Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

Post Lab 1	**Experiment 1- Linear Regression**
Q1	**Can you state the fundamental presumptions behind linear regression?** There are three crucial assumptions one has to make in linear regression. They are, 1. It is imperative to have a linear relationship between the dependent and independent A scatter plot can prove handy to check out this fact. 2. The independent variables in the dataset should not exhibit any multi-collinearity. In case they do, it should be at the barest minimum. There should be a restriction on their value depending on the domain requirement. 3. Homoscedasticity is one of the most critical It states that there should be an equal distribution of errors.
Q2	**What is Heteroscedasticity?** Heteroscedasticity is the exact opposite of homoscedasticity. It entails that there is no equal distribution of the error terms. You use a log function to rectify this phenomenon.
Q3	**What distinguishes R square from adjusted R square in particular?** In linear regression, you use both these values for model validation. However, there is a clear distinction between the two. R square accounts for the variation of all independent variables on the dependent variable. In other words, it considers each independent variable for explaining the variation. In the case of Adjusted R square, it accounts for the significant variables alone for indicating the percentage of variation in the model. By significant, we refer to the P values less than 0.05.
Post Lab 2	**Experiment 2- Multivariate Linear Regression**
Q1	**Is multivariate regression the same as linear regression?** **Multiple regression is a broader class of regressions that encompasses**

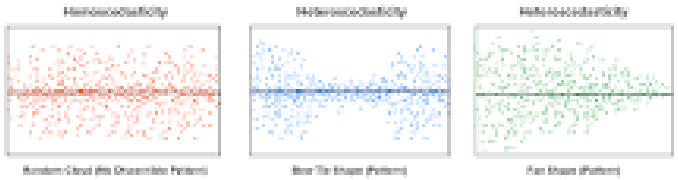
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	linear and nonlinear regressions with multiple explanatory variables. Whereas linear regress only has one independent variable impacting the slope of the relationship, multiple regression incorporates multiple independent variables.
Q2	**What is multivariate regression with example?** As the name implies, multivariate regression is **a technique that estimates a single regression model with more than one outcome variable**. When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.
Q3	**Why do we use multivariable regression models?** Multivariable regression models are used **to establish the relationship between a dependent variable (i.e. an outcome of interest) and more than 1 independent variable**. Multivariable regression can be used for a variety of different purposes in research studies.
Q4	**What are the assumptions of multivariate regression?** The relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. So the assumptions are: **independence; linearity; normality; homoscedasticity**.
Q5	**Is multivariate normality linear regression?**  **Multivariate Normality is the third assumption in assumptions of linear regression**. The linear regression analysis requires all variables to be multivariate normal. Means data should be normally distributed. As sample

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	sizes increase then the normality for the residuals is not needed.
Q6	**What is the difference between univariate and multivariate regression?** **Univariate analysis allows us to understand the distribution of values for one variable while multivariate analysis allows us to understand the relationship between several variables**
Q6	**What does multivariate analysis tell you?** Multivariate analysis can help companies **predict future outcomes, improve efficiency, make decisions about policies and processes, correct errors, and gain new insights**. Multivariate analysis often builds on univariate (one variable) analysis and bivariate (two variable) analysis
Q7	**Is multivariate analysis quantitative or qualitative?** Multivariate Analysis is an advanced statistical technique for pattern recognition, examining relationships, developing models and **can be used for both qualitative and quantitative data**.
Post Lab 3	**Experiment 3-Logistic Regression**
Q1	**What is a logistic function t?** **What is a logistic function's range of values?** $(z) = 1/(1+e^{-z})$ The values of a logistic function will range from 0 to 1. The values of Z will vary from -infinity to +infinity.
Q2	**What makes logistic regression so well-liked?** Logistic regression is famous because it can convert the values of logits (logodds), which can range from -infinity to +infinity to a range between 0 and 1. As logistic functions output the probability of occurrence of an event,

