MODEL ESTIMATION and CLASSIFIER ACCURACY MEASURE

Prof. Dr. Songül VARLI

MODEL ESTIMATION

- The model is first designed using training samples and then it is evaluated based on its performance on the test samples.

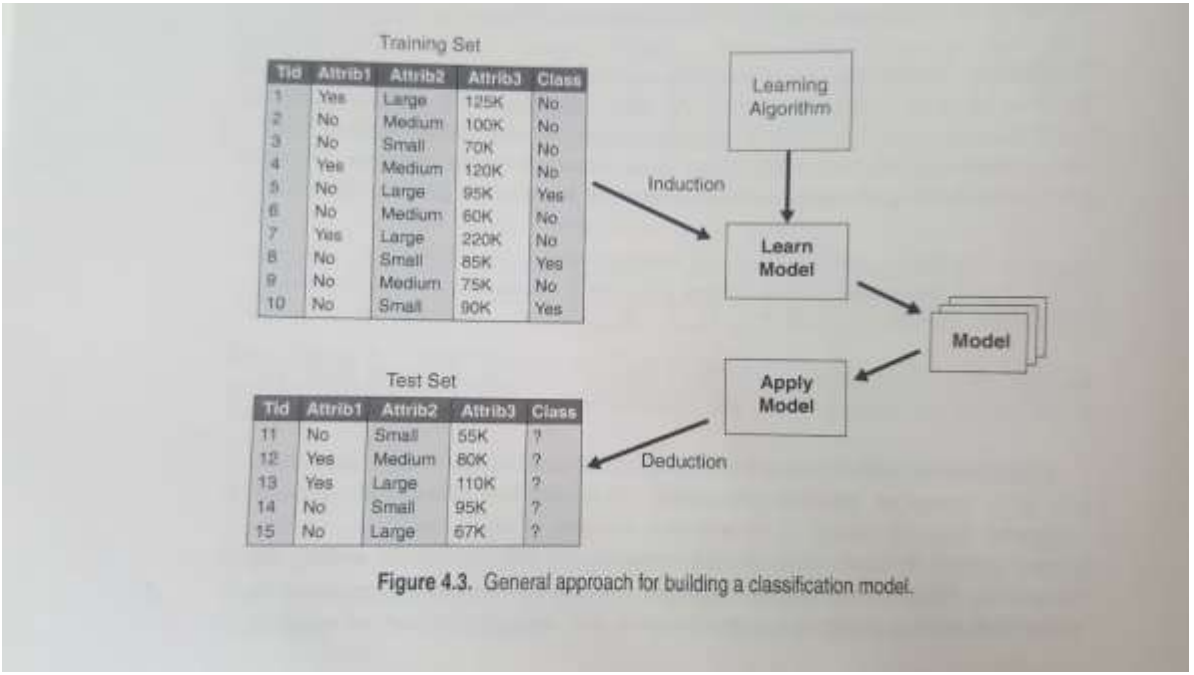
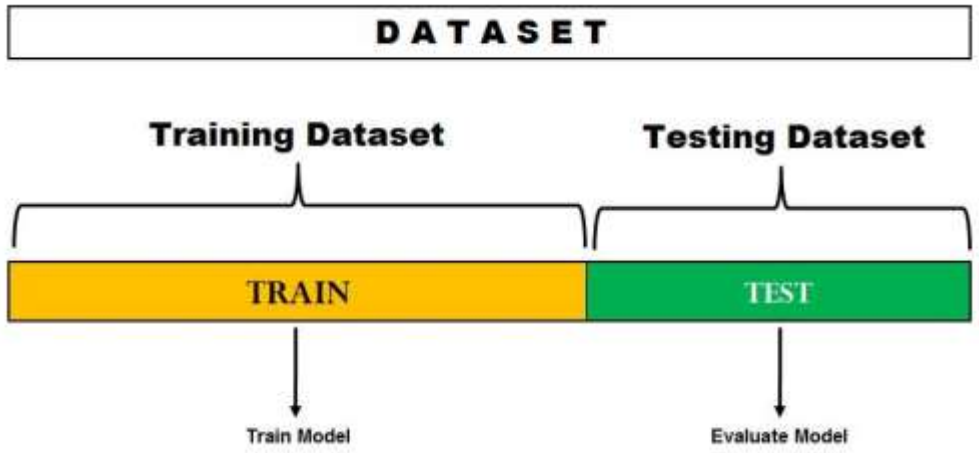
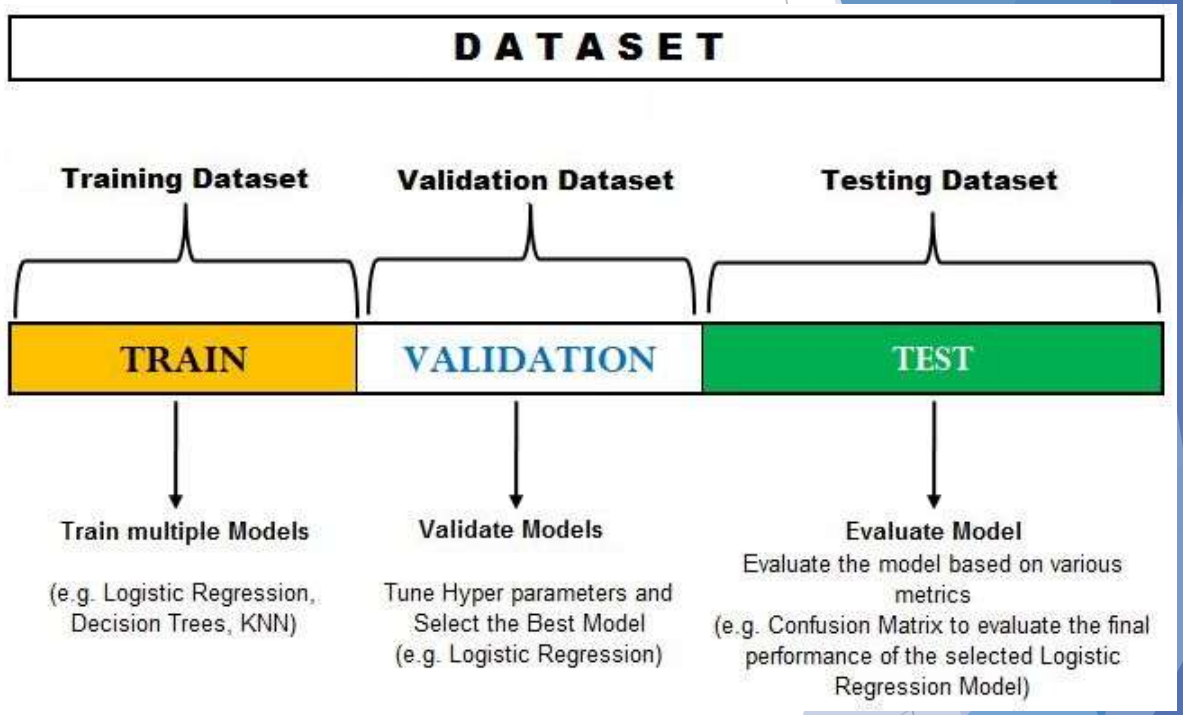


Figure 4.3. General approach for building a classification model.

The errors committed by a classisfication model are generally divided into two types: training errors and generalization errors.
Training error, also known as resubstitution error, is the number of misclassification error committed on training records, whereas generalization error is expected error of the model on previously unseen records.

Data Splitting





Model Overfitting

- A good classificaion model must not only fit the training data well, it must also accurately classisfy records it has never seen before.

- In other words, a good model must have low training error as well as low generalization error.
- It is important because a model that fits the the training data too well can have a poorer generalization error than a model with a higher training error. Such a situation is known as model overfitting.

MODEL ESTIMATION

How should the available samples be split to form training and test sets?

- If the training set is small, then the resulting model will NOT be very robust and will have low generalization ability.
- On the other hand, if the test set is small, then the confidence in the estimated error rate will be low

MODEL ESTIMATION

If the number of samples is smaller, then the designer of the data mining experiments has to be very careful in splitting data.

Data Splitting Methods:

1- Resubstitution Method
2- Holdout Method
3- Leave-one-out Method
4- Rotation Method (K-Fold Cross Validation)
5- Bootstrap Method

Data Splitting Methods:

- **1- Resubstitution Method:**

- All the available data are used for training as well as for testing. In other words, the training and testing sets are the same .
- The method is very seldom used in real World data mining applications.

Data Splitting Methods:

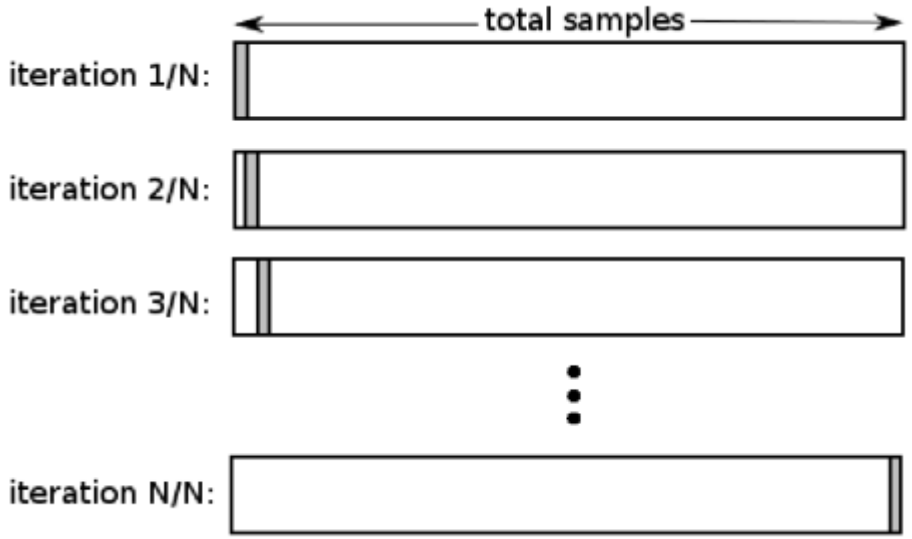
- **2-Hold-out Method:**

- Half of the data, or sometimes two-thirds of the data is used for thraininh and the remaining data is used for testing.
- Training and test sets are independent and the error estimation is pessimistic.

