EDU ENGINEERING

PIONEER OF ENGINEERING NOTES

TAMIL NADU'S BEST EDTECH PLATFORM FOR ENGINEERING

CONNECT WITH US

WEBSITE: www.eduengineering.net

TELEGRAM: @eduengineering

INSTAGRAM: @eduengineering

- **Regular Updates for all Semesters**
- **All Department** Notes AVAILABLE
- **Handwritten** Notes AVAILABLE
- **Past Year** Question Papers AVAILABLE
- **Subject wise** Question Banks AVAILABLE
- **Important Questions** for Semesters AVAILABLE
- Various **Author Books** AVAILABLE

5.1 Machine Learning Life Cycle

- The Machine Leaning (ML) model management and the delivery of highly performing model is as important as the initial build of the model by choosing right dataset. The concepts around model retraining, model versioning, model deployment and model monitoring are the basis for machine learning operations that helps the data science teams deliver highly performing models.
- The use of machine leaning has increased substantially in enterprise data analytics scenarios to extract valuable insights from the business data. Hence, it is very important to have an ecosystem to build, test, deploy and maintain the enterprise grade machine learning models in production environments.
- The ML model development involves data acquisition from multiple trusted sources, data processing to make suitable for building the model, choose algorithm to build the model, build model, compute performance metrics and choose best performing model.
- The model maintenance plays critical role once the model is deployed into production. The maintenance of machine learning model includes keeping the model up to date and relevant in tune with the source data changes as there is a risk of model becoming outdated in course of time.
- Machine learning model lifecycle refers to the process that covers right from source data identification to model development, model deployment and model maintenance. At high level, the entire activities fall under two broad categories, such as ML model development and ML model operations.
- The machine learning lifecycle process is shows in Fig. 5.1.1 and it includes the following phases :
  1. Business goal identification
  2. ML problem framing
  3. Data processing (Data collection, data preprocessing, feature engineering)
  4. Model development (Training, tuning, evaluation)
  5. Model deployment (Inference, prediction)
  6. Model monitoring.
- **Business goal :** An organization considering ML should have a clear idea of the problem and the business value to be gained by solving that problem. We must be able to measure business value against specific business objectives and success criteria.
- **ML problem framing :** In this phase, the business problem is framed as a machine learning problem : What is observed and what should be predicted (known as a

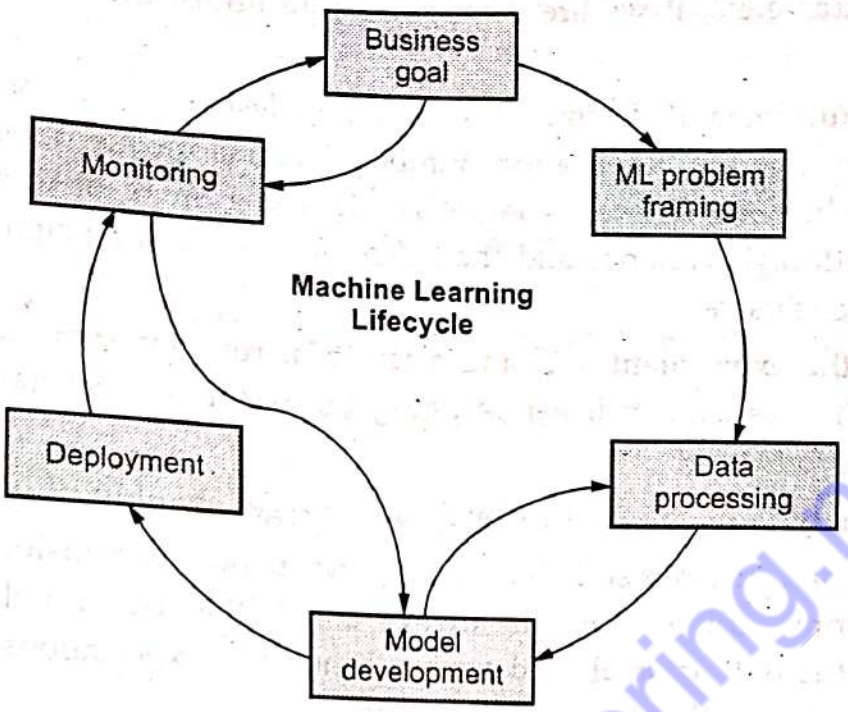


Fig. 5.1.1 Machine learning lifecycle process

label or target variable). Determining what to predict and how performance and error metrics must be optimized is a key step in this phase.

- **Data processing :** Training an accurate ML model requires data processing to convert data into a usable format. Data processing steps include collecting data, preparing data and feature engineering that is the process of creating, transforming, extracting, and selecting variables from data.
- **Model development :** Model development consists of model building, training, tuning and evaluation. Model building includes creating a pipeline that automates the build, train and release to staging and production environments.
- **Deployment :** After a model is trained, tuned, evaluated and validated, we can deploy the model into production. we can then make predictions and inferences against the model.
- **Monitoring :** Model monitoring system ensures your model is maintaining a desired level of performance through early detection and mitigation.

5.2 Guidelines for Machine Learning Experiments

- **Aim of the study :** What are the objectives (e.g. assessing the expected error of an algorithm, comparing two learning algorithm on a particular problem, etc.).
- **Selection of the response variable :** what should we use as the quality measure (e.g. error, precision and recall, complexity, etc.)
- **Choice of factors and levels :** What are the factors for the defined aim of the study (factors are hyperparameters when the algorithm is fix and want to find

best hyperparameters, if we are comparing algorithms, the learning algorithm is a factor).

- **Choice of experimental design :** Use factorial design unless we are sure that the factors do not interact. Replication number depends on the dataset size; it can be kept small when the dataset is large. Avoid using small datasets which leads to responses with high variance and the differences will not be significant and results will not be conclusive.
- **Performing the experiment :** Doing a few trial runs for some random settings to check that all is as expected, before doing the factorial experiment. All the results should be reproducible.
- **Statistical analysis of the data :** Conclusion we get should not be due to chance.
- Conclusions and recommendations : One frequently conclusion is the need for further experimentation. There is always a risk that our conclusions be wrong, especially if the data is small and noisy. When our expectations are not met, it is most helpful to investigate why they are not.

