- To train
use test

Scanned by CamScanner

11. **F measure (F_1 or F-score) :** Harmonic mean of precision and recall,

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

12. F_β : Weighted measure of precision and recall and assigns ß times as much weight to recall as to precision.

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

where β is a non-negative real number.

13. **Classifiers can also be compared with respect to :**
- Speed
- Robustness
- Scalability
- Interpretability

14. **Re-substitution error rate**
- Re-substitution error rate is a performance measure and is equivalent to training data error rate.
- It is difficult to get 0% error rate but it can be minimized, so low error rate is always preferable.

Syllabus Topic : Holdout

4.4.2 Holdout

- In holdout method, data is divided into training data set and testing data set (usually 1/3 for testing, 2/3 for training).

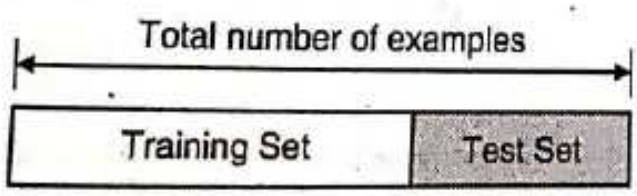

Fig. 4.4.1

To train the classifier, training data set is used and once the classifier is constructed then use test data set to estimate the error rate of the classifier.

Scanned by CamScanner

- If the training is more than better model is constructed and if the test data is more than more accurate the error estimates.
- **Problem :** The samples might not be representative. For example, some classes might be represented with very few instances or even with no instances at all.
- **Solution :** stratification is the method which ensures that both training and testing data have equal number of samples of same class.

Syllabus Topic : Random Sampling

4.4.3 Random Subsampling

- It is a variation of the holdout method.
- The holdout method is repeated k times.
- Each split randomly selects a fixed number example without replacement.

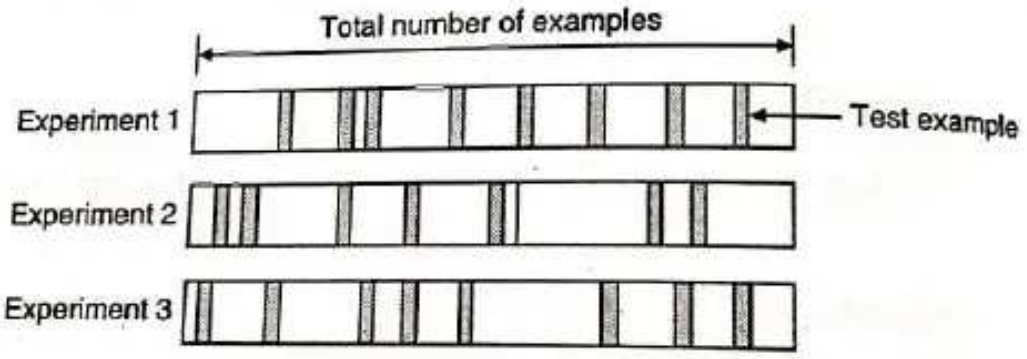

Fig. 4.4.2

- For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples.
- The overall accuracy is calculated by taking the average of the accuracies obtained from each iteration.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

Syllabus Topic : Cross-Validation

4.4.4 Cross-Validation (CV)

- Avoids overlapping test sets.

Scanned by CamScanner

k-fold cross-validation

- **First step :** Data is split into k subsets of equal size (usually by random sampling).

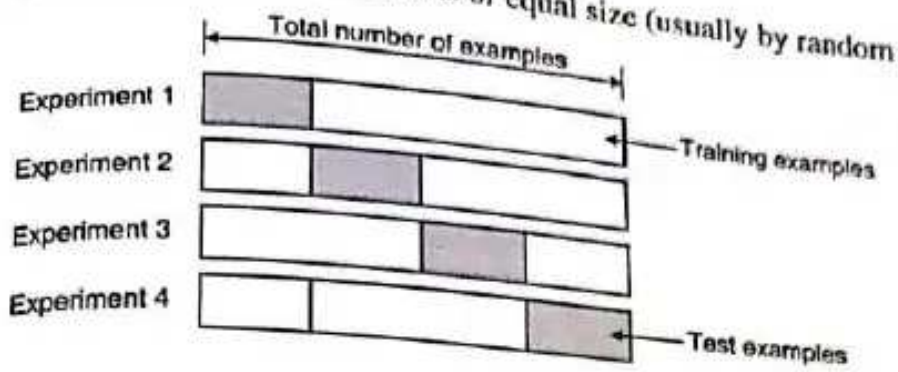

Fig. 4.4.3

- **Second step :** Each subset in turn is used for testing and the remainder for training.

