Assignment-3

Batch DS2311

Name: Alvin V Regi

1. **When implementing linear regression of some dependent variable y on the set of independent variables $\mathbf{x} = (x_1, \cdots, x_r)$, where r is the number of predictors, which of the following statements will be true?**

Answer-

Both a and b are true.
a) β_0, $\beta_1 ... \beta_r$ are the regression coefficients in linear regression, where β_0 is the intercept, and $\beta_1 ... \beta_r$ are the coefficients for the independent variables.
b) Linear regression uses the method of ordinary least squares to determine the best-fitting line by minimizing the sum of squared differences between predicted and actual values.

2. **What indicates that you have a perfect fit in linear regression?**

Answer-

The value $R^2 = 1$, which corresponds to SSR = 0
In linear regression, you're trying to create a line that predicts something (like a person's height) based on other things (like their weight). R^2 is like a score that tells you how well your prediction line fits the actual data.
If R^2 is 1, it means your prediction line fits perfectly, explaining all the variations in the data. Imagine you have a scatter plot of points, and your line goes exactly through every point. That's $R^2 = 1$.
Now, SSR is a measure of how wrong your predictions are. If your line is perfect, SSR is 0 because there's no difference between your predicted values and the real values.
So, a perfect fit in linear regression means $R^2 = 1$ and SSR = 0. In simple terms, your prediction is spot on.

3. **In simple linear regression, the value of what shows the point where the estimated regression line crosses the y axis?**

Answer-

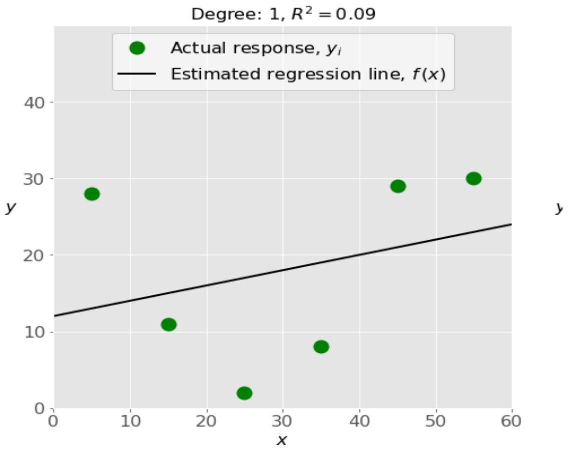
B0
Y=$B0$+$B1X$

- Y is the predicted or estimated value of the dependent variable.
- $B0$ is the y-intercept, which is the point where the regression line crosses the y-axis.
- $B1$ is the slope of the regression line.
- X is the independent variable.

4. **Check out these four linear regression plots:**

Answer-

The top-left plot.
An underfitted model occurs when the model is too simple to capture the underlying patterns in the data. A dataset that shows a clear, non-linear relationship, but your model is a simple straight line (a linear model).



5. There are five basic steps when you're implementing linear regression:

- **a. Check the results of model fitting to know whether the model is satisfactory.**
- **b. Provide data to work with, and eventually do appropriate transformations.**
- **c. Apply the model for predictions.**
- **d. Import the packages and classes that you need.**
- **e. Create a regression model and fit it with existing data.**

However, those steps are currently listed in the wrong order. What's the correct order?

Answer-

d. Import the packages and classes that you need.
b. Provide data to work with, and eventually do appropriate transformations.
e. Create a regression model and fit it with existing data.
a. Check the results of model fitting to know whether the model is satisfactory.
c. Apply the model for predictions.

6. **Which of the following are optional parameters to LinearRegression in scikit-learn?**

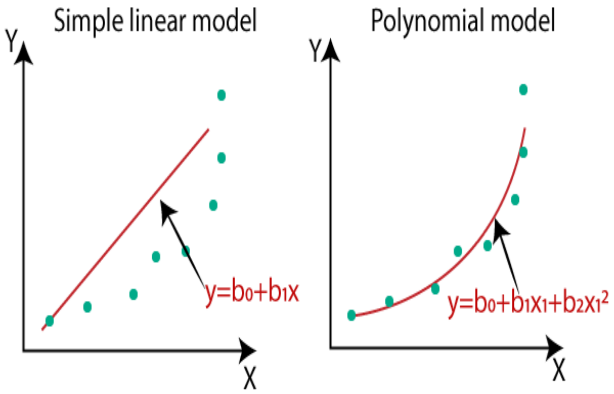
Answer-

In scikit-learn's Linear Regression class, the parameters fit_intercept, normalize, copy_X, and n_jobs are optional.