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	it can be applied to many real-life scenarios. It is for this reason that the logistic regression model is very popular. It is one of the most commonly asked logistic regression questions. Logistic regression is also predictive analysis just like all the other regressions and is used to describe the relationship between the variables. There are many real-life examples of logistic regression such as the probability of predicting a heart attack, the probability of finding if the transaction is going to be fraudulent or not, etc.
Q3	**How can a logistic regression model's probability be stated as a conditional probability?** P(Discrete value of Target variable | X1, X2, X3….Xk). It is the probability of the target variable taking up a discrete value (either 0 or 1 in case of binary classification problems) when the values of independent variables are given. For example, the probability an employee will attrite (target variable) given his attributes such as his age, salary, KRA's, etc.
Q4	***What are the outputs of the logistic model and the logistic function?*** The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities. Logistic model = α+1X1+2X2+….+kXk. The output of the same will be logits. Logistic function = f(z) = $1/(1+e^{-(\alpha+1X1+2X2+….+kXk)})$. The output, in this case, will be the probabilities.
Post Lab 4	**Experiment 4 – Decision Tree**
Q1	**Which types of problems are most suited for decision trees?** • Decision trees are most suitable for tabular data. • The outputs are discrete.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	• Explanations for decisions are required. • The training data may contain errors. • The training data may contain missing attribute values.
Q2	**What is the inductive bias of decision trees?** Shorter trees are preferred over longer trees. Trees that place high information gain attributes close to the root are preferred over those that do not.
Q3	**How does a decision tree handle missing attribute values?** One way to assign the most common value of that attribute to the missing attribute value. The other way is to assign a probability to each of the possible values of the attribute based on other samples.
Q4	**How does a decision tree handle continuous attributes?** By converting continuous attributes to a threshold-based boolean attribute. The threshold is decided by maximizing the information gain.
Q4	**What is Information Gain? What are its disadvantages?** Information gain is the reduction in entropy due to the selection of an attribute. Information gain ratio biases the decision tree *against* considering attributes with a large number of distinct values which might lead to overfitting. In order to solve this problem, information gain ratio is used.

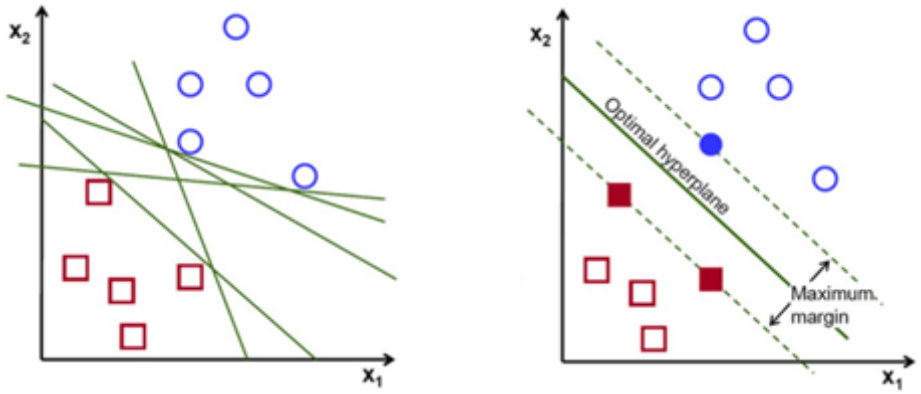
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

Post Lab 5	**Experiment 5 – Support Vector Machine**
Q1	**What is the geometric intuition behind SVM?** Explanation: If you are asked to classify two different classes. There can be multiple hyperplanes which can be drawn.  SVM chooses the hyperplane which separates the data points as widely as possible. SVM draws a hyperplane parallel to the actual hyperplane intersecting with the first point of class A (also known as Support Vectors) and another hyperplane parallel to the actual hyperplane intersecting with the first point of class B. SVM tries to maximize these margins. Eventually, this margin maximization improves the model's accuracy on unseen data.
Q2	**How would explain Convex Hull in light of SVMs?** Explanation: We simply build a convex hull for class A and class B and draw a perpendicular on the shortest distance between the closest points of both these hulls.