5.2.1 Dataset Preparation

- Machine learning is about learning some properties of a data set and applying them to new data. This is why a common practice in machine learning to evaluate an algorithm is to split the data at hand in two sets, one that we call a training set on which we learn data properties and one that we call a testing set, on which we test these properties.
- In training data, data are assign the labels. In test data, data labels are unknown but not given. The training data consist of a set of training examples.
- The real aim of supervised learning is to do well on test data that is not known during learning. Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy.
- The training error is the mean error over the training sample. The test error is the expected prediction error over an independent test sample.
- Problem is that training error is not a good estimator for test error. Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to over fitting and poor generalization.
- Training set : A set of examples used for learning, where the target value is known.
- Test set : It is used only to assess the performances of a classifier. It is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.
- Training data is the knowledge about the data source which we use to construct the classifier.

- In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a dataset is divided into a training set, a validation set (some people use 'test set' instead) in each iteration or divided into a training set, a validation set and a test set in each iteration.
- In machine learning, we basically try to create a model to predict the test data. So, we use the training data to fit the model and testing data to test it. The models generated are to predict the results unknown which is named as the test set.

5.3 Cross Validation (CV) and Resampling

- Validation techniques in machine learning are used to get the error rate of the ML model, which can be considered as close to the true error rate of the population. If the data volume is large enough to be representative of the population, you may not need the validation techniques.
- In machine learning, model validation is referred to as the process where a trained model is evaluated with a testing data set. The testing data set is a separate portion of the same data set from which the training set is derived. The main purpose of using the testing data set is to test the generalization ability of a trained model.
- Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern.
- In general, ML involves deriving models from data, with the aim of achieving some kind of desired behavior, e.g., prediction or classification.
- But this generic task is broken down into a number of special cases. When training is done, the data that was removed can be used to test the performance of the learned model on ``new'' data. This is the basic idea for a whole class of model evaluation methods called **cross validation.**
- Types of cross validation methods are holdout, K - fold and leave-one-out.
- The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximate fits a function using the training set only.
- K - fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed.

- Leave-one-out cross validation is K - fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximate is trained on all the data except for one point and a prediction is made for that point.

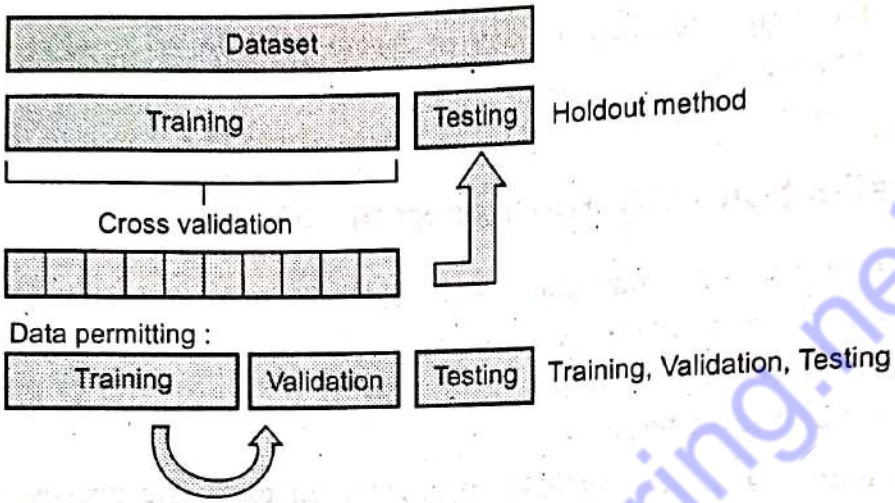


Fig. 5.3.1

5.3.1 K - Fold Cross Validation

- K - fold CV is where a given data set is split into a K number of sections/folds where each fold is used as testing set at some point.
- Lets take the scenario of 5-Fold cross validation (K = 5). Here, the data set is split into 5 folds.
- In the first interaction, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds has been used as the testing set.
- K - fold cross validation is performed as per the following steps :

1. Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as $f_1, f_2, \ldots f_k$.
2. For i = 1 to i = k

- Keep the fold f_i as validation set and keep all the remaining k – 1 folds in the cross validation training set.

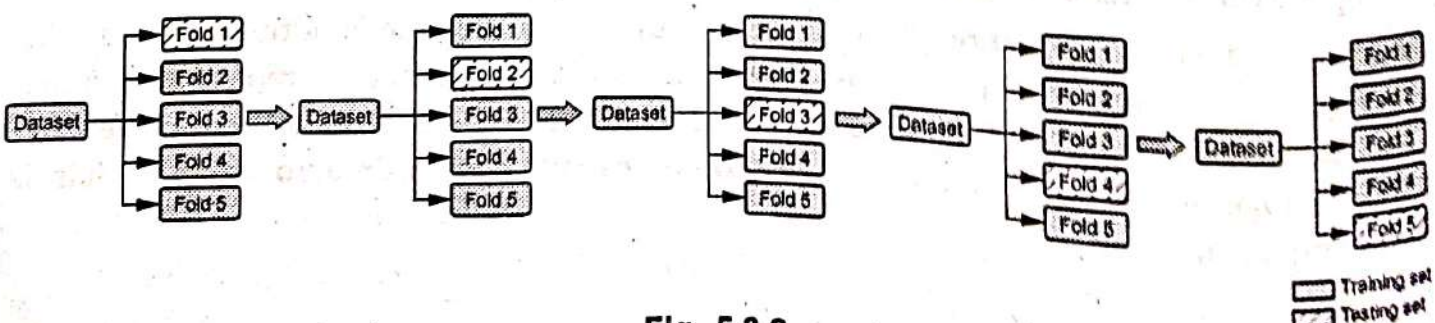


Fig. 5.3.2

3. Estimate the accuracy of your machine learning model by averaging the accuracies derived in all the k cases of cross validation.

- In the k - fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.
- The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once and gets to be in a training set $k-1$ times. The variance of the resulting estimating is reduced as k is increased.
- The disadvantage of this method is that the training algorithm has to be rerun scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.
- The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

5.3.2 Bootstrapping

- Bootstrapping is a method of sample reuse that is much more general than cross-validation. The idea is to use the observed sample to estimate the population distribution. Then samples can be drawn from the estimated population and the sampling distribution of any type of estimator can itself be estimated.
- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.
- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y respectively, where X and Y are random quantities. We will invest a fraction α of our money in X and will invest the remaining $1 - \alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize Var $(\alpha X + (1 - \alpha)Y)$.
- One can show that the value that minimizes the risk is given by,

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

where σ_X^2 = Var(X), σ_Y^2 = Var(Y) and σ_{XY} = Cov(X,Y)

- But the values of σ_X^2, σ_Y^2 and σ_{XY} are unknown.

- We can compute estimates for these quantities, σ_X^2, σ_Y^2 and $\hat{\sigma}_{XY}$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using,

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y and estimating α 1,000 times.
- We thereby obtained 1,000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.
- For these simulations the parameters were set to $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$ and $\sigma_{XY} = 0.5$ and so we know that the true value of α is 0.6.
- The mean over all 1,000 estimates for α is,

$$\hat{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r = 0.5996,$$

very close to $\alpha = 0.6$ and the standard deviation of the estimates is,

$$\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \hat{\alpha})^2} = 0.083$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $SE(\hat{\alpha}) \approx 0.083$.
- So roughly speaking, for a random sample from the population, we would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.
- There are three forms of bootstrapping which differ primarily in how the population is estimated.
  1. Nonparametric (Resampling)
  2. Semiparametric (Adding noise)
  3. Parametric. (Simulation)

1. **Nonparametric bootstrap :** In the nonparametric bootstrap a sample of the same size as the data is taken from the data with replacement. If we measure 10 samples, we create a new sample of size 10 by replicating some of the samples that we have already seen and omitting others.

2. **Semiparametric bootstrap :** The resampling bootstrap can only reproduce the items that were in the original sample. The semiparametric bootstrap assumes that the population includes other items that are similar to the observed sample by sampling from a smoothed version of the sample histogram. It turns out that this can be done very simply by first taking a sample with replacement from the observed sample and then adding noise.

3. **Parametric bootstrap :** Parametric bootstrapping assumes that the data comes from a known distribution with unknown parameters. We estimate the parameters from the data that you have and then you use the estimated distributions to simulate the samples.

5.4 Measuring Classifier Performance

- A binary classification rule is a method that assigns a class to an object, on the basis of its description.
- The performance of a binary classifier can be assessed by tabulating its predictions on a test set with known labels in contingency table or confusion matrix, with actual classes in rows and predicted classes in columns.
- Measures of performance need to satisfy several criteria :
  1. They must coherently capture the aspect of performance of interest;
  2. They must be intutive enough to become widely used, so that the same measures are consistently reported by researches, enabling community-wide conclusions to be drawn;
  3. They must be computationally tractable, to match the rapid growth in scale of modem data collection.
  4. They must be simple to report as a single number for each method-dataset combination.
- Performance metrics for binary classification are designed to captured tradeoffs between four fundamental population quantities : True positives, false positives, true negatives and false negatives.
- The evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix. The confusion matrix for a binary classification problem is shown below.

True class	Predicted class	
	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

- A confusion matrix contains about actual and predicted classifications done by a classification system. Performance of such systems is commonly using data in the matrix. Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.

2. **False negatives :** Examples predicted as negative, whose true class is positive.
3. **True poisitives :** Examples correctly predicted as pertaining to the positive class.
4. **True negatives :** Examples correctly predicted as belongings to the negative class.

- The evaluation measure most used in practice is the accuracy rate. It evaluates the effectiveness of the classifier by its percentage of correct predictions.

$$\text{Accuracy rate} = \frac{|\text{True negatives}| + |\text{True positives}|}{|\text{False negatives}| + \text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

- The complement of accuracy rate is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$\text{Error rate} = \frac{|\text{False negatives}| + |\text{False positives}|}{|\text{False negatives}| + |\text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

Error rate = 1 – (Accuracy rate)

- The recall and specificity measures evaluate the effectiveness of a classifier for each class in the binary problem. The recall is also known as sensitivity or true positive rate. Recall is the proportion of examples belonging to the positive class which were correctly predicted as positive.
- The specificity is a statistical measures of how well a binary classification test correctly identifies the negative cases.

$$\text{Recall (R)} = \frac{|\text{True positive}|}{||\text{True positive}| + |\text{False negative}|}$$

$$\text{Specificity} = \frac{|\text{True negative}|}{|\text{False positive}| + |\text{True positive}|}$$

- True positive Rate (TPR) is also called sensitivity, hit rate and recall.

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positive} + \text{Number of false positive}}$$

- A statistical measure of how well a binary classification test correctly identifies a condition. Probability of correctly labeling members of the target class.
- No single measures tells the whole story. A classifier with 90 % accuracy can be useless if 90 percent of the population does not have cancer and the 10 % that do are misclassified by the classifier. Use of multiple measures recommended.

5.4.1 Accuracy and ROC Curves

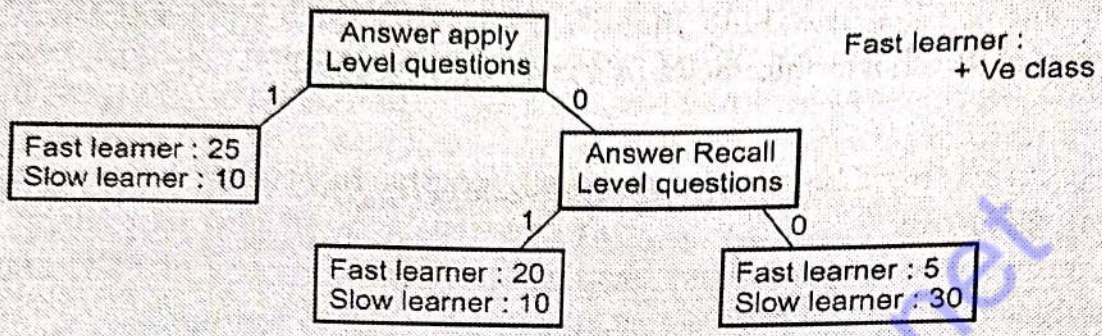
- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predicative model.

- Accuracy (ACC) measures the fraction of correct predictions. Precision measures the fraction of actual positives among those examples that are predicted as positive. Recall measures how many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall.

ROC Curve

- Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm fates over noisy channel. Recent years have seen an increase in the use of ROC graphs in the machine learning community.
- An ROC plot plots true positive rate on the Y-axis false positive rate on the X-axis; a single contingency table corresponds to a single point in an ROC plot.
- The performance of a ranker can be assessed by drawing a piecewise linear curve in an ROC plot, known as an ROC curve. The curve starts in (0, 0), finishes in (1, 1) and is monitorically non-decreasing in both axes.
- A useful technique for organizing classifiers and visualizing their performance. Especially useful for domains with skewed class distribution and unequal classification error costs.
- It allows to create ROC curve and a complete sensitivity/specificity report. The ROC curve is a fundamental tool for diagnostic test evaluation.
- In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100 Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter can distinguish between two diagnostic groups.
- Each point on an ROC curve connecting two segments corresponds to the true and false positive rates achieved on the same test set by the classifier obtained from the ranker by splitting the ranking between those two segments.
- An ROC curve is convex if the slopes are montonically non-increasing when moving along the curve from (0, 0) to (1, 1). A concavity in an ROC curve, i.e., two or more adjacent segments with increasing slopes, indicates a locally worse than random ranking. In this we would get better ranking performance by joining the segments involved in the concavity, thus creating a coarser classifier.

Example 5.4.1 *i) Find contingency table ii) Find recall iii) Precision iv) Negative recall v) False positive rate*



Solution : Contingency table

	Predicted			Total
Faster Learner	25	20	5	50
Slow Learner	10	10	30	50
Total	35	30	35	100

Actual

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Calculate precision and recall

Precision = 25/35 = 0.714

Recall = 25/30 = 0.833

False positive rate = (False positive) / (false positive + true negative)

= 10/(10 + 30) = 0.25

Example 5.4.2 *Consider following confusion matrix and calculate following i) Sensitivity of classifier ii) Specificity of classifier.*

Confusion Matrix		Predicted		Total
Actual	+	8	10	18
	–	4	8	12
Total		12	18	30

Solution : Given data : TP = 8, FN = 10, FP = 4, TN = 8

- Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR)

$$\text{Sensitivity (SN)} = \frac{TP}{TP+FN} = \frac{8}{8+10} = 0.444$$

- Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

$$\text{Specificity (SP)} = \frac{TN}{TN+FP} = \frac{8}{8+4} = 0.666$$

Example 5.4.3 *Consider the following 3-class confusion matrix. Calculate precision and recall per class. Also calculate weighted average precision and recall for classifier.*

		Predicted		
Actual	15	2	3
	7	15	8
	2	3	45

Solution :

		Predicted		
Actual	15	2	3	20
	7	15	8	30
	2	3	45	50
	24	20	56	100

$$\text{Classifier Accuracy} = \frac{15+15+45}{100} = \frac{75}{100} = 0.75$$

Calculate per-class precision and recall :

$$\text{First class} = \frac{15}{24} = 0.63 \quad \text{and} \quad \frac{15}{20} = 0.75$$

$$\text{Second class} = \frac{15}{20} = 0.75 \quad \text{and} \quad \frac{15}{30} = 0.50$$

$$\text{Third class} = \frac{45}{56} = 0.8 \quad \text{and} \quad \frac{45}{50} = 0.9$$

Example 5.4.4 *Prove that : i) FPR = 1 – TPR ii) FNR = 1 – TPR*

Solution : i) FPR = 1–TPR

False Positive Rate (FPR) = 1 – True Negative Rate (TNR)

FPR = FP/N = FP/ (FP + TN)

FPR = 1 – TNR

ii) FNR = 1 – TPR

False Negative Rate = 1 – True Positive Rate

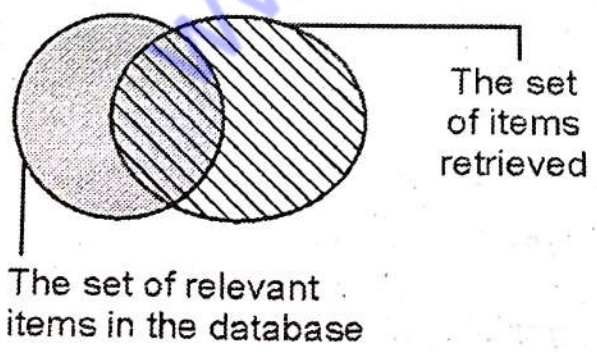
FNR = FN (FN + TP)

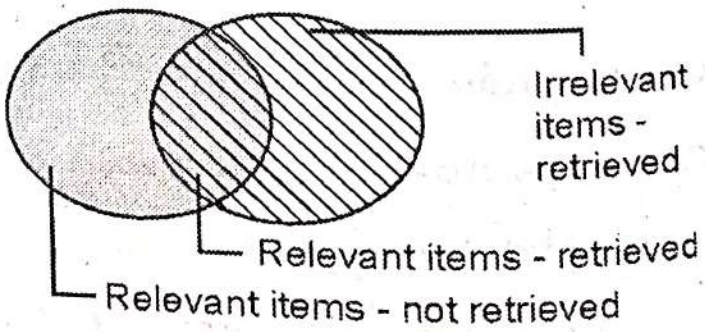
FNR = 1 – TPR

5.4.2 Precision and Recall

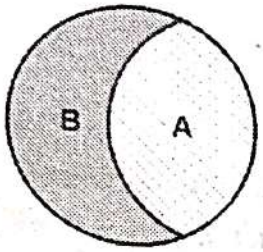
- **Relevance :** Relevance is a subjective notion. Different users may differ about the relevance or non-relevance of particular documents to given questions.
- In response to a query, an IR system searches its document collection and returns a ordered list of responses. It is called the retrieved set or ranked list. The system employs a search strategy or algorithm and measure the quality of a ranked list.
- A better search strategy yields a better ranked list and better ranked lists help the user fill their information need.
- Precision and recall are the basic measures used in evaluating search strategies. As shown in the first two figures, these measures assume :

1. There is a set of records in the database which is relevant to the search topic
2. Records are assumed to be either relevant or irrelevant.
3. The actual retrieval set may not perfectly match the set of relevant records.





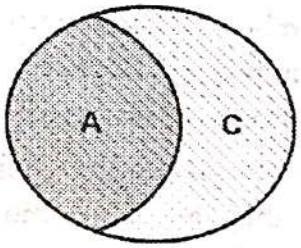
- **Recall** is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
B = Number of relevant records not retrieved.

$$\text{Recall} = \frac{A}{A+B} \times 100\ \%$$

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
C = Number of irrelevant records retrieved.

$$\text{Precison} = \frac{A}{A+C} \times 100\ \%$$

- As recall increases, the precision decreases and recall decreases the precision increases.

Example 5.4.5 *Assume the following :*
A database contains 80 records on a particular topic
A search was conducted on that topic and 60 records were retrieved.
Of the 60 records retrieved, 45 were relevant.
Calculate the precision and recall scores for the search.

Solution : Using the designations above :

A = The number of relevant records retrieved,

B = The number of relevant records not retrieved, and

C = The number of irrelevant records retrieved.

