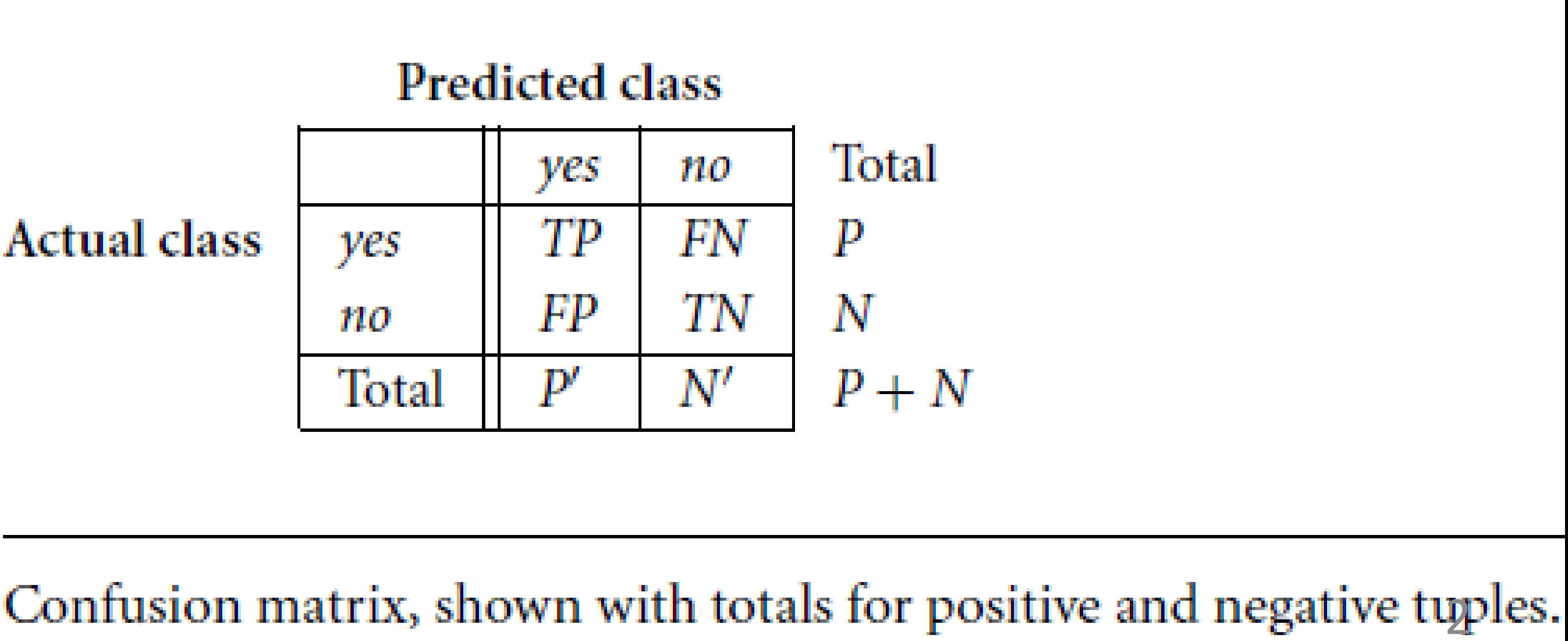
Module No : 03

Classification

Confusion Matrix

- The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes.
- TP and TN tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong.

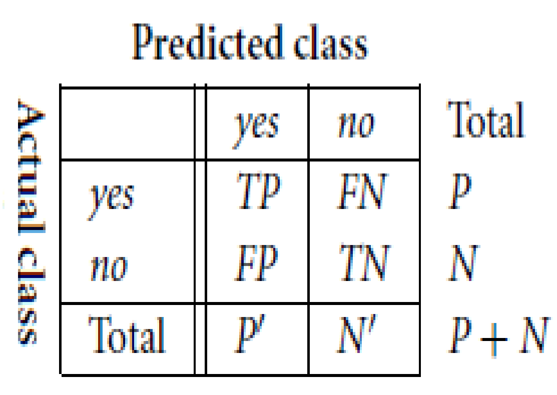


		Predicted class		
		yes	*no*	Total
Actual class	*yes*	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	$P + N$

Confusion matrix, shown with totals for positive and negative tuples.

