Q1. Explain the role of machine algorithm in Spam filtering

•E-mail provide perfect way to send millions of advertisement at no cost for the sender, and this unfortunate fact is nowaday extensively exploited by several organizations. • As a result, the e-mailboxes of millions of people get cluttered with all this so-called unsolicited bulk e-mail also known as "spam" or "junk mail". • Machine learning methods of recent are being used to successfully detect and filter spam emails. • Different categories of spam filtering1. Content Based Filtering Technique : Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbore , Neural Networks. 2. Case Base Spam Filtering Method : Case base or sample base filtering is one of the popular spam filtering methods. Firstly, all emails both non-spam & spam emails are extract from each user email using collection model.

Q2. Explain the role of machine learning algorithms in Natural Language processing = **• The role of machine learning and Al in natural language processing (NLP) and text analytics is to improve, accelerate and automate the underlying text analytics functions and NLP features that turn unstructured text into useable data and insights • Machine learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. • The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning. • It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning.• The most popular NLP machine learning algorithms are: 1.Support Vector Machines 2 Bayesian Networks 3 Maximum Entropy 4 Conditional Random Field 5 Neural Networks/Deep learning.**

Q3. What problems are faced by SVM when used with real datasets?
***1*. Unbalanced data – where the negative instances exceed the positive instances 2. Multilabel classification – SVM is design for binary classification, multilabel is computationally expensive 3. When large real datasets is used with support vector machine, it can extract a very large number of support vectors to increase accuracy and that can slow down the whole process. 4.To allow finding out a trade-off between precision and number of support vectors, Scikitlearn provides implementation called NuSVC, where the parameter nu (bounded between 0 and 1) can be used to control at same time number of support vectors & training errors. •NuSVC is defined as • class sklearn.svm.NuSVC(nu=0.5, kernel='rbf', degree=3, gamma=0.0, coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200)**

Q6. With reference to Hierarchical Clustering, explain the issue of connectivity constraints= **• Scikit-learn also allows specifying a connectivity matrix, which can be used as a constraint when finding the clusters to merge. • In this way, clusters which are far from each other (nonadjacent in the connectivity matrix) are skipped. • A very common method for creating such a matrix involves using the k-nearest neighbors graph function, that is based on the number of neighbors a sample :sklearn.datasets.make_circles(n_samples = 100, shuffle=True,noise=None, random_state=None, factor=0.8) • It makes a large circle containing a smaller circle in 2d. A simple toy dataset to visualize clustering and classification algorithms. • Parameters : 1. n_samples : int, optional (default=100) 2. shuffle : bool, optional (default=True) 3. noise : double or None (default=None) 4. random_state : int, Random State instance or None (default) 5. factor : 0 < double < 1 (default=.8) Scale factor between inner and outer circle**

Q5. Explain Evaluation methods for clustering algorithms =
1] Homogeneity • Homogeneity metric of a cluster labeling given ground truth. A clustering result satisfies homogeneity if all of its cluster contain only data point which are members of a single class. • This metric is independent of absolute value of the label: a permutation of the class or cluster label values won't change score value in any way. •To have homogeneity score, it's necessary to normalize this value considering initial entropy of class set H(C) : h=(1-(H(C|K)/H(C)) 2]Completeness•A complementaryrequirement is that each sample belonging to a class is assigned to the same cluster. • A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. set H(k) : k=(1-(H(C|K)/H(C)) 3] Adjusted Rand Index • The adjusted rand index measures the similarity between original class partitioning (Y) and the clustering. • If total number of samples in the dataset is n, the rand index is defined as : R= (a+b)/(n/2)

Q4. Explain role of machine learning the following common un-supervised learning problems: **i) Object segmentation : Object segmentation is the process of splitting up an object into a collection of smaller fixed-size objects in order to optimize storage and resources usage for large objects. S3 multi-part upload also creates segmented objects, withobject representing each part. ii) Similarity detection : In contrast to symmetry detection, automatic similarity detection is much harder and more time-consuming. The symmetry factored embedding & symmetry factored distance can be used to analyze symmetries in pointssets. A hierarchical approach was used for building graph of all subparts of object.**

Q7. With reference Clustering, explain issue of "Optimization of clusters"= **• The first method is based on the assumption that an appropriate number of clusters must produce a small inertia.•However, this value reaches its minimum (0.0) when the number of clusters is equal to the number of samples; therefore, we can't look for the minimum, but for a value which is a trade-off between the inertia and the number of clusters.• Given a partition of a proximity matrix of similarities into clusters, the program finds a partition with K classes that maximizes a fit criterion. Different options are available for measuring fit.•The default option (correlation) maximizes the the data matrix X and a structure matrix A in which a(i,j) = 1 if nodes i and j have been placed in the same class and a(i,j) = 0 otherwise.•Thus, a high correlation is obtained when the data values are high within-class and low between-class. This assume similarity data as input•For dissimilarity data,program maximizes negative of the correlation. Another measure of fit is the density function, which is simply the average data value within classes.•There is also a pseudo correlation measure that seeks to measure the difference between average value within classes & the average value between classes.**

Q8. Explain with example the variant of SVM, the Support vector regression **• Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). • The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. • First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. • In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem • But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration • However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.**

