HW 3: Deep Ensembles & Covariate Shift

The architecture of the model is as follows:

- MyClassifier: This class represents a single classifier model. It consists of three linear layers (fc1, fc2, fc3) followed by ReLU activation functions (relu). The input size of the first linear layer (fc1) is 2, corresponding to the two input features. The output size of fc1 is 64. The output size of fc2 is also 64, and the output size of fc3 depends on the number of classes (n_classes).

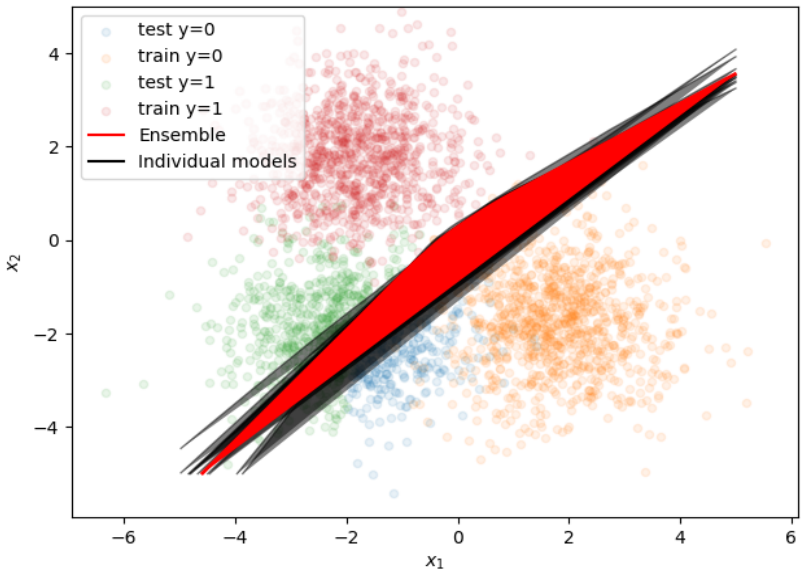
- In terms of nonlinearity, the ReLU (Rectified Linear Unit) activation function is applied after each linear layer in MyClassifier.

- MyEnsemble: This class represents an ensemble model that combines multiple MyClassifier models. It takes a list of MyClassifier models as input. In the forward pass, it computes the logits by passing the input through each model in the ensemble and averaging the logits across models using t.stack and mean operations.

Hyper-parameter choices -

batch size	learning rate	epochs	optimizer
100	0.01	10/model	Adam

Scatter plot of validation and test data, with all the classifier boundaries -



Metrics -

	Acc-train	Acc-val	Acc-test	ECE-train	ECE-val	ECE-test
Mean of models	0.996	0.995	0.819	0.004	0.006	0.082
Ensemble model	0.997	0.995	0.827	0.004	0.007	0.071

From the table of results, we can observe that the accuracy of the ensemble model is higher than the mean of the individual models across all three sets: train, validation, and test. This suggests that by pooling the predictions of various models, ensembling helps reduce the harmful impacts of covariate change.

In terms of calibration, the ECE values shed insight into the reliability of the predicted confidence from the model. Better calibration is shown by a lower ECE, which means that the projected confidence and the empirical accuracy are closely aligned. The ensemble model outperforms the mean of the individual models on both the validation and test sets, as shown by a comparison of the ECE values. This shows that ensembling aids in improving the model's calibration, even in the presence of covariate shift.

The observation is further supported by the test data's confidence calibration curve. Compared to the mean of the individual models, the ensemble model's curve fits the data better. This suggests that the anticipated confidence of the ensemble model is more accurately calibrated and offers more dependable estimates of accuracy. Overall, ensembling enhances calibration and accuracy, reducing the adverse effects of covariate change.

