AccelerateAI

Data Science Global Bootcamp

Data Normalization & Probability: Assignment 02

Instruction: There are 6 questions. Q1 through Q3 are on Encoding/Normalization/Scaling and Q4 through Q6 are on Probability.

Encoding/Normalization/Scaling

Q1: Refer to the dataset in GitHub: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/01%20Python/gapminder.csv

The "Gapminder" dataset contains the health, life expectancy and GDP information for multiple countries categorized by "Region". Use encoding techniques (One hot encoding and Label encoding) for the category column "Region" and provide your solution in the form of Jupyter notebook file. Explain which encoding technique should you use and why?
For Q2 and Q3, please follow below description: Let's refer free "Wine" Dataset that is deposited on the UCI machine learning repository. You can refer to GitHub for this - https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/01%20Python/wine_data_UCI.csv
The Wine dataset consists of 3 different classes/qualities where each row corresponds to a particular wine sample. The "quality" features indicate the class/quality of wine and it is represented as (1, 2, 3) and rest of the columns correspond to 13 different attributes (features).
Consider the features - the wine quality i.e. "quality", "Alcohol" (percent/volume) and "Malicacid" (g/l).
Q2: Do you think feature scaling is required? If yes - Why, If no - why?
Q3: If you feel feature scaling is required, then perform Standardization and Normalization and provide your result. What is the difference you observe between these two methods?
Provide your solution in the form of Jupyter notebook file wherever applicable.

See Python file for solution.

Probability

Q4: Facebook has a content team that labels pieces of content on the platform as spam or not spam. 90% of them are diligent raters and will correctly label 95% of the time. The remaining 10% are non-diligent raters and will label 50% of the content incorrectly. Assume the pieces of content are labeled independently from one another, for every rater. Given that a piece has been rated as non-spam, what is the probability that is it actually non-spam?

In this problem we have to find the marginal probability.

Let P(NS) – probability of non-spam

P(D) – probability of rating diligently = 0.9,

P(ND) – probability of rating non-diligently = 0.1

Given:

P(NS|D) = 0.95, P(NS|ND) = 0.50

Probability of P(NS) = P(NS|D) P(D) + P(NS|ND) P(ND)

= 0.95*0.9 + 0.5*0.1 = **0.905**

Q5: If the probability of seeing a car on the highway in 30 minutes is 0.95, what is the probability of seeing a car on the highway in 10 minutes? (Assume a constant default probability)

This is a bit tricky one. Some of you have considered this to be a repeat of 3 independent event of seeing the car in 10 minutes. However, the question is not about seeing one or more car, it is about seeing 1 car.

This type of phenomenon is governed by Poisson distribution.

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Where λ = r*T r – rate of event, T-time period.

Let r = count of event per unit time (unit time = 10 min)

We are given that probability of observing 1 car in 30 min = 0.95

$\lambda = 3 * r$

P(X=1, $\lambda = 3 * r$) $= \frac{3*re^{-3r}}{1!}$ = 0.95

=> Solving this using numerical method gives us r = 1/3

Probability of observing 1 car in 10 min ($\lambda = r * 1 = 1/3$)

P(X=1, $\lambda = 1/3$) = 1/3 * exp(-1) / 1! = 0.2388

Q6: A machine produces items of which 1% at random are defective. How many items can be packed in a box while keeping the chance of one or more defectives in the box to be no more than 0.5? What are the expected value and standard deviation of the number of defectives in a box of that size?

This is governed by binomial distribution.

Let p – probability that item is defective = 0.01

P(X=k) = $\binom{n}{k} p^k q^{n-k}$

P(one of more defective) = 1- P(zero defective)

$P(k=0) = {}^{n}C_{k}(0.99)^{n} * (0.01)^{0} \quad < \quad 0.5$

⇨ n log(0.99) < log(0.5)
⇨ n < 68.9

The expected value = n*p = 68 * 0.01 = 0.68

The variance = n*p*q = 68*0.01*0.99 = 0.673