Confusion Matrix

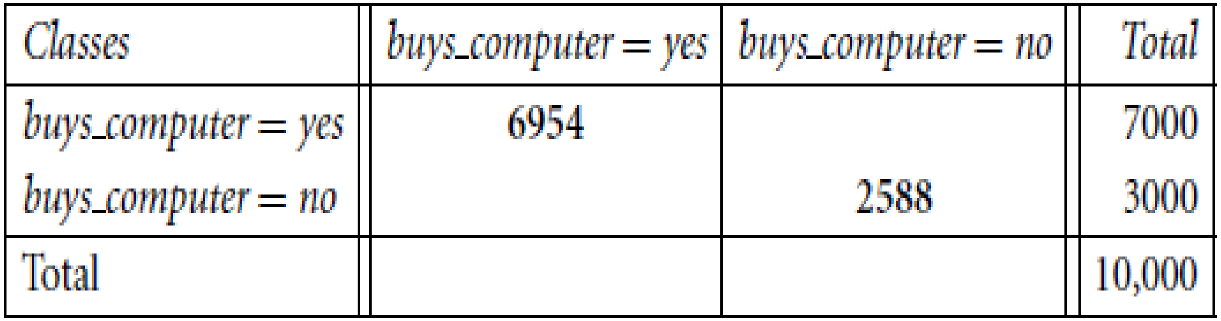
- E.g. suppose in a data set of the customers who buys the computer, there are total 10000 tuples, out of that 7000 are positive and 3000 are negative and our model has predicated 6954 are positive and 2588 are negative, so prepare confusion matrix