In this example A = 45, B = 35 (80 – 45) and C = 15 (60 – 45).

$$\text{Recall} = \frac{45}{45+35} \times 100\ \%$$

$$\text{Recall} = \frac{45}{80} \times 100\ \%$$

$$\textbf{Recall} = \textbf{56.25 \%}$$

$$\text{Precision} = \frac{A}{A+C} \times 100\ \%$$

$$\text{Precision} = \frac{45}{45+15} \times 100\ \% = \frac{45}{60} \times 100$$

Precision = 75 %

Example 5.4.6 *20 found documents, 18 relevant, 3 relevant documents are not found, 27 irrelevant are as well not found. Calculate the precision and recall and fallout scores for the search.*

Solution : Precision : 18/20 = 90 %

Recall : 18/21 = 85.7 %

Fall-out : 2/29 = 6.9 %

- Recall is a non-decreasing function of the number of docs retrieved. In a good system, precision decreases as either the number of docs retrieved or recall increases. This is not a theorem, but a result with strong empirical confirmation.
- The set of ordered pairs makes up the precision-recall graph. Geometrically when the points have been joined up in some way they make up the precision-recall curve. The performance of each request is usually given by a precision-recall curve. To measure the overall performance of a system, the set of curves, one for each request, is combined in some way to produce an average curve.
- Assume that set R_q containing the relevant document for q has been defined. Without loss of generality, assume further that the set R_q is composed of the following documents :

$$R_q = \{ d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \}$$

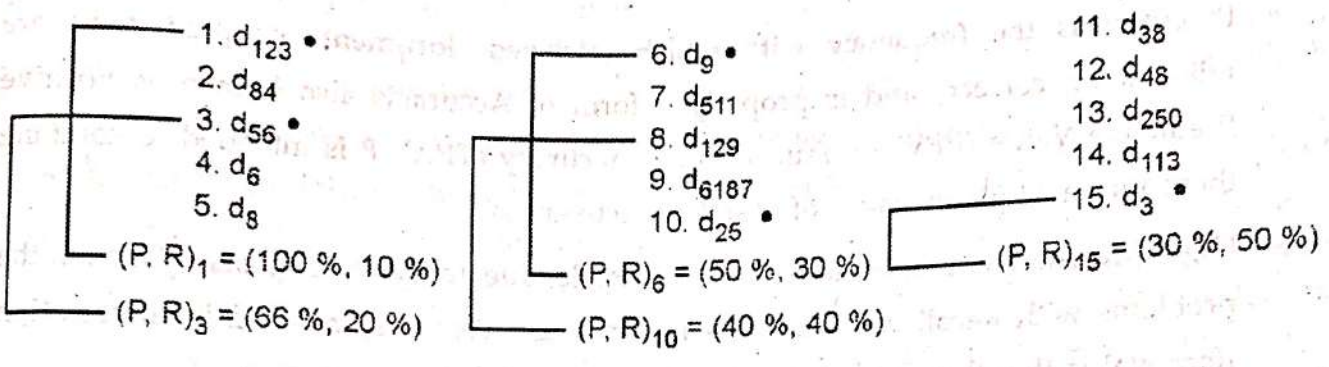
There are ten documents which are relevant to the query q.

- For the query q, a ranking of the documents in the answer set as follows.

Ranking for query q :

1. d_{123} *	6. d_9 *	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} *	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} *	15. d_3 *

- The documents that are relevant to the query q are marked with star after the document number. Ten relevant documents, five included in Top 15 .



- Fig 5.4.1 shows the curve of precision versus recall. By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a precision-recall curve.

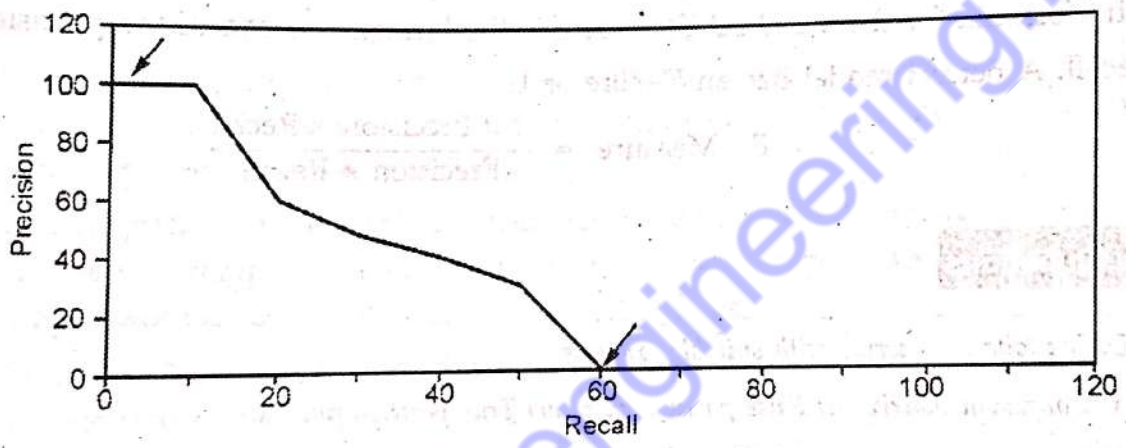


Fig. 5.4.1 Precision versus recall curve

- The precision versus recall curve is usually plotted based on 11 standard recall level: 0 %,10 %,....,100 %.
- In this example : The precisions for recall levels higher than 50 % drop to 0 because no relevant documents were retrieved. There was an interpolation for the recall level 0 %.
- Since the recall levels for each query might be distinct from the 11 standard recall levels.

5.4.3 F - Measure

- The F measure is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. The F - measure or F - score is one of the most commonly used "single number" measures in Information Retrieval, Natural Language Processing and Machine Learning.
- F-measure comes from Information Retrieval (IR) where Recall is the frequency with which relevant documents are retrieved or 'recalled' by a system, but it is known elsewhere as Sensitivity or True Positive Rate (TPR).

- Precision is the frequency with which retrieved documents or predictions are relevant or 'correct', and is properly a form of Accuracy, also known as Positive Predictive Value (PPV) or True Positive Accuracy (TPA). F is intended to combine these into a single measure of search 'effectiveness'.
- High precision and low accuracy is possible due to systematic bias. One of the problems with Recall, Precision, F - measure and Accuracy as used in Information Retrieval is that they are easily biased.
- The F-measure balances the precision and recall. The result is a value between 0.0 for the worst F-measure and 1.0 for a perfect F - measure.
- The formula for the standard F1 - score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1.