7. **While working with scikit-learn, in which type of regression do you need to transform the array of inputs to include nonlinear terms such as x^2?**

Answer-

In polynomial regression, we describe the relationship between the independent variable x and the dependent variable y using an nth-degree polynomial in x.



8. **You should choose stats models over scikit-learn when:**

Answer-

You need more detailed results.

Stats models provides extensive statistical details, including p-values and confidence intervals, making it a better choice when you require a deeper analysis of your regression model.

9. **_________ is a fundamental package for scientific computing with Python. It offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. It provides a high-level syntax that makes it accessible and productive.**

Answer-

Numpy

10. **_________ is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics that allow you to explore and understand your data. It integrates closely with pandas data structures.**

Answer-

Seaborn

11. **Among the following identify the one in which dimensionality reduction reduces.**

a) Performance
b) statistics
c) Entropy
d) Collinearity

Answer-

Collinearity.

Collinearity refers to the situation where two or more features in a dataset are highly correlated, and dimensionality reduction methods can help mitigate this issue. By reducing collinearity, dimensionality reduction can improve the stability and interpretability of models, leading to better performance in some cases.

12. **A Which of the following machine learning algorithm is based upon the idea of bagging?**

a) Decision Tree
b) Random Forest
c) Classfication
d) SVM

Answer-

Random Forest

Random Forest is a machine learning algorithm based on the idea of bagging . In Random Forest, multiple decision trees are trained on different subsets of the training data, and the predictions from these trees are combined to make a final prediction.

13. **Choose a disadvantage of decision trees among the following.**

a) Decision tree robust to outliers
b) Factor analysis
c) Decision Tree are prone to overfit
d) all of the above

Answer-

Decision Trees are prone to overfit.

One of the disadvantages of decision trees is that they can easily become too complex and fit the training data too closely, leading to overfitting. This means that the model may perform well on the training data but may not generalize well to new, unseen data.

14. **What is the term known as on which the machine learning algorithms build a model based on sample data?**

Answer-

Training data

Machine learning algorithms learn patterns from the training data to make predictions or decisions on new, unseen data.

15. **Which of the following machine learning techniques helps in detecting the outliers in data?**

Answer-

Anomaly detection

The machine learning technique that specifically helps in detecting outliers in data is "Anomaly detection." Anomaly detection focuses on identifying patterns that deviate from the norm or expected behaviour.

16. **Identify the incorrect numerical functions in the various function representation of machine learning.**

Answer-

Case-based.

17. **Analysis of ML algorithm needs**

Answer-

Both A and B

- **Statistical Learning Theory:** Focuses on understanding the statistical properties of learning algorithms, such as their ability to generalize from a limited set of training data to unseen data.
- **Computational Learning Theory:** Examines the computational aspects of learning algorithms, including their efficiency, complexity, and feasibility in terms of computational resources.

18. **Identify the difficulties with the k-nearest neighbor algorithm.**

Answer-

Both A and B

a) Curse of dimensionality: As the number of features or dimensions increases, the data becomes increasingly sparse, making it more challenging for the algorithm to find meaningful nearest neighbors.
b) Calculate the distance of the test case for all training cases: K-nearest neighbors involves calculating the distance between the test case and all training cases, which can be computationally expensive, especially for large datasets.

19. **The total types of the layer in radial basis function neural networks is ______**

Answer-

2
In a radial basis function neural network, there are typically two types of layers:

1. Input Layer
2. Radial Basis Function Layer

20. **Which of the following is not a supervised learning**

Answer-

PCA

Principal Component Analysis (PCA) is not a supervised learning algorithm; it is an unsupervised learning technique used for dimensionality reduction.