Predicted class

Actual class

	yes	*no*	Total
yes	*TP*	*FN*	*P*
no	*FP*	*TN*	*N*
Total	*P'*	*N'*	*P* + *N*

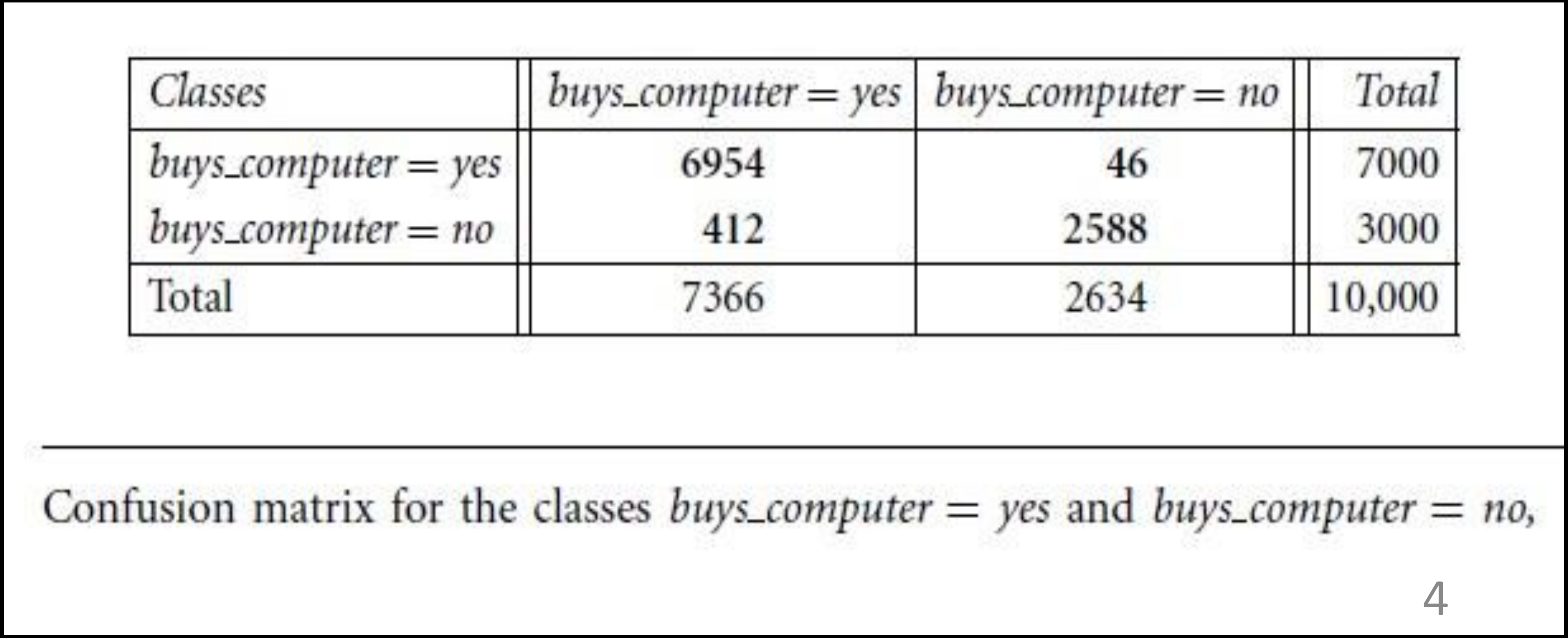


Classes	*buys_computer* = *yes*	*buys_computer* = *no*	*Total*
buys_computer = *yes*	6954		7000
buys_computer = *no*		2588	3000
Total			10,000

Confusion matrix for the classes *buys_computer* = *yes* and *buys_computer* = *no*,