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

Review Questions

1. *Define following terms with suitable example :*
   i) Confusion matrix ii) False positive rate iii) True positive rate.
2. *What is a contingency table/matrix ? What is the use of it ?*
3. *Explain true positive, true negative false positives, false negatives and class ratio.*
4. *What is a contingency table ? What does it represent ?*
5. *What is multiple linear regression ? How will it be different from simple linear regression ?*

5.5 Multiclass Classification

- Multiclass classification is a machine learning classification task that consists of more than two classes, or outputs. For example, using a model to identify animal types in images from an encyclopedia is a multiclass classification example because there are many different animal classifications that each image can be classified as. Multiclass classification also requires that a sample only have one class.
- Each training point belongs to one of N different classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs.

- There are many scenarios in which there are multiple categories to which points belong, but a given point can belong to multiple categories. In its most basic form, this problem decomposes trivially into a set of unlinked binary problems, which can be solved naturally using techniques for binary classification.
- Common model for classification is the Support Vector Machine (SVM). An SVM works by projecting the data into a higher dimensional space and separating it into different classes by using a single (or set of) hyperplanes. A single SVM does binary classification and can differentiate between two classes. In order to differentiate between K classes, one can use (K – 1) SVMs. Each one would predict membership in one of the K classes.

5.5.1 Weighted Average

- **Mean Average Precision (MAP)** is also called average precision at seen relevant documents. It determine precision at each point when a new relevant document gets retrieved. Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved.
- Avoids interpolation, use of fixed recall levels. MAP for query collection is arithmetic averaging. Average precision - recall curves are normally used to compare the performance of distinct IR algorithms.
- Use P = 0 for each relevant document that was not retrieved. Determine average for each query, then average over queries :

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(doc_i)$$

where Q_j = Number of relevant document for query j

N = Number of queries

$P(doc_i)$ = Precision at i^{th} relevant document

Precision - recall appropriateness :

- Precision and recall have been extensively used to evaluate the retrieval performance of IR algorithms. However, a more careful reflection reveals problems with these two measures :
- First, the proper estimation of maximum recall for a query requires detailed knowledge of all the documents in the collection.
- Second, in many situations the use of a single measure could be more appropriate.
- Third, recall and precision measure the effectiveness over a set of queries processed in batch mode.

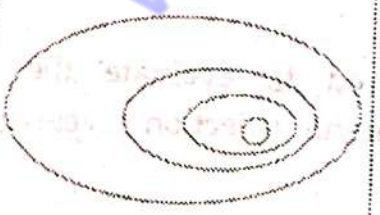
- Fourth, for systems which require a weak ordering though, recall and precision might be inadequate.

Single value summaries :

- Average precision - - recall curves constitute standard evaluation metrics for information retrieval systems. However, there are situations in which we would like to evaluate retrieval performance over individual queries. The reasons are two fold :
  1. First, averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
  2. Second, we might be interested in investigating whether a algorithm outperforms the other for each query.
- In these situations, a single precision value can be used.

5.5.2 Multiclass Classification Techniques

- Each training point belongs to one of N different classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs. The multi-class classification problem refers to assigning each of the observations into one of k classes.
- A common way to combine pair wise comparisons is by voting. It constructs a rule for discriminating between every pair of classes and then selecting the class with the most winning two-class decisions. Though the voting procedure requires just pair wise decisions, it only predicts a class label.
- Example of multi-label classification is as follows :

1. Is it eatable ? 2. Is it sweet ? 3. Is it a fruit ? 4. Is it a banana ?	1. Is it a banana ? 2. Is it an apple ? 3. Is it an orange ? 4. Is it a pineapple ?	1. Is it a banana ? 2. Is it yellow ? 3. Is it sweet ? 4. Is it round ?
		
Nested/Hierarchical	Exclusive/Multi-class	General/Structured

- Fig. 5.5.1 and 5.5.2 shows binary and multiclass classification.

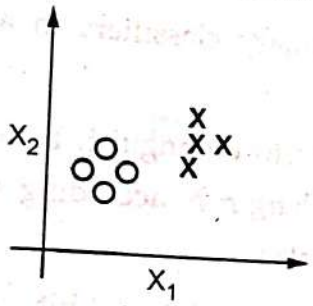


Fig. 5.5.1 Binary classification

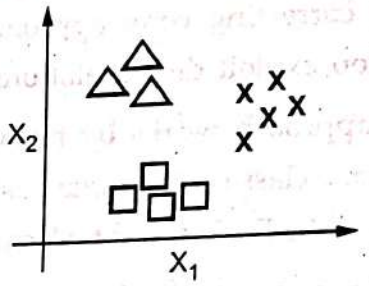


Fig. 5.5.2 Multiclass classification

- Multiclass classification through binary classification :

1. One Vs All (OVA) :

- For each class build a classifier for that class vs the rest. Build N different binary classifiers.
- For this approach, we require N = K binary classifiers, where the k^{th} classifier is trained with positive examples belonging to class k and negative examples belonging to the other K – 1 classes.
- When testing an unknown example, the classifier producing the maximum output is considered the winner, and this class label is assigned to that example.
- It is simple and provides performance that is comparable to other more complicated approaches when the binary classifier is tuned well.

2. All-Vs-All (AVA) :

- For each class build a classifier for those class vs the rest. Build N (N – 1) classifiers, one classifier to distinguish each pair of classes i and j .
- A binary classifier is built to discriminate between each pair of classes, while discarding the rest of the classes.
- When testing a new example, a voting is performed among the classifiers and the class with the maximum number of votes wins.

3. Calibration

- The decision function f of a classifier is said to be calibrated or well-calibrated if P (x is correctly classified | f(x) = s) ≃ s
- Informally f is a good estimate of the probability of classifying correctly a new datapoint x which would have output value x. Intuitively if the "raw" output of a classifier is g you can calibrate it by estimating the probability of x being well classified given that g(x)=y for all y values possible.

4. Error-Correcting Output-Coding (ECOC)

- Error correcting code approaches try to combine binary classifiers in a way that lets you exploit de-correlations and correct errors.
- This approach works by training N binary classifiers to distinguish between the K different classes. Each class is given a codeword of length N according to a binary matrix M . Each row of M corresponds to a certain class.
- The following table shows an example for K = 5 classes and N = 7 bit code words.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7
Class 1	0	0	0	0	0	0	0
Class 2	0	1	1	0	0	1	1
Class 3	0	1	1	1	1	0	0
Class 4	1	0	1	1	0	1	0
Class 5	1	1	0	1	0	0	1

- Each class is given a row of the matrix. Each column is used to train a distinct binary classifier. When testing an unseen example, the output codeword from the N classifiers is compared to the given K code words, and the one with the minimum hamming distance is considered the class label for that example.

