1**.Write the difference between linear regression logistic regression polynomial regression**?

|  |  |  |  |
| --- | --- | --- | --- |
| Aspect | Linear regression | Logistic regression | Polynomial regression |
| Type of output | Continuous | Binary (0 or 1) | Continuous |
| Nature of relationship | linear | Logistic (S-shaped) | Non-linear |
| Equation | Y = mx+ b | P=1/1+e(mx+b) | y = b+m1x+m2x2+…+ mnxn |
| Use case | Predicting scales, house prices | Spam detection, binary classification | Modeling complex relationships |
| Error function | Mean squared error | Log Loss (Cross-Entropy) | MSE or other regression loss |
| Applicability to classification | Not suitable | Binary and multi-class classification | Primarily regression |
| Handling outliers | Sensitive | Less sensitive | Sensitive, especially at higher degrees |
| Computational complexity | Less complex | Moderate complexity | Increases with higher degrees |
| Interpretability | Direct interpretation of coefficients | Log-odds interpretation of coefficients | Complex with higher-degree polynomials |
| Overfitting | Prone to underfitting | Moderate risk, can be controlled | Prone to overfitting, especially at higher degrees |

2. Write 3 examples of datasets?

1. **MNIST Dataset:**
   * **Description:** The MNIST dataset is a collection of 28x28 pixel grayscale images of handwritten digits (0 through 9). It is a popular dataset used in machine learning for training image classification models.
   * **Application:** It is commonly used for developing and testing algorithms for image classification tasks.
2. **IMDB Movie Reviews Dataset:**
   * **Description:** This dataset consists of movie reviews from the Internet Movie Database (IMDb). It includes labeled sentiment for each review, indicating whether the review expresses a positive or negative sentiment.
   * **Application:** It is often used in natural language processing (NLP) tasks such as sentiment analysis, where the goal is to predict the sentiment of a given text.
3. **Iris Flower Dataset:**
   * **Description:** The Iris dataset contains measurements of four features (sepal length, sepal width, petal length, and petal width) for 150 iris flowers, representing three different species: serosa, versicolor, and virginica.
   * **Application:** It is a classic dataset in the field of machine learning and is often used for practicing classification algorithms. The goal is to predict the species of iris based on the provided measurements.
4. Write 10 difference between machine learning classification and machine learning clustering?

Top of Form

1. v**Objective:**
   * *Classification:* The main objective is to predict the class label or category of a given input based on past observations.
   * *Clustering:* The main objective is to group similar data points together based on their inherent patterns or similarities without predefined categories.
2. **Supervision:**
   * *Classification:* It is a supervised learning task where the algorithm is trained on labeled data, meaning the training dataset includes input-output pairs.
   * *Clustering:* It is an unsupervised learning task, and the algorithm works on unlabeled data. There are no predefined output classes; the algorithm discovers patterns on its own.
3. **Output:**
   * *Classification:* The output is a discrete class label or category for each input.
   * *Clustering:* The output is the grouping or clustering of data points, with no predefined labels.
4. **Training:**
   * *Classification:* Requires a labeled training dataset for the algorithm to learn the relationship between inputs and corresponding class labels.
   * *Clustering:* Does not require labeled data; the algorithm identifies patterns and structures in the data without explicit guidance.
5. **Use Cases:**
   * *Classification:* Commonly used in tasks such as spam detection, image recognition, and sentiment analysis.
   * *Clustering:* Used in tasks like customer segmentation, anomaly detection, and grouping similar documents.
6. **Algorithm Types:**
   * *Classification:* Algorithms include decision trees, support vector machines, and neural networks.
   * *Clustering:* Algorithms include k-means, hierarchical clustering, and DBSCAN.
7. **Evaluation:**
   * *Classification:* Performance is often evaluated using metrics such as accuracy, precision, recall, and F1 score.
   * *Clustering:* Evaluation is more subjective and may involve metrics like silhouette score or visual inspection of clusters.
8. **Examples:**
   * *Classification:* Predicting whether an email is spam or not, based on its content.
   * *Clustering:* Grouping customers based on their purchasing behavior without predefined categories.
9. **Handling Outliers:**
   * *Classification:* Outliers are typically considered during training, as they may be part of the labeled dataset.
   * *Clustering:* Outliers can significantly impact clustering results, and some algorithms are more sensitive to them than others.
10. **Applications in Feature Engineering:**

* *Classification:* Feature engineering is crucial to improve the predictive performance of the model.
* *Clustering:* Feature engineering may be less critical, as the algorithm focuses on finding inherent patterns rather than predicting specific outcome

Top of Form