Machine Learning: Clustering vs. Classification and Regression vs. Classification

Clustering vs. Classification

Clustering is an unsupervised learning method that groups a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups.

Classification is a supervised learning approach that involves predicting the category or class of a given input based on labelled training data.

Key difference between them is listed below:

Clustering	Classification
Unsupervised learning	Supervised learning
Groups data based on similarity	Predicts the category of data
No predefined labels	Uses predefined labels
Used for exploratory data analysis	Used for predictive modeling
Algorithm examples: K-means, Hierarchical clustering	Algorithm examples: Logistic Regression, Decision Trees
Outcome is not certain, changes with algorithm	Outcome is a specific class label
Used to understand the structure of data	Used to make decisions based on data
Sensitive to scale of data	May require feature scaling for better performance
Example: Grouping customers based on shopping behavior	Example: Predicting if an email is spam or not

Regression vs. Classification

Regression deals with predicting a continuous output variable based on input variables. It is a type of supervised learning like classification, but whereas classification predicts categorical outcomes, regression predicts numeric outcomes.

Regression	Classification
Predicts a continuous value	Predicts a categorical value
Outcome is a quantity	Outcome is a class label
Example: Predicting house prices	Example: Identifying if a tumor is malignant or benign
Evaluated by MSE, RMSE	Evaluated by accuracy, F1-score
Models include Linear Regression	Models include SVM, Neural Networks
Sensitive to outliers	Less sensitive to outliers
Can have multiple types of regression (linear, polynomial)	Can be binary or multi-class classification
Used in economics, real estate	Used in email filtering, medical diagnoses
Requires scaling of data for better performance	Often requires encoding of labels
Focus on the relationship between variables	Focus on separating data into classes