- The advantage is that all the examples are used for both training and testing.
- The error estimates are averaged to yield an overall error estimate.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

- **Leave-one-out cross validation**
  - If dataset has N examples, then N experiments to be performed for Leave-one-out cross validation.
  - For every experiment, training uses N-1 examples and remaining example for testing.
- The average error rate on test examples gives the true error.

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

- **Stratified cross-validation:** Subsets are stratified before the cross-validation is performed.
- **Stratified ten-fold cross-validation**
  - This gives accurate estimate of evaluation.
  - The estimate's variance get reduced due to stratification.
  - Ten-fold cross-validation is repeated ten times and finally the results are averaged

based on the previous 10 results.

Scanned by CamScanner

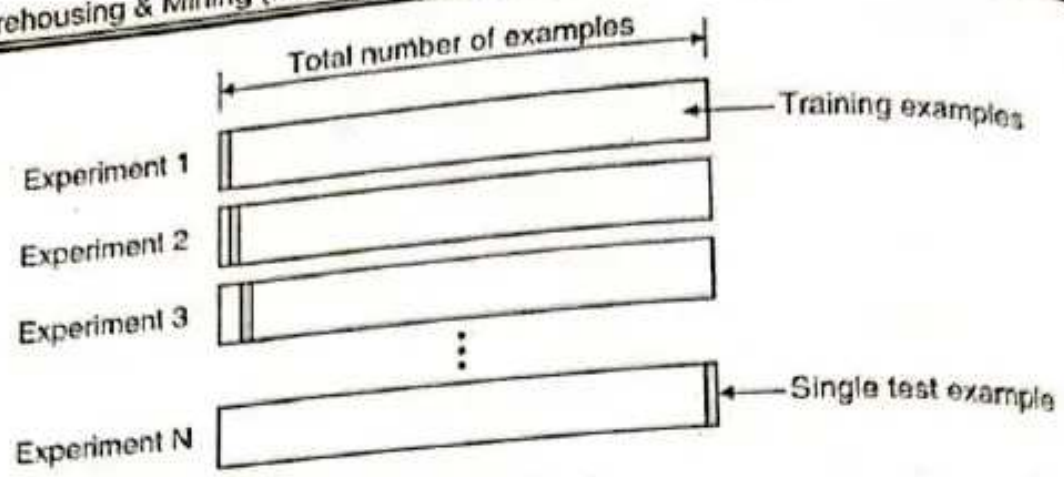

Fig. 4.4.4

Syllabus Topic : Bootstrap

4.4.5 Bootstrapping

- CV uses sampling of data set without replacement. Once the tuple or instance is selected it cannot be selected again for training or test data.
- The bootstrap uses sampling with replacement to get the training set.
- **Training set :** A dataset of k instances is sampled with replacement k times to form the training set of k instances.
- **Test set :** This is separate dataset from the original dataset which is not the part of training dataset.
- Bootstrapping is the best error estimator for small datasets.

Syllabus Topic : Clustering

4.5 What is Clustering ?

4.5.1 What is Clustering ?

→ **(MU - Dec. 2010, May 2012, Dec. 2012, Dec. 2013)**

- Clustering is an unsupervised learning problem.

A cluster contains da
We can show this w

- From Fig. 4.5 Geometrical di belong to whi clustering.
- The other ki cluster are a descriptive c

Applications

Clustering a

- **Marketing** database, c be identifi
- **Biology** classes ba
- **Librarie** ordering
- **Insuran** identifi
- **City**