Atharva Prashant Pawar (9427) Comps-A [Batch D]

17/10/23

ML Postlab - 1

1. linear regression assumes that there is a linear relationship b/w the independent variables & the dependent variable. It also assumes that errors are normally distributed, have constant variance, & are independent.

2. Heteroscedasticity is a violation of the assumption in linear regression where the variance of the error terms is not constant accross all levels of the independent variable.
It leads to unequal spread of residuals.

3. R - squared (R²) measures the proportion of the variance in the dependent variable explained by the independent variables.
Adjusted R-squared (R² adjusted) considers the no. of predictors, penalizing the inclusion of unnecessary variables, providing a more accurate measure of model fit.

Atharva Prashant Pawar (9427) Comps-A Batch-D

ML PostLab - 2

17/10/23

1. No, multivariate regression is an extension of linear regression. Linear regression deals with one dependent variable & one independent variable, while multivariate regression involves multiple dependent & independent variables.

2. Multivariate regression involves predicting multiple dependent variables using multiple independent variables. eg: predicting both a student's math & english scores based on study hours & attendance.

3. Multivariable regression models allow's us to analyze & predict complex relationships b/w multiple independent variables & multiple dependent variables, providing a more comprehensive understanding of real-world scenarios.

4. Assumptions include linearity, independency of errors, homoscedasticity multivariate normality & no multicollinearity among independent var.

5. Multivariate normality is a specific assumption for multivariate regression, which requires that the residuals are normally distributed across all dependent variables.

6. Univariate Regrn	Multivariate Regrn.
① Univariate regr has one dependent variable.	Multivariate regr. involves multiple dependent variables.
② Univariate uses one Independent variable.	Multivariate uses multiple independent variables
③ Univaritate is simpler.	Multivariate is more complex, considering interdependencies.

7. Multivariate analysis helps identify complex relationships b/w multiple variables, enabiling a deeper understanding of how they collectively affect one anothe.

8. Multivaritate analysis is primarly quantitative It uses statistical techniqu. to analyze relationships & make numerical predictions based on multiple variables.

Atharva Prashant Pawar (9427) Comps-A (Batch-D)

17/10/23

ML - PostLab - 3

1. What is a logistic fun. f ?

...

⇒ $(z) = 1/(1+e^{-z})$

The values of a logistic. fun. will range from 0 to 1. The values of z will vary from ~~fn~~ $-\infty$ to $+\infty$.

2. Logistic regression is popular because it's effective for binary classification, interpretable, handles both categorical & numerical features & provides probabilities of class membership.

3. The logistic regression model estimates the conditional probability of an event. eg: case 1 occuring given the values of the i/p features.

4. The o/p of a logistic model is a Ⓟ b/w 0 & 1. the logistic fun. (sigmoid) maps i/p values to this Ⓟ range which aids in classification.

Atharva Prashant Pawar (9427) Comps-A [Batch D]

7/10/23

ML Post Lab - 4 Decision Tree

1. Which types of problems are most suited for decision trees?

→ Dicision trees are well-suited for classification & regression problems. They work best when the relationship b/w features & the target variables are non-linear or can be split into discrete decision points.

2. The inductive bias of decision trees is their preference for simple, easily interpretable models. They tend to favor shorter, more specific branches over complex ones, which can help with overfitting.

3. Decision trees can handle missing attributes values by using ~~technique~~ techniques like imputation, where they estimate missing values based on available data, or they can skip missing attributes during tree traversal.

4. Decision trees handles continuous attributes by selecting a threshold that best separates the data into two subsets. This threshold is choosen based on criteria like Gini impurity or info. gain.

5. Info. Gain (IG) measures the reduction in uncertainty about the target variable when a dataset is split based on an attribute. Disadvantages include a bias towards attributes with many values & overfitting on noisy data.

Atharva Prashant Pawar (9427) Comps-A [Batch-D]

17/10/23

ML Post Lab -5

1. SVM can be used for both classification as well as regression. It helps to find the hyperplane that maximaly separates different classes of data while mainteing The largest margin b/w the classes.

2. In the context of SVM, convex Hall is the outer boundary formed by the support vector & is critical in defining the margine & SVMs decision boundary. It represents the region in which SVM finds the optimal hyperplanes for classifications.

3. Hard margin:
It seeks to find hyperplane that perfectly separates 2 classes of data points without any misclassification.

4. Soft Margin:
This allows misclassification to a certain extent.

4. Hinge Loss:

This loss is used in ML, mainly in SVM & binary classification tasks.

It is designed to quantify the error.

$$\text{Loss}(y, f(x)) = \max(0, 1-y * f(x))$$

5. "Kernel Trick" is a fundamental concept in ML. It is used to implicitly map data from a lower-dimensional space to a higher-dimensional space without explicitly computing the transformation.

6. Explain about SVM regression.

- It is used for regression tasks. while traditional SVMs are designed for classification tasks, it also helps to predicting continuous numeric values.

Atharva Prashant Pawar (9427) Comps-A [Batch-D]

17/10/23

ML PostLab : 6

① Similarity - based clustering is a technique in unsupervised learning algo.
It uses similarity measures to compare data points & groups points into clusters. based on their dissimilarity or similarity.

② Significance testing in clustering is crucial for validating & ensuring the reliablity of the obtained clusters aiding in their interpretation & making informed decision about clustering methods & parameter.

③ 1) Customer Segmentation
2) Image compression
3) Healthcare.

④

	Hard Clustering :	Soft Clustering.
1.	In this each data point belongs exclusivly to one point.	Some points may belong to multiple clusters.
2.	Each point is assigned to a single cluster	They are associated with a set of clusters.
3.	kmeans, Hierarchical	Fuzzy k means, GMM

(5) It is difficult to determine the optimal no. of clusters.

The algo. are sometimes sensitive to order of data.

Results may change based on how data is arranged.

(6) It's difficult as from data you won't come to know how many clusters are there & if data is clusterable or not..

Clusters may vary in density & may not be of equal size.

(7)

	Portion clustering:	Hierarchical Clustering
1.	It aims to divide the datasets into a set of non-overlapping clusterings where each data points belongs exclusively to one cluster.	It constructs a tree like hierarchy of clusters where data points can belong to multiple clusters at diff levels.
2.	Need to specify no of clusters.	No need to specify
3.	kmean, kmedoids	Agglomerative & divisive clustering.

Atharva Prashant Pawar (9427) Comps-A [Batch D]

17/10/23

ML PostLab - 7

1. Weak learners are models that perform slightly better than random guessing or chance on a classification or regression task.
The model: are characterized by their limited predictive power when used individual.

2. The key idea behind a Random forest is that by combining multiple trees to add randomness.
It reduces overfitting caused by decision tree.

3. Bagging involves multiple base models on different random subsets. of the training data, created through bootstrapping.
Bagging reduces variance & helps to prevent overfitting.
Boosting uses multiple base models & sequentially trains them.
Each base model is trained to correct errors.
Boosting is effictive at reducing bias.

4. Stacking is an ensemble learning technique that combines the predictions from multiple base models.
It leverages strength from various models & combines them.

5. It combines multiple algs in a hierarchical fashion to make predictions. It is especially good at dealing with complex or noisy datasets.

6. Meta-Learning focuses on training models how to learn.

The idea here is to leverage the knowledge gained from previous tasks to facilitate faster & more accurate learning on new, unseen tasks.