Example 5.5.1 *Consider the following three-class confusion matrix.*

		Predicted		
		15	2	3
Actual		7	15	8
		2	3	45

Calculate precision and recall per class. Also calculate weighted average precision and recall for the classifier.

Solution :

		Predicted		
	15	2	3	20
Actual	7	15	8	30
	2	3	45	50
	24	20	56	100

$$\text{Classifier accuracy} = \frac{15+15+45}{100} = \frac{75}{100} = 0.75$$

Calculate per-class precision and recall :

$$\text{First class} = \frac{15}{24} = 0.63 \quad \text{and} \quad \frac{15}{20} = 0.75$$

$$\text{Second class} = \frac{15}{20} = 0.75 \quad \text{and} \quad \frac{15}{30} = 0.50$$

$$\text{Third class} = \frac{45}{56} = 0.8 \quad \text{and} \quad \frac{45}{50} = 0.9$$

Example 5.5.2 *Prove with an example : Accuracy = 1 – Error rate.*

Solution : Accuracy is the percent of correct classifications. Error rate is the percent of incorrect classifications. Classification accuracy is a misleading measure of performance when the data are not perfectly balanced. This is because a classifier may take advantage of an imbalanced dataset and trivially achieve a classification accuracy equal to the fraction of the majority class.

Review Questions

1. *Explain construction of multi-class classifier,*
   i) One Vs all approach ii) One Vs one approach
   iii) Error correcting output codes approach.
2. *Explain any two approaches to construct multiclass classifier.*

5.6 t - Test

- When a small sample (size < 30) is considered, the tests are inapplicable because the assumptions we made for large sample tests, do not hold good for small samples.
- In case of small samples it is not possible to assume,
  i) That the random sampling distribution of a statistics normal
  ii) The sample values are sufficiently close to population values to calculate the S.E. of estimate.
- Thus an entirely new approach is required to deal with problems of small samples. But one should note that the methods and theory of small samples are applicable to large samples but its converse is not true
- When sample sizes are small, as is often the case in practice, the Central Limit Theorem does not apply. One must then impose stricter assumptions on the population to give statistical validity to the test procedure. One common assumption is that the population from which the sample is taken has a normal probability distribution to begin with.
- Degree of freedom (df) : By degree of freedom we mean the number of classes to which the value can be assigned arbitrarily or at will without voicing the restrictions or limitations placed.
- For example, we are asked to choose any 4 numbers whose total is 50. Clearly we are at freedom to choose any 3 numbers say 10, 23, 7 but the fourth number, 10 is fixed since the total is 50 [50 – (10 + 23 + 7) = 10]. Thus we are given a restriction, hence the freedom of selection of number is 4 – 1 = 3.
- The degree of freedom (df) is denoted by ν(nu) or df and it is given by $\nu = n - k$, where n = number of classes and k = number of independent constrains.

5.6.1 t - Test for Single Mean

- When the sample values come from a normal distribution, the exact distribution of "t" was worked out by W. S. Gossett. He called it a **t - distribution.**
- Unfortunately, there is not one t - distribution. There are different t - distributions for each different value of n. If n = 7 there is a certain t - distribution but if n = 13 the t - distribution is a little different. We say that the variable t has a t - distribution with n–1 degrees of freedom.
- Suppose a simple random sample of size n is drawn from a population. If the population from which the sample is taken follows a normal distribution, the distribution of the random variable,

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

follows **Student's t - Distribution** with n – 1 degrees of freedom.

- The sample mean is $\bar{x}$ and the sample standard deviation is s.
- The **degrees of freedom** are the number of free choices left after a sample statistic such as is calculated. When you use a t - distribution to estimate a population mean, the degrees of freedom are equal to one less than the sample size.

  d.f. = n – 1

Assumptions :

1. Population is normal although this assumption can be relaxed if sample size is "large".
2. Random sample was drawn from the population of interest.

- Based on the comparison of calculated 't' value with the theoretical 't' value from the table, we conclude :

Shape of student's t - distribution

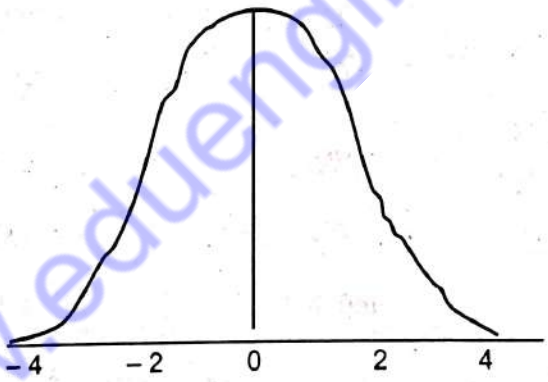


Fig. 5.6.1

5.6.2 Properties of Students t - Distribution

1. The t - distribution is different for different degrees of freedom.
2. The t - distribution is centered at 0 and symmetric about 0.
3. The total area under the curve is 1. The area to the left of 0 is 1/2 and the area to the right of 0 is 1/2.
4. As the magnitude of t increases the graph approaches but never equals 0.
5. The area in the tails of the t - distribution is larger than the area in the tails of the normal distribution.
6. The shape of the t-distribution is dependent on the sample size n.

6. As sample size n increases, the distribution becomes approximately normal.
7. The standard deviation is greater than 1.
8. The mean, median, and mode of the t-distribution are equal to zero.
9. The area in the tails of the t - distribution is a little greater than the area in the tails of the standard normal distribution, because we are using s as an estimate of σ, thereby introducing further variability.
10. As the sample size n increases the density of the curve of t get closer to the standard normal density curve. This result occurs because as the sample size n increases, the values of s get closer to σ, by the law of large numbers.

T - critical values

- Critical values for various degrees of freedom for the t - distribution are (compared to the normal)

n	Degrees of freedom	$t_{0.025}$
6	5	2.571
16	15	2.131
31	30	2.042
101	100	1.984
1001	1000	1.962
Normal	"Infinite"	1.960

5.6.3 t - Test for Correlation Coefficients

- The correlation coefficient, ρ (rho), is a popular statistic for describing the strength of the relationship between two variables.
- The correlation coefficient is the slope of the regression line between two variables when both variables have been standardized by subtracting their means and dividing by their standard deviations. The correlation ranges between plus and minus one.
- When ρ is used as a descriptive statistic, no special distributional assumptions need to be made about the variables (Y and X) from which it is calculated.

