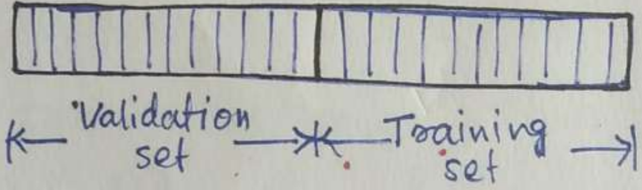
Module 3

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.

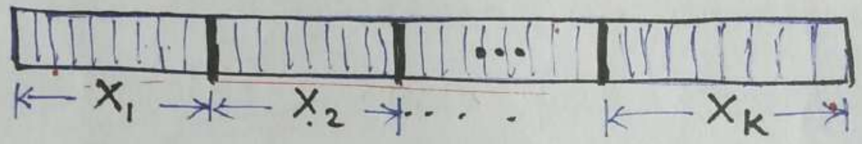
→ Higher variance given the smaller size of the data.

2) K-fold cross-validation

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

* 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 $\quad V_1 = X_1 \qquad T_1 = X_2 \cup X_3 \cup \dots \cup X_k. \qquad P_1 = ?$

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

— — — — — ,

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

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

Problems with this approach:

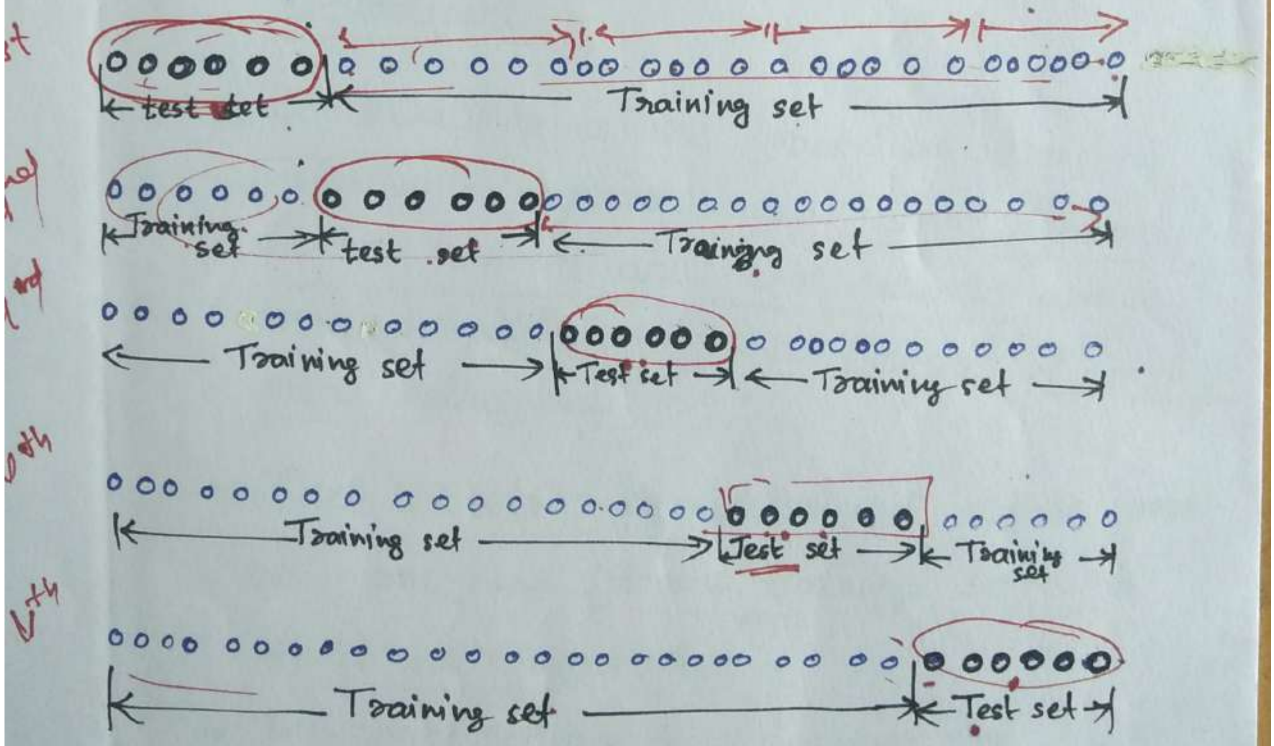
→ To keep the training set large, we allow validation sets to be small.

→ 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~~ for training.
* We get N pairs and hence N iterations are performed. (V_i, T_i)

4) Bootstrapping

* Also known as bootstrap sampling, bootstrap, ~~bo~~ 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 the respective training datasets.

Measuring Classifier Performance

Test set → Model → Performance ?

- True positive (TP)
- False positive (FP)
- True negative (TN)
- False negative (FN)

→ Confusion matrix ?

Performance:

- → Precision
- → Recall
- → Accuracy
- → Error rate.
- → Sensitivity
- → Specificity
- → F-measure.
- → ROC & AUC

→ +ve
→ −ve

TP: +Ve sick. → Classifier → +Ve sick. } True +Ve.

FP: −Ve Healthy. → Classifier. → +Ve sick — False +Ve

TN: −Ve Healthy → Classifier. → −Ve Healthy — True −Ve

FN: +Ve sick → Classifier. → −Ve Healthy. — False −Ve

7

Confusion Matrix

(2x2)

- TP
- FP
- FN
- TN

	Actual condition is +ve	Actual condition is -ve
Predicted condition is +ve	TP	FP
Predicted condition is -ve	FN	TN

Precision & Recall

Precision - It is the ratio between true positives and all the predicted positives.

* Precision is a measure of how many samples are correctly identified as +ve out of all the samples which are predicted as +ve.

$$\text{Precision} = \frac{\text{No. of TP}}{\text{Total No. of Predicted +ve}}$$

$$= \frac{TP}{TP + FP}$$

Recall – It is the ratio between true positives and all the actual positives.

* Recall is a measure of how many samples are correctly identified as +ve out of all the samples which are actually +ve.

$$\text{Recall} = \frac{\text{No. of TP}}{\text{Total no. of actual positives}}$$

$$= \frac{TP}{TP+FN}$$

Other measures:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error rate} = 1 - \text{Accuracy}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \begin{cases} \text{Recall} \\ \text{TPR} \end{cases}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F-measure} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

(Precision & Recall)

$$= \frac{2 \times \text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC & AUC

- Precision & Recall -
Problems

Problem 1

Suppose a computer program for recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats. Of the eight dogs identified, five actually are dogs while the rest are cats. Compute the precision and recall of the computer program.

Ans)

	Actual dogs	Actual cats	
Predicted dogs	T.P 5	FP 3	8
Predited cats	FN 7	TN	
	12		

dog → +ve
cat → −ve

$$\text{precision} = \frac{TP}{\text{No. predicted positive}}$$

$$= \frac{TP}{TP + FP} = \frac{5}{5+3} = \frac{5}{8}$$

$$\text{Recall} = \frac{TP}{\text{Actual Positive}} = \frac{TP}{TP + FN}$$

$$= \frac{5}{5+7} = \frac{5}{12}$$