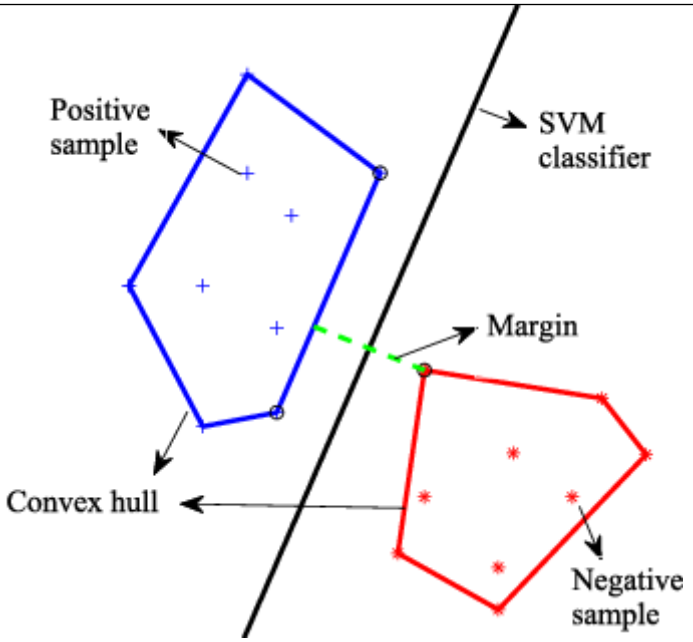
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	
Q3	**What do know about Hard Margin SVM and Soft Margin SVM?** Explanation: If a point Xi satisfies the equation ***Yi(WT*Xi +b) ≥ 1,*** then Xi is correctly classified else incorrectly classified. So we can see that if the points are linearly separable then only our hyperplane is able to distinguish between them and if any outlier is introduced then it is not able to separate them. So these type of SVM is called ***hard margin SVM*** *(since we have very strict constraints to correctly classify each and every data point).* To overcome this, we introduce a term ***(ξ)*** (pronounced as Zeta) $y_i(w^T x_i + b) \geq 1 - \xi_i$ *if ξi= 0, the points can be considered as correctly classified.* *if ξi> 0 , Incorrectly classified points.*

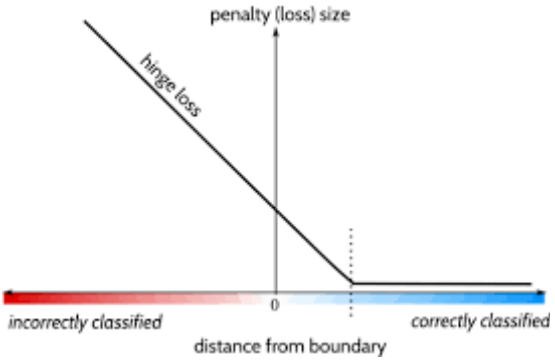
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

Q4	**What is Hinge Loss?** Explanation: Hinge Loss is a loss function which penalises the SVM model for inaccurate predictions. If ***Yi(WT*Xi +b) ≥ 1***, hinge loss is '**0**' i.e the points are correctly classified. When ***Yi(WT*Xi +b)*** **< 1**, then hinge loss increases massively. As ***Yi(WT*Xi +b)*** increases with every misclassified point, the upper bound of hinge loss {**1-** ***Yi(WT*Xi +b)***} also increases exponentially. Hence, the points that are farther away from the decision margins have a greater loss value, thus penalising those points.  We can formulate hinge loss as **max[0, 1-** ***Yi(WT*Xi +b)*****]**
Q5	**What's the "kernel trick" and how is it useful?** Explanation: Earlier we have discussed applying SVM on linearly separable data but it is very rare to get such data. Here, kernel trick plays a huge role. The idea is to map the non-linear separable data-set into a higher dimensional space where we can find a hyperplane that can separate the samples.

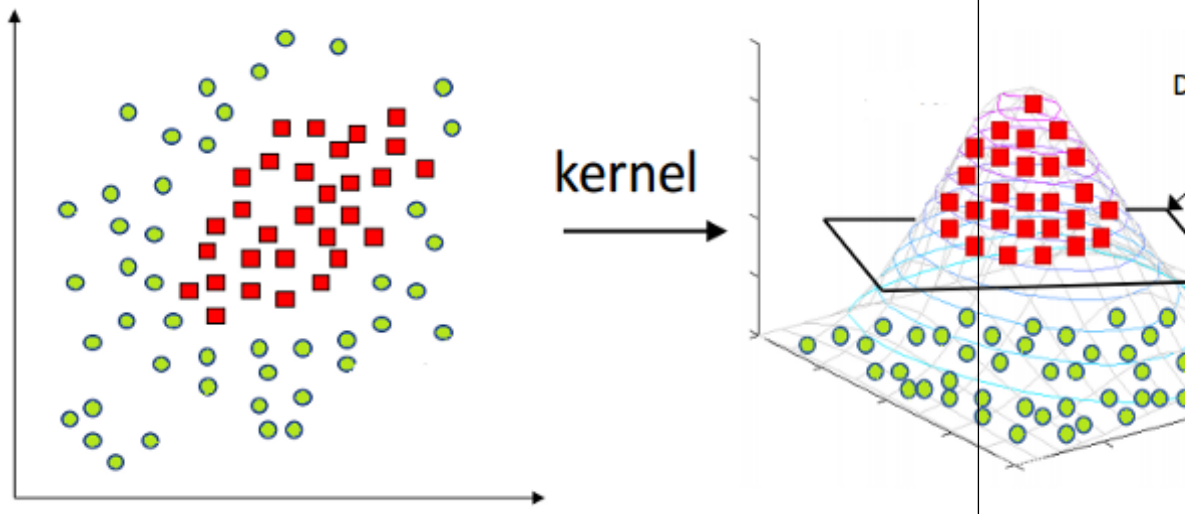
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	 It reduces the complexity of finding the mapping function. So, **Kernel function defines the inner product in the transformed space.** Application of the kernel trick is not limited to the SVM algorithm. Any computations involving the dot products (x, y) can utilize the kernel trick.
Q6	**Explain about SVM Regression?** Explanation: The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because the 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