Confusion Matrix

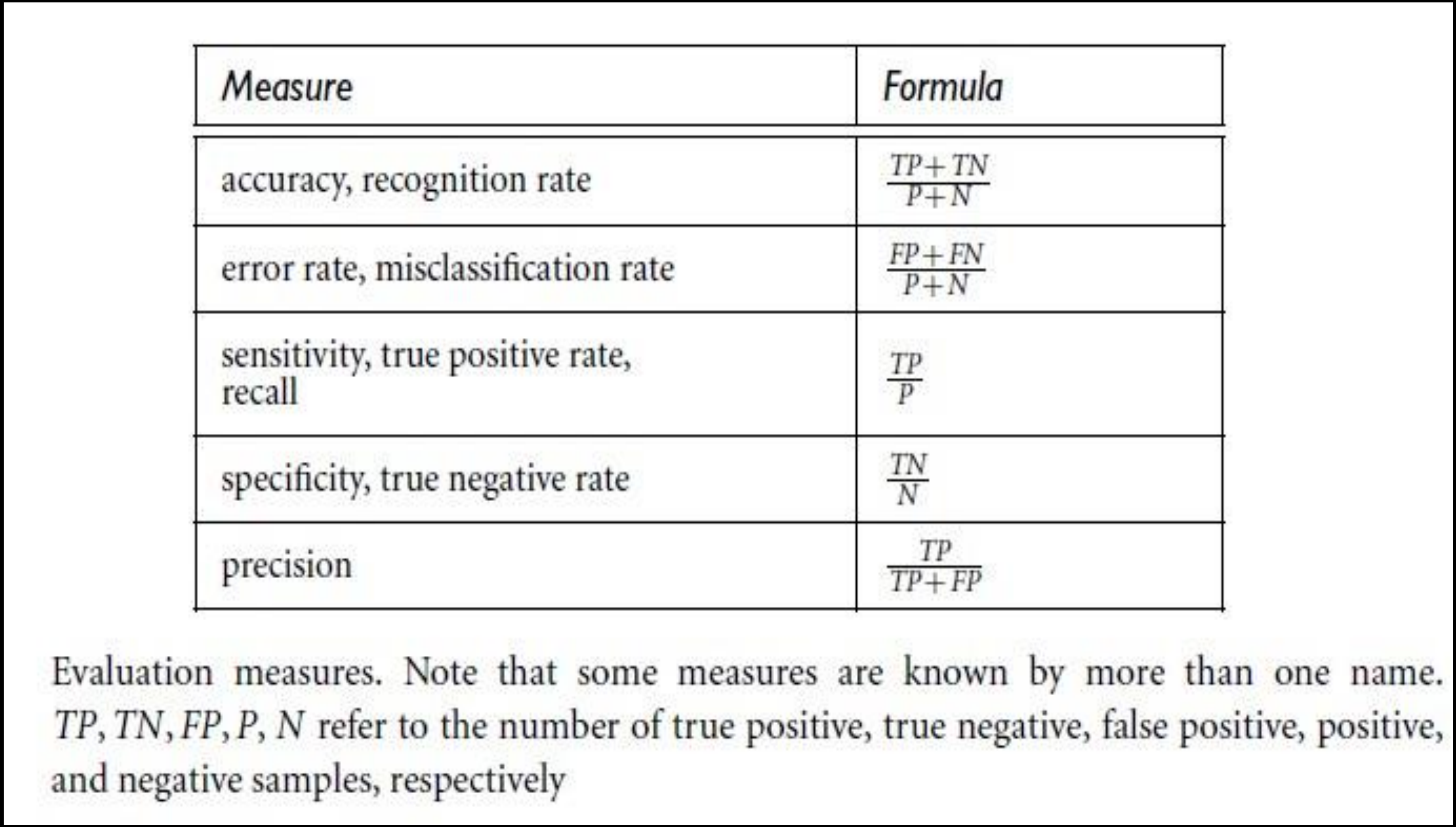
- E.g. suppose in a data set of the customers who buys the computer, there are total 10000 tuples, out of that 7000 are positive and 3000 are negative and our model has predicated 6954 are positive and 2588 are negative, so the confusion matrix will be



