Resampling

- The method of repeatedly drawing samples from a training ~~da~~ set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.

1) Bootstrapping :-

- Also known as bootstrap sampling, bootstrap or random sampling with replacement.

- Bootstrapping is the process of computing performance measures using several randomly selected training and test datasets which are selected through a process of sampling with replacement.

- The bootstrap procedure will create one or more new training datasets some of which are repeated.

- The corresponding test datasets are then constructed from the set of examples that were not selected for that respective training dataset.

- The objective is to estimate the true sampling distribution of a quantity T.

① we take new samples from true population, compute T, and accumulate all of the values of T into the sampling distribution.

② taking a ~~f~~ new samples is expensive, so instead

i take a single sample, and,
, use it to estimate the population.

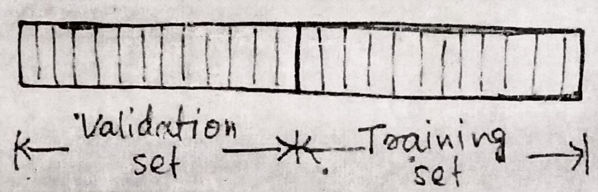
(3) we then take samples from this estimated population,
and compute T, from each,

(4) and, accumulate all values of T into an estimate
of the sampling distribution.

Module 3

Cross validation and re-sampling methods

* To test the performance of a classifier, we need to have a number of training/validation set pairs.



* Cross validation methods are used for generating multiple training-validation sets from a given dataset.

* Cross validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

Different methods are –

→ 1) Hold out method.

→ 2) K-fold cross validation.

→ 3) Leave-one-out cross validation (LOOCV)

~~→ 4) Bootstrapping~~

1) Holdout Method

* Simplest kind of cross validation.

* The dataset is seperated into two sets, called the training set and the testing set.

* The algorithm fits a function using the training set only. Then the function is used to predict the output values for the data in the testing set.

Advantages

→ Simple and easy to run

→ Lower computational cost as it only needs to be run once.

Disadvantages

→ Only work on large dataset

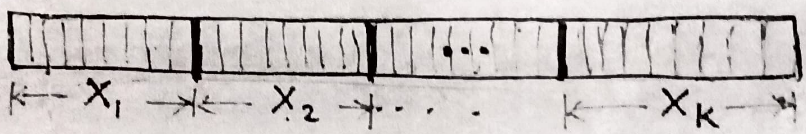
→ Higher variance given the smaller size of the data.

2) K-fold cross-validation

* The dataset X is divided randomly into K equal-sized parts, X_i; $i = 1, \ldots, K$.

* To generate each pair, we keep one of the k parts out as the validation set V_i, and combine ~~the remaining~~ the remaining $k-1$ parts to form the training set, T_i.

* Doing this k times, we get k pairs (V_i, T_i).



1st $\quad V_1 = X_1 \quad T_1 = X_2 \cup X_3 \cup \ldots \cup X_k. \quad P_1 = ?$

2nd $\quad V_2 = X_2 \quad T_2 = X_1 \cup X_3 \cup \ldots \cup X_k \quad P_2 = ?$

— — — —

kth $\quad V_k = X_k \quad T_k = X_1 \cup X_2 \ldots \cup X_{k-1} \quad P_k ?$

$$\boxed{P = \sum_{i=1} P_i}$$

Problems with this approach:

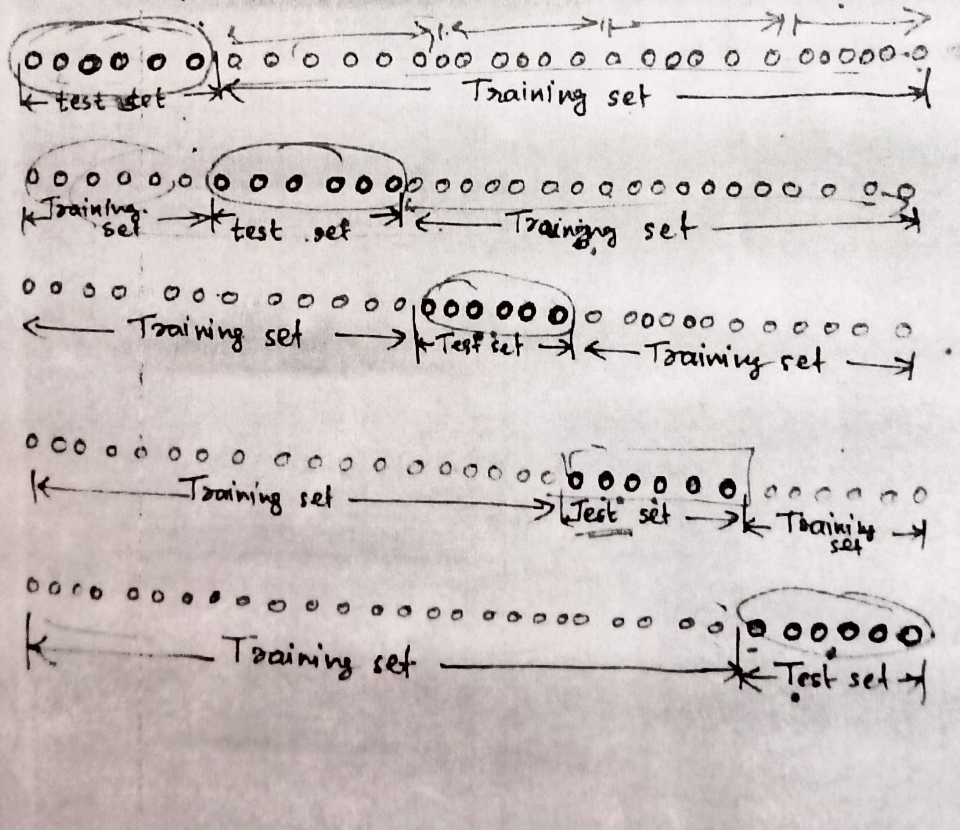
→ To keep the training set large, we allow validation sets to be small.

→ Every two training sets share $k-2$ parts.

* K is typically 10 or 30. As k increases, the percentage of training instances increases and we get more robust estimators; But the validation set becomes smaller. Also the cost of training the classifier increases as K increases.

Example:

Consider a dataset containing 30 samples. And let K = 5. Then we divide dataset into 5 folds, each fold containing 6 samples.



3) Leave - one-out cross validation (L.O.O.C.V)

* Given a dataset of N instances, only one instance is left out as the validation set and remaining N-1 instances are used for training.
* We get (N pairs) and hence N iterations are performed. (V_i, T_i)

4) Bootstrapping