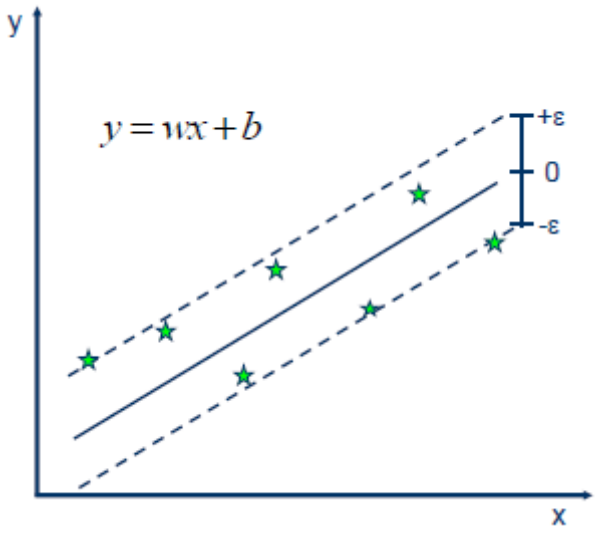
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions



- Solution:

$$\min \frac{1}{2}\|w\|^2$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon$$

$$wx_i + b - y_i \le \varepsilon$$

Post Lab 6	**Experiment 6 – Clustering Algorithm**
Q1	**What is *Similarity-based Clustering*?** Clustering, when the data are similar pairs of points is called **similarity-based clustering**. • A typical example of similarity-based clustering is community detection in social networks, where the observations are individual links between people, which may be due to friendship, shared interests, and work relationships. The *strength* of a link can be the frequency of interactions, for example, communications by e-mail, phone, or other social media, co-authorships, or citations. • In this clustering paradigm, the points to be clustered are not assumed to be part of a vector space. Their attributes (or features) are incorporated into a single dimension, the *link strength*, or *similarity*, which takes a numerical value SijS_{ij}Sij for each pair of points i, j. Hence, the natural representation for this problem is by means of the similarity matrix given below: $S = [S_{ij}]_{i,j=1}^{n}$ The similarities are symmetric $S_{ij} = S_{ji}$ and nonnegative $S_{ij} \ge 0$.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

Q2	**Why do you need to perform *Significance Testing* in *Clustering*?** • **Significance testing** addresses an important aspect of cluster validation. Many cluster analysis methods will deliver clusterings even for homogeneous data. They assume implicitly that clustering has to be found, regardless of whether this is meaningful or not. A critical and challenging question in cluster analysis is whether the identified clusters represent important underlying structure or are artifacts of natural sampling variation. • **Significance testing** is performed to distinguish between a clustering that reflects meaningful *heterogeneity* in the data and an artificial clustering of *homogeneous* data. • Significance testing is also used for more specific tasks in cluster analysis, such as; estimating the number of clusters, and for interpreting some or all of the individual clusters, to show the significance of the individual clusters.
Q3	**Give examples of using *Clustering* to solve real-life problems** • **Identifying cancerous data:** Initially we take known samples of a cancerous and non-cancerous dataset, and label both the samples dataset. Then both the samples are mixed and different clustering algorithms are applied to the mixed samples dataset. It has been found through experiments that a cancerous dataset gives the best results with unsupervised non-linear clustering algorithms. • **Search engines:** Search engines try to group similar objects in one cluster and the dissimilar objects far from each other. It provides results for the searched data according to the nearest similar object which is clustered around the data to be searched.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	• **Wireless sensor network's based application:** Clustering algorithm can be used effectively in *Wireless Sensor Network's based application*. One application where it can be used is in *Landmine detection*. The clustering algorithm plays the role of finding the Cluster heads (or cluster center) which collects all the data in its respective cluster.
Q4	***What's the difference between hard and soft clustering?*** Hard clustering means that each data point is assigned to a specific cluster, and soft clustering means that each data point is assigned a probability of belonging to each cluster.
Q5	***What are the challenges associated with clustering?*** There are a few challenges associated with clustering, the main one being that it can be difficult to determine the optimal number of clusters to use. If too few clusters are used, then important information may be lost. If too many clusters are used, then the data may become too fragmented and difficult to interpret. Another challenge is that some clustering algorithms can be sensitive to the order of the data, meaning that the results can vary depending on how the data is arranged.
Q6	***Why is it difficult to determine the optimal number of clusters in a dataset?*** There are a few reasons why it can be difficult to determine the optimal number of clusters in a dataset. One reason is that there is no guarantee that the data will be well-behaved and will cluster nicely. Another reason is that even if the data does cluster nicely, there is no guarantee that the clusters will be of equal size or that they will be spaced evenly. Finally, the optimal number of clusters may depend on the application or the specific

