Sample Questions for Upcoming Exam of Probability

Topic:

1- Text analysis using multinomial and MLE.

Dataset: A text dataset containing multiple documents with labeled categories.

Instructions:

1. Load the dataset provided.

2. Answer all the questions using Python in a Notebook.

3. Use appropriate libraries such as numpy, pandas, matplotlib, scipy.stats, nltk, and sklearn.

4. Clearly explain your approach in markdown cells where required.

Questions

Section 1: Understanding the Multinomial Distribution

1. Load the text dataset into a pandas DataFrame and display the first five rows.

2. Compute the frequency of unique words in each document.

3. Fit a multinomial distribution to the word frequency data and display the estimated parameters.

Section 2: Maximum Likelihood Estimation (MLE) for Text Data

5. Estimate the parameters of a multinomial distribution using Maximum Likelihood Estimation (MLE) for each document category.

6. Given a new document, compute the likelihood of it belonging to each category using the estimated parameters.

7. Interpret the results and explain how MLE helps in text classification.

Section 3: Inference and Application

11. Analyze the word distributions across different categories and visualize them using bar charts.

12. Comment on how multinomial distribution and MLE can be applied in real-world text analysis problems like spam detection and sentiment analysis.

Sample Questions for Upcoming Exam of Probability

Topic:

1- Multivariate normals

Dataset:

A Kaggle dataset containing patient data with 20+ medical parameters and diagnosis labels.

Instructions:

1. Load the dataset provided.

2. Answer all the questions using Python in a Notebook.

3. