Assignment_6_task_6.4

Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla
Matriculation No. 256245

Assignment 6 Task 6.4

Cross validation is a way of splitting your entire dataset into training and validation data. It helps us form
a better picture of which model has the least amount of error. Sometimes we may wish to re-train the model after
testing, this second cycle of training is would be done on the entire dataset. This step has the following advantages:

1. Maximizing data utilization: helps the model find even more potential patterns and nuances possibly leading to a better final model
2. Helps finalize hyperparameters: cross-validation helps finalize the best hyperparameters (e.g. regularization strength, number of layers, learning rate etc...) Once these are fixed we re-train the model on the entire dataset.
3. The models trained on cross-validation are intermediates and not quite ready for production. This is because keeping all things equal, a model trained on more data will have better performance.

Though this does mean that even if we train the model on all the data after fixing hyperparameters, this tuning would have been done for 70% of the data rather than all of it, and there is no guarantee that the hyperparameters selected for 70% of the data will be equally effective for 100% of the data, this is why we still have "data wastage".

K-fold cross-validation is a type of cross-validation where we divide our entire dataset into 'k' disjointed subsets. The size of each subset can be arbitrary (but is usually kep equal). The model is then trained on all except
one of these subsets. The final remaining subset is used for testing. This cycle of training the model is done multiple times changing the hidden subset eveytime. Finally the error of testing each version of the model is
averaged to produce the generalization error of the model. The model with the lowest generalization error is
chosen.

Leave-one-out cross-validation is yet another type of cross-validation, where due to scarcity of training data
we train the model on all but one sample of the data, choosing a different sample to exclude everytime.

The reason we use such training methods is for the following:

1. Maximize use of training data

2. Reduction in model variance. This is possible in some cases, where the training data missed some minority samples, or had too many outliers skewing the model's performance.
3. Helps us pick more accurate hyperparameters. Since we're able to use more data to train intermediate models we are also able to finalize better hyperparameters for use in the final production model.

All of these reasons result in a more stable, robust and less biased estimate of performance.

Cross-validation also helps us detect bias and variance in models. If a model consistently shows an poor performance and high average error across all k-folds, then we can conclude that the model has high bias and may never really be able to capture the target function.

Similarly, if the model inconsistent performance on the k-folds or has it's best performance and worst performance are very far apart then the model has high variance and may be too unreliable for capturing the target function.