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	goal that you are trying to achieve.
Q7	***What are the differences between partitioning and hierarchical clustering?*** Partitioning clustering is a method of clustering data points into a set number of groups, while hierarchical clustering is a method of creating a hierarchy of clusters, with each cluster containing a subset of the data points. Partitioning clustering is typically faster than hierarchical clustering, but hierarchical clustering can produce more accurate results.
Post Lab 7	**Experiment 7- Ensemble Techniques**
Q1	**What are *Weak Learners*?** Answer In ensemble learning theory, we call **weak learners** (or **base models**) models that can be used as building blocks for designing more complex models by combining several of them. Most of the time, these basics models perform not so well by themselves either because they have a high bias (low degree of freedom models, for example) or because they have too much variance to be robust (high degree of freedom models, for example).
Q2	**How is a *Random Forest* related to *Decision Trees*?** • ***Random forest*** is an ***ensemble learning*** method that works by constructing a multitude of ***decision trees***. A random forest can be constructed for both classification and regression tasks. • Random forest **outperforms** decision trees, and it also does not have the habit of *overfitting* the data as decision trees do.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	• A decision tree trained on a specific dataset will become very deep and cause overfitting. To create a random forest, decision trees can be trained on different subsets of the training dataset, and then the different decision trees can be averaged with the goal of *decreasing the variance*.
Q3	**What are the differences between *Bagging* and *Boosting*?** • **Bagging** mostly aims at reducing variance. • **Boosting** is mainly focused on reducing **bias**. The base models that are considered for boosting are models with a low variance but high bias. • **Bagging** can be *parallelized*. The different models are fitted independently from each other. • **Boosting** can *not* be *parallelized*, and it can become too expensive to fit sequentially several complex models.
Q4	**How does *Stacking* work?** The idea of **stacking** is to learn several different weak learners and combine them by training a *meta-model* to output predictions based on the multiple predictions returned by these weak models. If a stacking ensemble is composed of L weak learners, then to fit the model the following steps are followed: • Split the training data into two folds. • Choose L weak learners and fit them to the data of the first fold. • For each of the L weak learners, make predictions for observations

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	in the second fold. • Fit the meta-model on the second fold, using predictions made by the weak learners as inputs. The process of *stacking* is shown in the figure below: 
Q5	**What is a *Super Learner Algorithm*?** • When a problem has already been fit with many different algorithms on the dataset, and some algorithms have been evaluated many times with different configurations, using all the models instead of just the best model from the group is the intuition behind the **super learner** ensemble algorithm. • The **super learner algorithm** involves first pre-defining the k-fold split of your data, then evaluating all different algorithms and algorithm configurations on the same split of the data. All out-of-fold predictions are then kept and used to train an algorithm that learns how to best combine the predictions. • The super learner technique is an example of the general method called *stacking*.
Q6	**What is *Meta-Learning*?** • **Meta-learning** refers to *learning about learning*. In machine learning, it most commonly refers to machine learning algorithms that learn from the output of other machine learning algorithms.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	• Meta-learning can be used in ensemble learning, e.g., fitting a meta-model on the out-of-fold predictions from the k-fold cross-validation.
Q7	**What are the differences between *Homogeneous* and *Heterogeneous Ensembles*?** • **Homogeneous** ensemble consists of members having a single-type base learning algorithm. Popular methods like *bagging* and *boosting* generate diversity by sampling from or assigning weights to training examples but generally utilize a single type of base classifier to build the ensemble. • **Heterogeneous** ensemble consists of members having different base learning algorithms such as *SVM*, *ANN*, and *Decision Trees*. A popular heterogeneous ensemble method is stacking, which is similar to boosting. • **Homogeneous** ensemble work by applying the same algorithm on all the estimators. These algorithms should not be fine-tuned. In contrast to *heterogeneous ensembles*, a large number of estimators are used. • **Heterogeneous** ensembles use different fine-tuned algorithms. They work well with *small* amount of estimators. The number of algorithms should always be odd to avoid ties.
Post Lab 8	**Experiment 8 - Dimensionality Reduction**
Q1	**What is *Principal Component Analysis (PCA)*?** • The ***Principal Component Analysis* (PCA)** is the process of computing principal components and using them to perform a