Q9.Define Bayes Theorem. Elaborate Naive Bayes Classifier working with example.= **• In machine learning, we try to determine the best hypothesis from some hypothesis space H, given the observed training data D. • In Bayesian learning, the best hypothesis means the most probable hypothesis, given the data D plus any initial knowledge about the prior probabilities of the various hypotheses in H. • Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.• Bayes' theorem is a method to revise the probability of an event given additional information.• Bayes's theorem calculates a conditional probability called a posterior or revised probability. If A and B are two random variables P(A/B) = (P(B/A)P(A))/P(B)• In the context of classifier hypothesis h and training data I.p(h/I) = (P(i/h)P(h))/ P(I) • Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.It is highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. • A Naive Bayes Classifier is a program which predicts a class value given a set of attributes. • For each known class value, 1. Calculate probabilities for each attribute, conditional on the class value.2. Use the product rule to obtain a joint conditional probability for the attributes.3. Use Bayes rule to derive conditional probabilities for the class variable. • Once this has been done for all class values, output the class with the highest probability.**

Q10.Explain the Lasso, and ElasticNet and ridge types of regression.
1]Lasso • One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero. • Thus, the final model will include all p predictors, which creates a challenge in model interpretation. A more modern machine learning alternative is the lasso. • The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero. • Lasso : Lasso is a regularized regression machine learning technique that avoids over fitting of training data and is useful for feature selection. • The lasso is a shrinkage method like ridge, with subtle but important differences.• The only difference from Ridge regression is that the regularization term is in absolute value. But this difference has a huge impact on the trade-off. 2] ElasticNet • ElasticNet, which combines both Lasso and Ridge into a single model with two penalty factors : One proportional to L1 norm and the other to L2 norm. • Elastic net is a related technique. Use elastic net when you have several highly correlated variables. Lasso provides elastic net regularization when you set the alpha name-value pair to a number strictly between 0 and 1. • Elastic net can generate reduced models by generating zero-valued coefficients. Empirical studies have suggested that the elastic net technique can outperform lasso on data with highly correlated predictors. • The elastic net technique solves this regularization problem. For an strictly between 0 and 1 and a nonnegative λ , elastic net solves the problem 3]ridge regression=•it an extension to linear Regression •This technique shrinks regression Coefficient ,which result in variable with minor Constribution resulting in their Coefficients Close to zero •The amount of penatly can be fine tune with lambda a constant •when lambda=0 term of penalty will no effect r & ridge regression equivalent with ordinary least square Coefficient • As lambda increase to large infinite value the Shrinkage penalty grow & ridge regression will get close to zero. • As compare to ordinary Least Square regression •Ridge regression is highly Senssetive. •This shrinkage achieved by the term L2-norm which is used in penalizing the regression model.

Q.11 Explain concepts of weak and eager learner. Ans.: Concepts of weak and eager learner=**•Eager learning is a learning method in which the system tries to construct a general, input-independent target function during training of the system, as opposed to lazy learning, where generalization beyond the training data is delayed until a query is made to the system.• Combining several weak learners to give a strong learner. It is a kind of multiclassifier systems and meta-learners. Ensemble typically applied to a single type of weak learner. •Lazy learning (e.g, instance-based learning): Simply stores training data (or only minor processing) and waits until it is given a test tuple •Eager learning (the above discussed methods): Given a set of training tuples, constructs a classification model before receiving new (e.g., test) data to classify.**

Q.12 List the advantages and disadvantages of decision tree. =
• Advantages:1. Decision trees can handle both nominal and numeric attributes.2. Decision tree representation is rich enough to represent any discrete value classifier.3. Decision trees are capable of handling datasets that may have errors.4. Decision trees are capable of handling datasets that may have missing values. • Disadvantages 1. Most of the algorithms require that the target at-tribute will have only discrete values.2. Most decision-tree algorithms only examine single field at time 3. Decision trees are prone to errors in classification problems with many class.

Q.13 Explain Ada Boost algorithms in detail **=1. It combines weak classifier algorithm to form strong classifier, then by selecting, training set at every iteration multiple classifiers are combined which gives good accuracy score2. Correct amount of weight can be assigned in final voting, which improves accuracy. •Algorithm Step 1:Choose the training set and train the algorithm. Step 2: Retrains the algorithm iteratively by selecting another set of training data based on the accuracy of previous training.Step 3: The weight-age depends on the accuracy achieved for each trained classifier.Step 4:It assigns weight to each training item.Step 5:Higher weights are assigned to misclassified item so that those items can appear in the training subset of next classifier with higher probability Step 6: After training weights is assigned to each classifier also baseon accuracy. Step 7: Higher weights are assigned to more accurate classifier get more impact on the final outcome.**

*Q14.Impurity measure***• defines how well e classes are separated. In general the impurity measure should satisfy: Largest when data are split evenly for attribute values.Pi =1/Number of clasess. Should be 0 when all data belong to the same class.• Two methods are used for measuring impurity: Gini Index and Entropy based measure. •Entropy based measure I(D) =Entropy (D) ΣPilog Pi • Gini Index measure :I(D) = Gini(D) = 1-ΣP^2i •Here, the sum is always extended to all classes. This is a very common measure and it is used as a default value by scikit-learn.**

*Q15.Explain K-Means algorithms =***1. In k-means, k is the number of clusters given by user and objects are classified into k clusters based on their attributes.2. K-means is one of the simplest unsupervised learning algorithms.3.Define K centroids for K clusters which are generally far away from each other.4.Group the objects into clusters based on the distance with respect to centroid 5. After this first step, again calculate the new centroid for each cluster based on the elements of that cluster.6. Follow the same method and group the elements based on new centroid.7. At every step, the centroid changes and elements move from one cluster to another.8. Do the same process till no element is moving from one cluster to another i.e. till two consecutive steps with the same centroid and the same elements are obtained. 9. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.**