- When hypothesis tests are made, you assume that the observations are independent and that the variables are distributed according to the bivariate-normal density function.
- However, as with the t-test, tests based on the correlation coefficient are robust to moderate departures from this normality assumption.
- The population correlation ? is estimated by the sample correlation coefficient r. Note we use the symbol R on the screens and printouts to represent the population correlation.
- t - test for correlation coefficients formul

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

With degrees of freedom equal to n – 2.

- The steps to be followed for the t - test for correlation coefficient is listed below :

1. State the null hypothesis and alternative hypothesis.

   $H_0 = \rho = 0$

   $H_a = \rho \neq 0$

   Here ρ is the population correlation coefficient.

2. State the significance level.
3. Find the test statistic of correlation coefficient with the above-defined formula.
4. To make a decision, use the critical value approach or the p - value approach
5. Finally, state the conclusion.

- The above test is conducted with the supposition that the association is linear between the variables and originate from a normal distribution that is bivariate.
- The t-test is always used for population correlation coefficient of zero. So, in order to test the population correlation coefficient other than zero, z-test for correlation coefficient is used to test the significance of the correlation coefficient.

5.7 McNemar's Test

- The McNemar test is a non-parametric test for paired nominal data. It is used for finding a change in proportion for the paired data. It compare the performance of two classifiers on N items from a single test set.
- McNemar's test is used to compare the performance of two classifiers on the same test set. This test works if there are a large number of items on which A and B make different predictions.

- McNemar's test is applied to 2 × 2 contingency tables with matched pairs of subjects to determine whether the row and column marginal frequencies are equal.
- The three main assumptions for the test are :
  1. We must have one nominal variable with two categories and one independent variable with two connected groups.
  2. The two groups in the dependent variable must be mutually exclusive.
  3. Sample must be a random sample.

5.8 K - fold CV Paired t Test

- We use k - fold cross-validation to get K training/validation set pairs. To train the two classification algorithms on the training sets T_i; where i = 1; ...; K and test on the validation sets V_i.
- The error percentages of the classifiers on the validation sets are recorded as p_i^1 and p_i^2.
- If the two classification algorithms have the same error rate, then we expect them to have the same mean, or equivalently, that the difference of their means is 0.
- The difference in error rates on fold i is $p_i = p_i^1 - p_i^2$. This is a paired test; that is, for each i, both algorithms see the same training and validation sets.
- When this is done K times, we have a distribution of p_i containing K points. Given that p_i^1 and p_i^2 are both (approximately) normal, their difference p_i is also normal. The null hypothesis is that this distribution has 0 mean
  ($H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$)
- We define :

$$m = \frac{\sum_{t=1}^{N} p_i}{K}, \quad S^2 = \frac{\sum_{t=1}^{K} (p_i - m)^2}{K-1}$$

Under the null hypothesis that $\mu = 0$, we have a statistic that is t-distributed with K – 1 degrees of freedom :

$$\frac{\sqrt{k}\,(m-0)}{S} = \frac{\sqrt{k}\,(m)}{S} \sim t_{k-1}$$

- Thus the K - fold cv paired t - test rejects the hypothesis that two classification algorithms have the same error rate at significance level α if this value is outside the interval $(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$.
- If we want to test whether the first algorithm has less error than the second, we need a one-sided hypothesis and use a one-tailed test :

$$H_0 : \mu \geq 0 \text{ vs. } H_1 : \mu < 0$$

- If the test rejects, out claim that the first one has significantly less error is supported.

- Advantage is that each test set is independent of others. But the training sets still overlap. This overlap may prevent the test from obtaining a good estimate of the amount of variation that would be observed if each training set were completely independent of previous training sets.
- The variance in the t statistic maybe sometimes underestimated, the means are occasionally poorly estimated and this may result in large t values.

5.9 Two Marks Questions with Answers

Q.1 Define Bootstrapping.

Ans. : Bootstrapping is a method of sample reuse that is much more general than cross-validation. The idea is to use the observed sample to estimate the population distribution. Then samples can be drawn from the estimated population and the sampling distribution of any type of estimator can itself be estimated.

Q.2 What is confusion matrix ?

Ans. : The evaluation measures in classification problems are defined from a matrix with the number of examples correctly and incorrectly classified for each class, named confusion matrix.

Q.3 What is cross-validation ?

Ans. : Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

Q.4 Explain McNemar's test.

Ans. :

- The McNemar test is a non-parametric test for paired nominal data. It is used for finding a change in proportion for the paired data. It compare the performance of two classifiers on N items from a single test set.
- McNemar's test is used to compare the performance of two classifiers on the same test set. This test works if there are a large number of items on which A and B make different predictions.

Q.5 What is K - fold cross-validation ?

Ans. : K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called **folds**. For each learning set, the prediction function uses k – 1 folds and the rest of the folds are used for the test set.

Q.6 What is a T - test ?

Ans. : The t - test compares the means (averages) of two populations to determine how different they are from each other. The test generates a t-score and p-value, which quantify exactly how different each population is and the likelihood that this difference can be explained by chance or sampling error.

Q.7 List the applications of cross-validation.

Ans. :

- This technique can be used to compare the performance of different predictive modeling methods.
- It has great scope in the medical research field.
- It can also be used for the meta-analysis, as it is already being used by the data scientists in the field of medical statistics.

Q.8 Explain merits and demerits of t-test.

Ans. : Merits :

1. Easy to gather data.
2. Determine source data.
3. Essential for generalization.

Demerits :

1. It may contains small amount of noise.
2. If the data collected violates the assumption of the t - test, then the output is unreliable.
3. T-test cannot be used for multiple comparisons

❑❑❑

EDU ENGINEERING

PIONEER OF ENGINEERING NOTES

TAMIL NADU'S BEST EDTECH PLATFORM FOR ENGINEERING

CONNECT WITH US

WEBSITE: www.eduengineering.net

TELEGRAM: @eduengineering

INSTAGRAM: @eduengineering

- **Regular Updates for all Semesters**
- **All Department** Notes AVAILABLE
- **Handwritten** Notes AVAILABLE
- **Past Year** Question Papers AVAILABLE
- **Subject wise** Question Banks AVAILABLE
- **Important Questions** for Semesters AVAILABLE
- Various **Author Books** AVAILABLE