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	change of basis on the data. • The **Principal Component** of a collection of points in a *real coordinate space* are a sequence of p unit vectors, where the i-th vector is the direction of a line that best fits the data while being *orthogonal* to the i - 1 vectors. The best-fitting line is defined as the line that minimizes the average squared distance from the points to the line. • ***PCA*** is commonly used in *dimensionality reduction* by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.
Q2	**Explain how do you understand *Dimensionality Reduction*** **Dimensionality Reduction** is typically choosing a basis or mathematical representation within which you can describe most but not all of the variance within your data, thereby retaining the relevant information, while reducing the amount of information necessary to represent it. There are a variety of techniques for doing this including but not limited to PCA, ICA, and Matrix Feature Factorization. These will take existing data and reduce it to the most discriminative components. These all allow you to represent most of the information in your dataset with fewer, more discriminative features.
Q3	**How does the *Curse of Dimensionality* affect Machine Learning models?** As we increase the **number of dimensions**, our data becomes more **sparse**; every new dimension increases the volume of the feature space, giving our data a **higher differentiation** chance and therefore, the possibility of it becoming more spread out in a higher dimensional space than in a lower one. This means that if we need more **training samples** of a kind for our

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	model to be able • to learn about them and be able to predict them well, generalizing, in the future. • As we increase the number of dimensions, especially for parametric models, we **increase the time** it takes to **train** them. • Introducing features that don't add much value to our models and therefore increasing the number of dimensions, makes our model learn from these ***noisy*** or **irrelevant features** and can lead to a reduction in its performance. • More features and thus dimensions lead to models that are more complex and harder to interpret than those with a low number of features.
Q4	**What methods of *Hyperparameters Tuning* do you know?** • **Random Search**: here, we create a *grid of possible values* for hyperparameters. Each iteration tries a random combination of hyperparameters from this grid, records the performance, and lastly returns the combination of hyperparameters that provided the best performance. • **Grid search**: we also create a grid of possible values for hyperparameters, but in each iteration, we try a combination of hyperparameters in a *specific order* and the model it fits on each and every combination of hyperparameters possible and records the model performance. Finally, it returns the best model with the best hyperparameters.

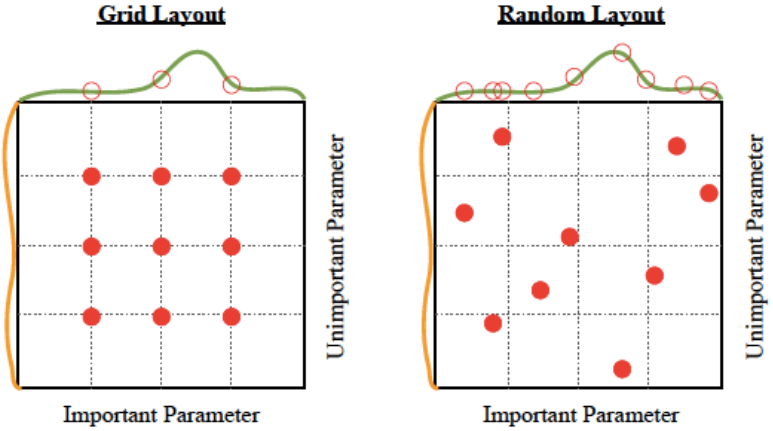
Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Machine Learning Post Lab Questions

	 • **Bayesian Optimization**: The idea here is to minimize the loss function of our model by changing *model parameters*. This approach helps us to find the minimal point in the minimum number of steps.