Classes	*buys_computer = yes*	*buys_computer = no*	*Total*
buys_computer = yes	6954	46	7000
buys_computer = no	412	2588	3000
Total	7366	2634	10,000

Confusion matrix for the classes *buys_computer* = *yes* and *buys_computer* = *no*,

Classifiers performance evaluation measures

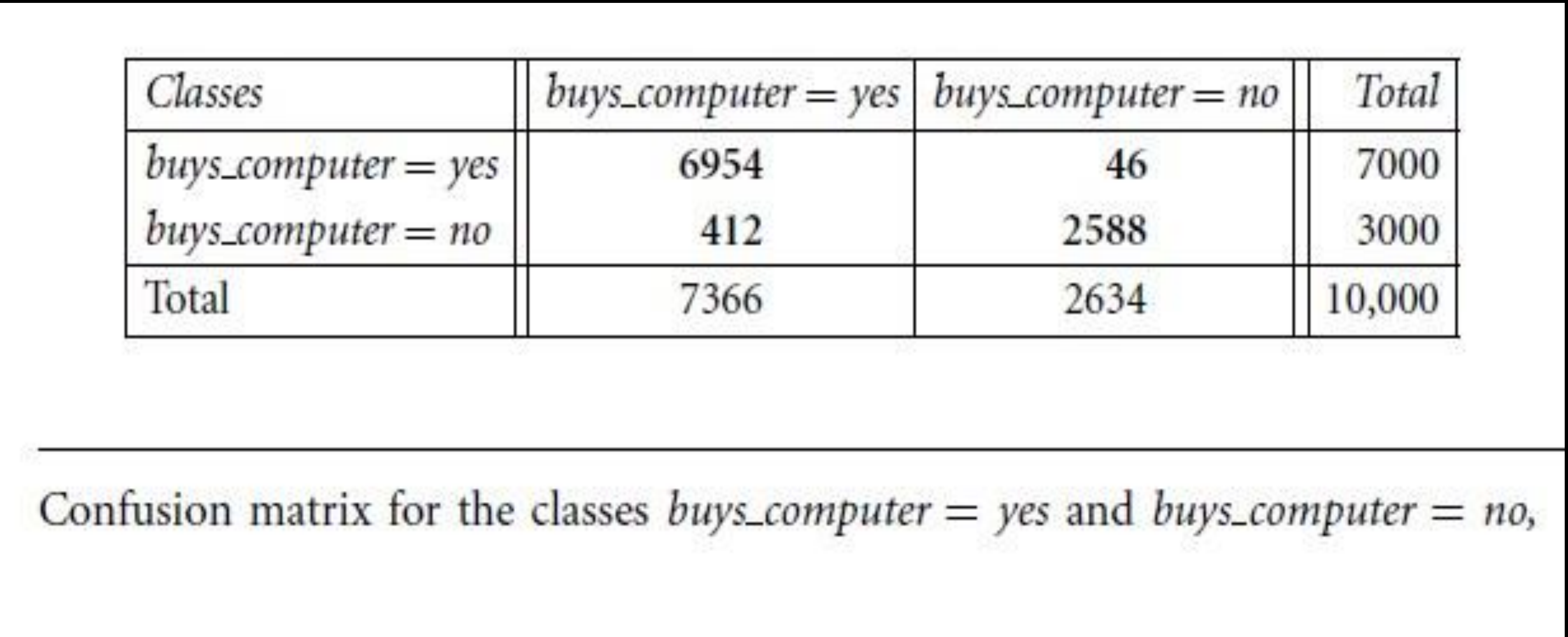


Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$

Evaluation measures. Note that some measures are known by more than one name. TP, TN, FP, P, N refer to the number of true positive, true negative, false positive, positive, and negative samples, respectively

Classifiers performance evaluation measures

- Find all evaluation measures for the following confusion matrix

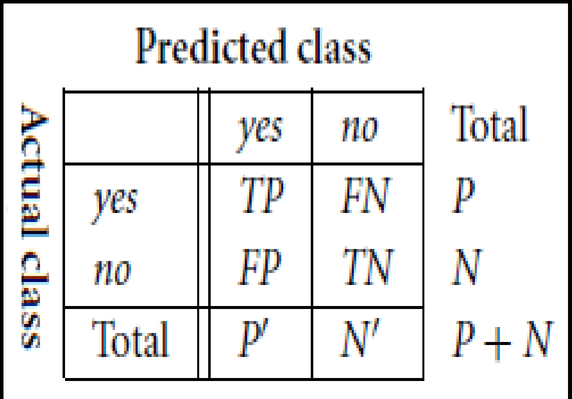


Classes	*buys_computer = yes*	*buys_computer = no*	*Total*
buys_computer = yes	6954	46	7000
buys_computer = no	412	2588	3000
Total	7366	2634	10,000

Confusion matrix for the classes *buys_computer = yes* and *buys_computer = no*,

Confusion Matrix

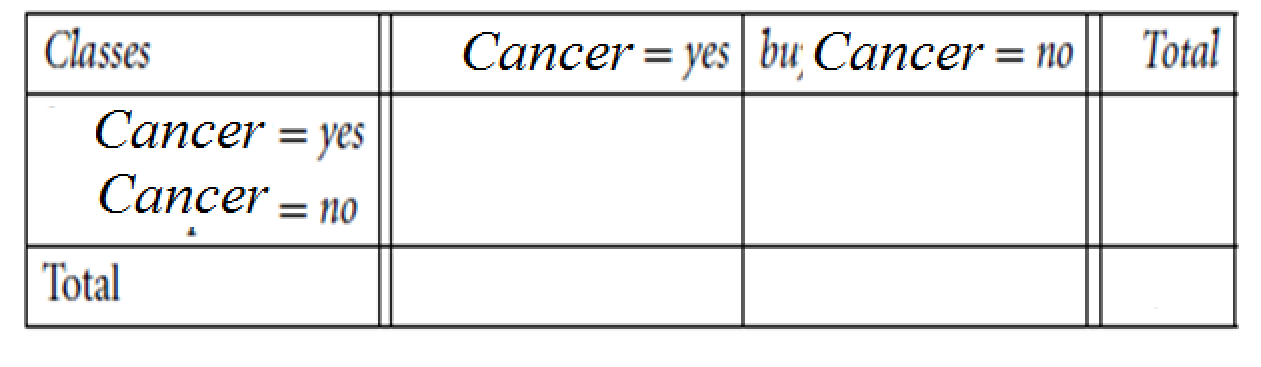
- E.g. suppose in a data set of the cancer, there are total 10000 tuples, out of that 300 are positive and 9700 are negative and our model has predicated 90 are positive and 9560 are negative, so prepare confusion matrix and Find all evaluation measures for the confusion matrix



