Introduction to Machine Learning

Week 3 Graded Project

Topics Covered:

Part - I

- **Probability**
- **Bayes Theorem**
- **Normal Distribution**
- **Binomial Distribution**
- **Poisson Distribution**

Part – II

- **Linear Regression**
- **Exploratory Data Analysis**
- **Descriptive Statistics**
- **Log Transformations**
- **Dealing with categorical data**

Part I (25 Points)

Probability and Bayes Theorem (7 points)

Q1. A consumer research survey sampled 200 men to find out whether they prefer to drink plain water or soft drink. 80 out of these 200 men prefer a soft drink. What is the probability that a randomly chosen man will prefer a soft drink? (2 points)

Q2. From a full deck of 52 cards, 1 card is drawn randomly. What is the probability that the card is either a spade or a king? (2 points)

Q3. A drilling company has estimated a 40% chance of striking oil for their new well. A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests. Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful? (3 points)

Normal Distribution (10 points)

Q1. There are 1000 students in a class. The average Score and the Variance of the Score of the class is 240 and 400 respectively. The Scores of the students are normally distributed. Rahul, a student in the class belongs to the 95th percentile in the class. What is the actual Score Rahul has got? (3 points)

Q2. The mean score on a college placement exam is 500 with a standard deviation of 100. Ninety-five percent of the test takers score above what? (3 points)

Q3. A speed-data of some cars is given. The speeds are normally distributed with a mean of 70 km/hr and a standard deviation of 10 km/hr. (4 points)

a) What is the probability that a car picked at random is travelling at more than 100 km/hr?
b) What is the probability that the car speed is between 80 Km / hr and 100 Km / hr

Binomial Distribution (4 points)

Q1. You flip a fair coin 10 times. What is the probability of getting 8 or more heads? (2 points)

Hint - Use stats.binom.pmf() function for this.

Q2. My Bank has a large Credit Card portfolio. Based on empirical data, they have found that 60% of the customers pay their bill on time. If a sample of 10 accounts is selected from the current database, construct the Probability Distribution of accounts paying on time. (2 points)

Poisson Distribution (4 points)

Q1. Assume a poisson distribution with lambda = 5.0. What is the probability that

a) X <= 1?
b) X > 1?

(2 points)

Q2. The number of defects per month in a manufacturing plant is known to follow a Poisson distribution, with a mean of 2.5 defects a month. (2 points)

a) What is the probability that in a given month, no defects occur?
b) That at least one defect occurs?

Hint: Use Poisson distribution equation, find X = 0, Given lambda = 2.5

Part II (25 Points)

Insurance Claim Prediction

Context

A key challenge for the insurance industry is to charge each customer an appropriate premium for the risk they represent. The ability to predict a correct claim amount has a significant impact on insurer's management decisions and financial statements. Predicting the cost of claims in an insurance company is a real-life problem that needs to be solved in a more accurate and automated way. Several factors determine the cost of claims based on health factors like BMI, age, smoker, health conditions and others. Insurance companies apply numerous techniques for analyzing and predicting health insurance costs.

Attribute information:

age : Age of the policyholder
sex: Gender of policyholder
bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight
children: Number of children of the policyholder
smoker: Indicates policy holder is a smoker or a non-smoker (non-smoker=0;smoker=1)
region: The region where the policy holder belongs to (northeast, northwest, southeast, southwest)
claim: Claim amount
bloodpressure: Blood pressure reading of policyholder
diabetes: Suffers from diabetes or not (non-diabetic=0; diabetic=1)
regular_ex: Regularly exercise or not (no-exercise=0; exercise=1)

Steps

1. Import the data and perform the following checks and **write down your insights** at every step. (8 points)
    a. Shape of the data
    b. Data types of attributes
    c. 5-point summary of the relevant attributes
    d. Missing values
    e. Correlation among the attributes
    f. Outliers (display a boxplot)
    g. Remove outliers (using IQR)
    h. Distribution of the target column(“claim”)
2. Transform the column “claim” using log transformation (hint: use np.log(‘column’) and append the transformed column to the dataframe under the column name “log_claim” - optionally you can check the effect of the transformation by plotting histogram of “claim” before and after transformation. (2 points)
3. Encode the categorical variables. In case a column has more than 2 categories, use one-hot encoding. (2 points)
4. Separate out the dependent variable(“claim”) from the independent variables(exclude claim and log_claim from the rest of the variables). (1 point)
5. Split the data into testing and training sets (X_train, y_train, X_test, y_test). (1 point)
6. Train a linear regression model using the training data and print the r_squared value of the prediction on the test data. (2 points)
7. Plot a scatter plot between the actual values and the predicted values for the test set (because plain numbers might not give the entire picture). (1 point)
8. Comment on the performance of the model. (1 point)
9. Repeat steps 4, 5, 6,7 and 8 except, this time use “log_claim” as your dependent variable (note: “claim” cannot be among the predictors). (5 points)
10. Compare the performance of the models trained using the skewed dependent variable as it is and log transformed variable - write your comments and conclude the project. (2 points)

Learning Outcome:

Linear Regression
Exploratory Data Analysis
Descriptive Statistics
Log Transformations
Dealing with categorical data