracy(around .99) with minimum number of layers. Plot the accuracy vs #layers graph to prove that your model is best.

3.c)Create a dataloader object which will create batches of data.

3.d)Train the classifier with batch sizes (16,32,64,128), learning rates (.001,.002,.003.01,.02,.03), optimizers (Adam, SGD) using GridSearch to find best configuration. Use earlystopping criteria to stop training.

3.e)Report prediction accuracy and confusion matrix

3.f)Randomly select 5 maps from testing set, plot them on map, predict classes using their PCA data and report them along with their true label.

(Note: Refer the .ipynb file uploaded in the files section)