Predicted class

Actual class

	yes	no	Total
yes	TP	FN	P
no	FP	TN	N
Total	P'	N'	$P+N$



Classes	$Cancer = yes$	bu; $Cancer = no$	Total
$Cancer = yes$ $Cancer = no$			
Total			

Confusion Matrix

Predicted

165	NO	Yes	
NO	50 [TN]	10 [FP]	60
Yes	5 [FN]	100 [TP]	105
	55	110	

ACTUAL

- Accuracy $= \dfrac{TP + TN}{Total}$

$= (100 + 50)/165$

$= \underline{0.91}$

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

or

$\dfrac{FP + FN}{Total}$

$= \underline{0.09}$

- Recall :- $\dfrac{TP}{\text{actual Yes}}$

$= \dfrac{100}{105} = \underline{0.95}$

- Precision $= \dfrac{TP}{\text{Predicted yes}}$

$= \dfrac{100}{110}$

$= \underline{0.64}$

- Evaluation measures for the confusion matrix

1. Accuracy:
2. Error rate:
3. Sensitivity: ability to correctly label the positive as positive
4. Specificity: ability to correctly label the negative as negative
5. Precision: % of positive tuples labelled as positive

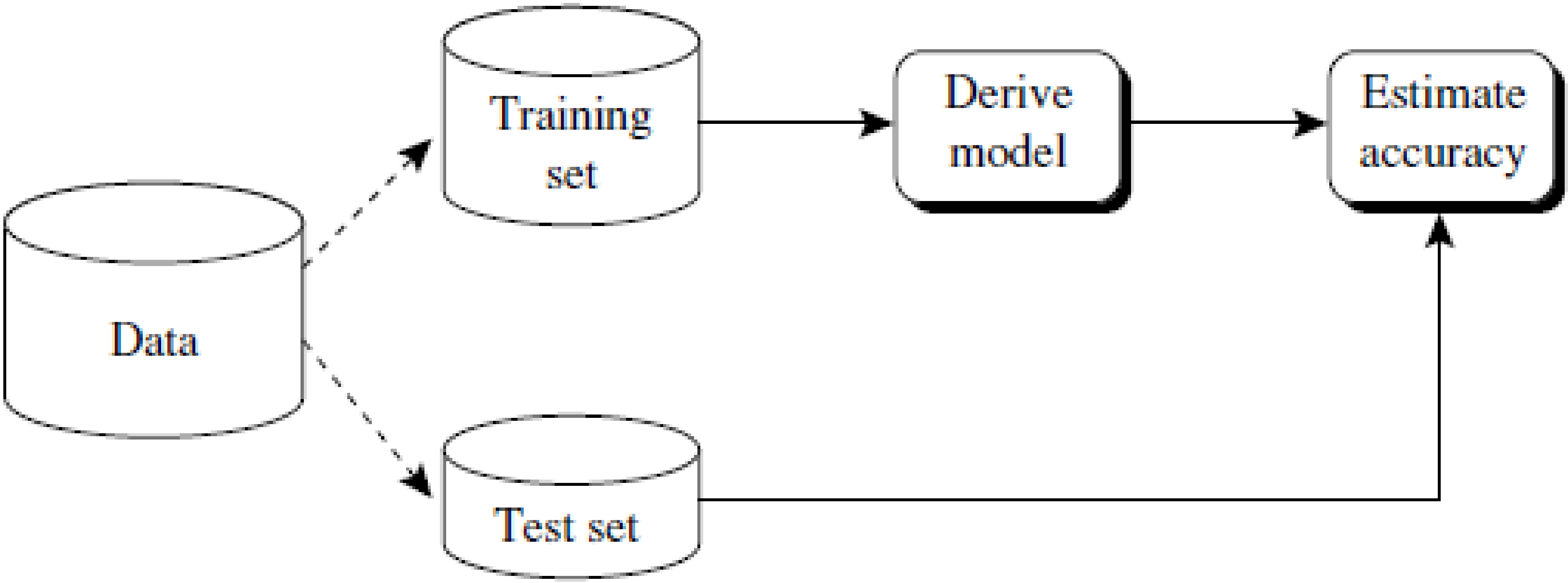
Model evaluation and selection methods

1. Holdout
2. Random sampling
3. Cross validation
4. Bootstrap
5. ROC Curves (Receiver operating characteristic curves)

Holdout

- In this method, the given data are randomly partitioned into two independent sets, a training set and a test set.
- Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set.
- The training set is used to derive the model. The model’s accuracy is then estimated with the test set.
- The estimate is pessimistic(negative) because only a portion of the initial data is used to derive the model.