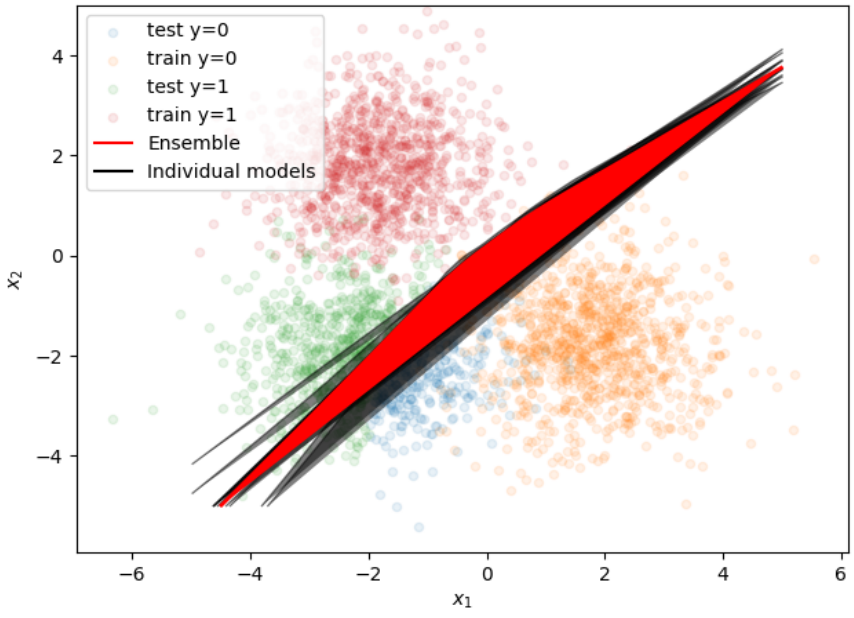
Extra credit -

Mixup - Mixup augmentation is used to bump up test accuracy of the models. Model architecture and hyper-parameter were same as baseline. Alpha value for mixup is taken 0.2. It worked better for the test accuracy but not much difference for ECE-test for both mean and ensemble of the models. It doesn't improve the calibration of the models.

Metrics -

	Acc-train	Acc-val	Acc-test	ECE-train	ECE-val	ECE-test
Mean of models	0.995	0.995	0.843	0.035	0.038	0.079
Ensemble model	0.994	0.994	0.852	0.035	0.038	0.074

