In Class Activity

Classification vs Regression

HARSH SIDDHAPURA

1230169813

03/28/2024

PART - 1

Problem	Classification	Regression
House Price Prediction	Determine if a house is expensive or affordable. House features (e.g., square footage, location). Binary label (expensive or affordable).	Estimate the sale price of a house. Same features as in classification. Sale price (continuous value).
Natural Language Understanding	Categorize news articles into topics. Text content of articles, metadata. Multi-class labels (e.g., politics, sports, entertainment).	Predict the length of an article. Same text features. Article length (continuous value).
Fault Detection in Manufacturing	Detecting faulty products on the assembly line. Sensor readings, production process data. Binary label (faulty or not).	Estimate the severity of defects. Same features as in classification. Severity score (continuous value).
Employee Attrition	Identify employees likely to leave the company. Employee demographics, performance metrics, job satisfaction. Binary label (attrition or not).	Predict the time until an employee leaves. Same features as in classification. Time until attrition (continuous value).
Click-Through Rate (CTR) Prediction	Predict whether a user will click on an ad. Ad features (e.g., ad content, placement), user behavior (e.g., past clicks). Binary label (click or not).	Estimate the probability of a click. Same features as in classification. Continuous value between 0 and 1 representing the likelihood of a click.
Market Segmentation	Group customers into segments. Customer demographics, behavior. Multi-class labels (segments).	Predict sales volume for each segment. Same features as in classification. Sales volume (continuous value).
Weather Forecasting	Predict rain/no-rain. Historical weather data, satellite images. Binary label (rain or not).	Forecast temperature or precipitation. Same features as in classification. Temperature or precipitation (continuous value).
Anomaly Detection	Detect fraudulent transactions.	Estimate the severity of

	Transaction details, user behavior. Binary label (fraud or not).	anomalies. Same features as in classification. Anomaly severity (continuous value).
Sentiment Analysis	Determine sentiment (positive/negative/neutral) from text. Text reviews, social media posts. Multi-class labels (sentiment categories).	Predict a sentiment score. Same text features. Sentiment score (continuous value).
Medical Diagnosis	Diagnose diseases based on patient data. Medical test results, patient history. Binary labels (disease or not).	Estimate blood sugar levels based on patient features. Same features as in classification. Blood sugar level (continuous value).
Recommendation Systems	Recommend products (e.g., movies) to users. User preferences, item features. Binary relevance (user likes or not).	Predict user ratings for items. Same features as in classification. User rating (continuous value).
Image Recognition	Recognize objects or animals in images. Pixel values of images. Multi-class labels (e.g., cat, dog, car).	Estimate the age of a person from their photo. Facial features, image metadata. Age (continuous value).
Credit Risk Assessment	Determine if a loan applicant is high-risk or low-risk. Applicant's financial history, income, and other relevant factors. Binary label (high-risk or low-risk).	Predict the credit score of an applicant. Same features as in classification. Credit score (continuous value).
Customer Churn Prediction	Identify customers likely to churn (leave a service). Customer behavior, usage patterns, and demographics. Binary label (churn or not).	Predict the time until a customer churns. Same features as in classification. Time until churn (continuous value).
Spam Detection	Predict whether an email is spam or not. Features could include email content, sender information, and metadata. Binary label (spam or not).	Estimate the probability of an email being spam. Same features as in classification. Continuous value between 0 and 1 representing spam probability.

PART - 2

1. In the Spam Detection dataset, each data record represents an individual email. These emails can be categorized into two main classes: legitimate emails, commonly referred to as "ham," and spam emails. Each data record contains information about a specific email, including its content, sender information, metadata, and any other relevant attributes that might help in distinguishing between spam and legitimate emails. The dataset is structured such that each email is represented as a separate data entry, allowing for the analysis and classification of emails based on their characteristics. The goal of the dataset is typically to develop machine learning models or algorithms that can accurately classify incoming emails as either spam or ham, thereby aiding in the task of email filtering and ensuring that users receive only relevant and non-malicious content in their inbox.

2. Some key attributes (features) X of the dataset:
    - **Word Frequencies:** These features capture the occurrence frequency of specific words or phrases in the email content. Words associated with spam (e.g., “free,” “discount,” “urgent”) tend to have higher frequencies.
    - **Characteristics of Sender:** These features relate to the sender’s identity and behavior:
        - **Sender Domain:** The domain from which the email originates (e.g., gmail.com, spammydomain.com).
        - **Sender Reputation:** A measure of how trustworthy or suspicious the sender is based on historical data.
        - **Sender Frequency:** How often the sender appears in the dataset.
    - **Metadata:** Information about the email itself:
        - **Timestamp:** Date and time when the email was sent.
        - **Subject Line:** The subject of the email.
        - **Number of Recipients:** How many recipients received the email.
    - **Length of Email:** The total number of characters or words in the email. Longer emails might exhibit different patterns.
    - **Presence of Attachments or Links:** A binary indicator (1 if attachments or links exist, 0 otherwise). Spammers often include malicious links or attachments.

3. The primary goal is to predict whether an email is spam or not. The target variable (Y) is a binary label which should be estimated:
    - Y = 1: Indicates that the email is spam.
    - Y = 0: Indicates that the email is legitimate (ham).

In summary, the dataset contains email records, each described by various features, and the objective is to build predictive models that accurately classify emails as spam or not. These features help capture patterns indicative of spam behavior, allowing us to make informed predictions.