Holdout



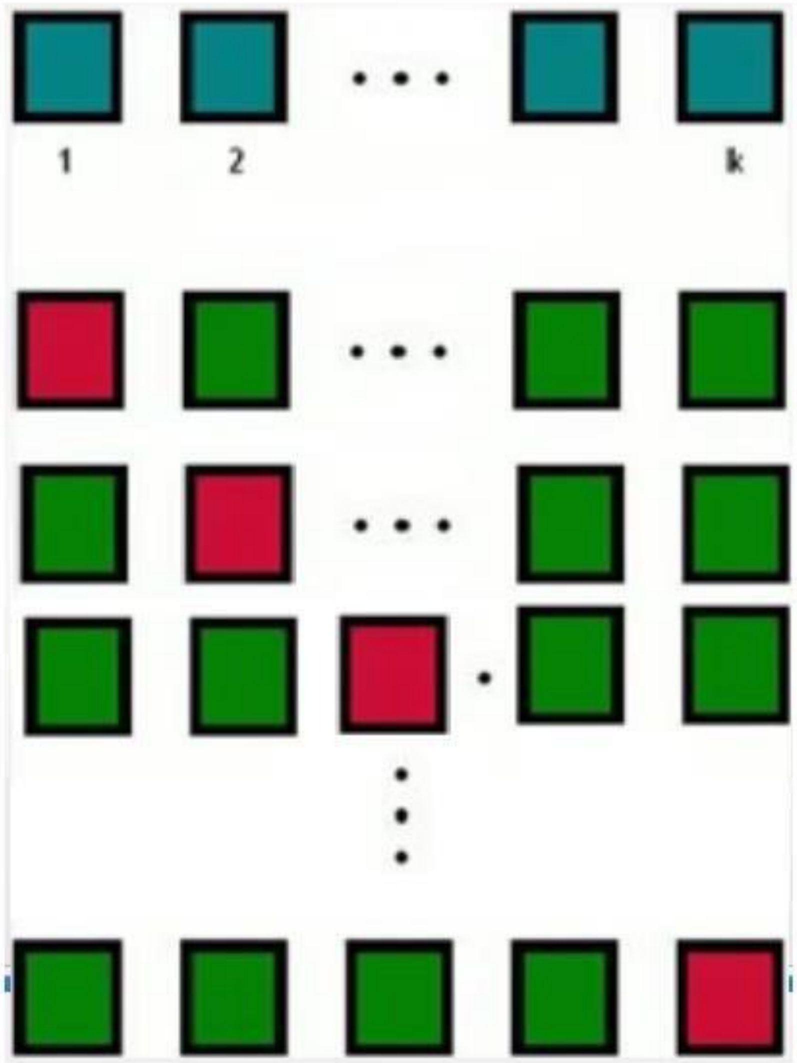
Estimating accuracy with the holdout method.

Random sub-sampling

- Random subsampling is a variation of the holdout method in which the holdout method is repeated k times.
- The overall accuracy estimate is taken as the average of the accuracies obtained from each iteration.

Cross-validation

- In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2, ... , Dk, each of approximately equal size.
- Training and testing is performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the model.
- That is, in the first iteration, subsets D2,....., Dk collectively serve as the training set to obtain a first model, which is tested on D1
- the second iteration is trained on subsets D1, D3, ... , Dk and tested on D2 and so on...
- Each fold is used the same number of times for training and once for testing
- the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data


1
2
k

Bootstrap Video

- Bootstrap randomly selects a tuple from the original data set
- Add that tuple into the training dataset and again send it back to the original dataset
- Repeat this process N times (N is the total number of tuples in the original dataset)
- The bootstrap is allowed to select the same tuple more than once.
- We use the training data set to train the model and test dataset to obtain an accuracy estimate of the model