DSBDA – U5

Holdout method:

- Used for evaluating the performance of prediction models.
- Dataset is divided into two separate subsets:
  - Training
  - Testing
- Training set is usually larger than testing set. It is used for training the model by adjusting weights & params. (60-70%)
- Testing set is usually ~~%~~ smaller than training set. It is used for testing/evaluating the final performance of the model after training. (20-30%)
- ~~Holdout~~ Validation set is an optional set created ~~for used~~ which is used for to tune Hyperparameters & select best model config. Done after training but before testing phase. (10-20%)
- The split is usually 60/40, 70/30 or 80/20.
- Using a validation test improves model performance w/o biasing final evaluation.
- Test set provides an unbiased estimate of model performance for real-world use

↳ Eg.: If 1k samples, 700 for training, 300 testing.

↳ Adv: · Simple
· Fast
· Suitable for large datasets

Random subsampling method:

↳ Extension of holdout method.

↳ Involves performing holdout process multiple times w/ diff. random split of data.

↳ Process →
- Repeatedly split the dataset into training & testing sets randomly.
- Train & evaluate model for each split.
- Compute avg. performance over all iterations.

↳ Adv → - Reduces variance & biases caused by single train-test split.
- Flexible no. of trials.

↳ Dis → - More computation
- Some data pts. may never be selected in training/testing

(I) K-fold cross-validation:

↳ Extension of random subsampling method.

↳ More robust evaluation method that splits the dataset into k* equal-sized folds (subsets)

↳ Process :→
- The model is trained on k-1 folds & tested on remaining folds.
- This is repeated k times, each time using diff. fold as the test set.
- Results are averaged to provide more reliable estimate of model performance.

Adv → - Reduces bias
- More reliable
- Robust
- Uses all data pts. for ~~testing~~ training/testing.

Disadv → - More computationally expensive.