Data Splitting Methods:

- **3-Leave-one-out method:**

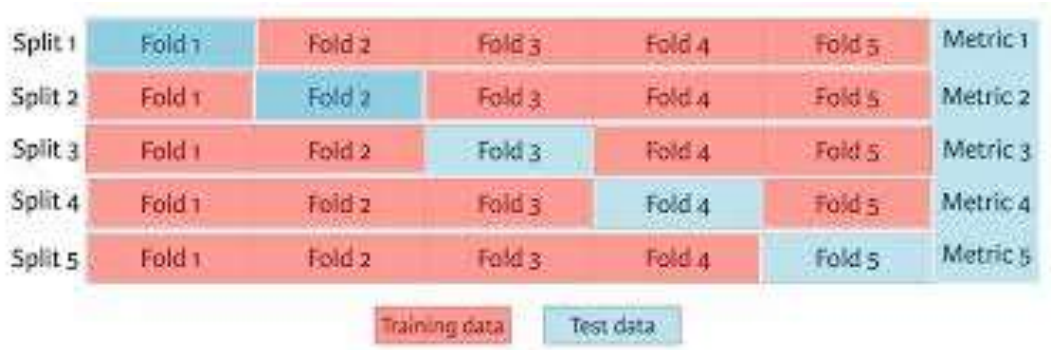
- A model is designed using (n-1) samples for training and evaluated on the one remaining sample. This is repeated n times with different training sets of size (n-1).
- This approach has large computational requirements because n different models have to be designed and compared.



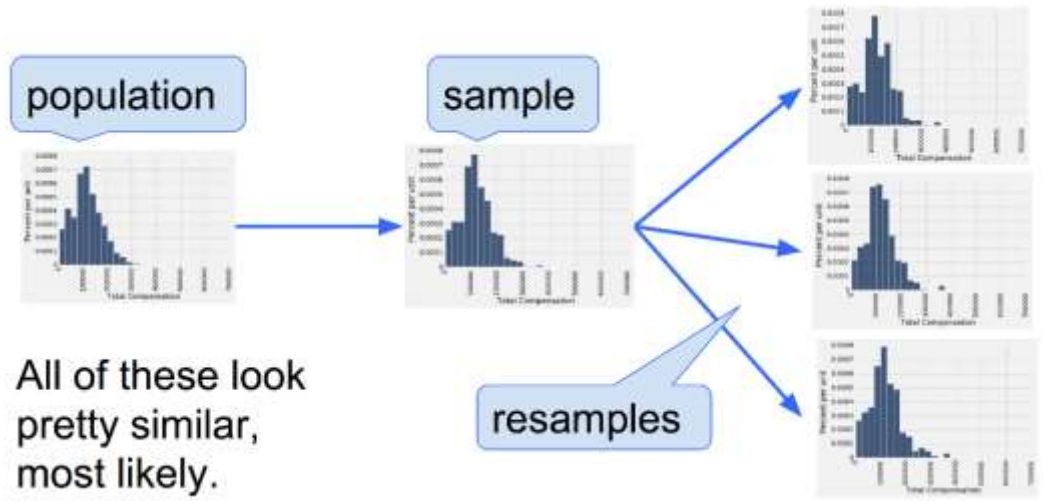
Data Splitting Methods:

- **4- k-Fold Cross Validation Method:**

- This approach is comprimise between holdout and leave-one-out methods. It divides the available samples into P disjoint subsets, where 1<= P<=n
- (P-1) substs are used for training and the remaining subset for testing
- This is the most popular method in practice, especially for problems where the number of samples is relatively small.



Data Splitting Methods:



- **5- Bootstrap method:**
- This method resamples the available data with replacements to generate a number of «fake» data sets of the same size as the given data set.
- These new training sets can be used to define so called bootstrap estimates of the error rate.

Model Estimation

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Problem with Accuracy

- Consider a 2-class problem
  - Number of Class NO examples = 990
  - Number of Class YES examples = 10

- If a model predicts everything to be class NO, accuracy is 990/1000 = 99 %
  - This is misleading because the model does not detect any class YES example
  - Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

Alternative Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Alternative Measures

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2*1*0.5}{1+0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	1	9
	Class=No	0	990

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F - measure (F)} = \frac{2*0.1*1}{1+0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	10	40

$$\text{Precision } (p) = 0.8$$
$$\text{Recall } (r) = 0.8$$
$$\text{F - measure } (F) = 0.8$$
$$\text{Accuracy } = 0.8$$

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	1000	4000

$$\text{Precision } (p) = \sim 0.04$$
$$\text{Recall } (r) = 0.8$$
$$\text{F - measure } (F) = \sim 0.08$$
$$\text{Accuracy } = \sim 0.8$$