Scatter plot of validation and test data, with all the classifier boundaries -

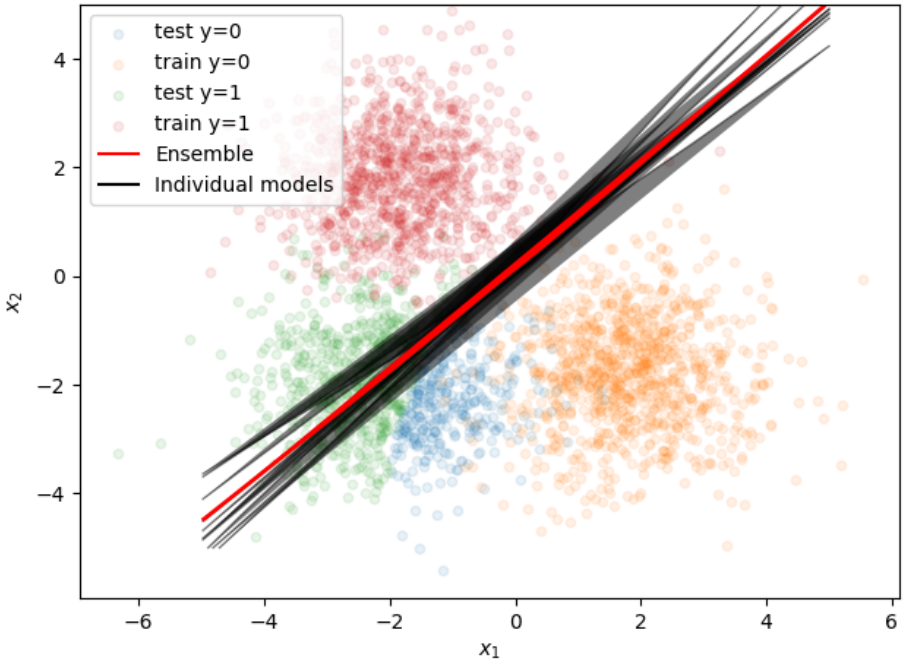


Stochastic Weight Averaging - SWA is a method used to improve the generalization of neural networks by averaging the weights of the model over multiple epochs with a cyclical learning rate schedule. For this, base learning rate is taken 0.001 and max learning rate is taken 0.1. Also, momentum is added to SGD of 0.9. All other hyperparameters and model architecture were kept same. It worked worse for the test accuracy and ECE-test for both mean and ensemble of the models.

Metrics -

	Acc-train	Acc-val	Acc-test	ECE-train	ECE-val	ECE-test
Mean of models	0.994	0.994	0.759	0.007	0.008	0.100
Ensemble model	0.994	0.996	0.751	0.007	0.009	0.096

Scatter plot of validation and test data, with all the classifier boundaries -

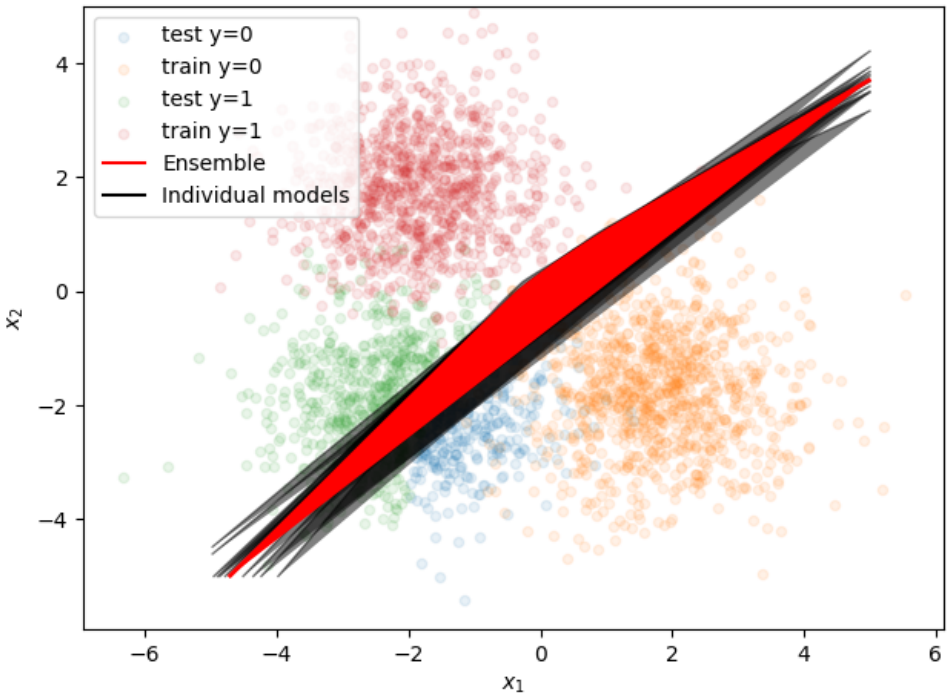


Monte Carlo - Dropout - By applying dropout during inference and ensembling the predictions over multiple samples, the test accuracy doesn't improve nor the ECE-test. Only the calibration or ECE for train and validation sets improved. Model architecture and hyper-parameter were same as baseline.

Metrics -

	Acc-train	Acc-val	Acc-test	ECE-train	ECE-val	ECE-test
Mean of models	0.996	0.996	0.810	0.003	0.006	0.104
Ensemble model	0.996	0.996	0.815	0.002	0.005	0.099

Scatter plot of validation and test data, with all the classifier boundaries -


test y=0
train y=0
test y=1
train y=1
Ensemble
Individual models
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
x_1
x_2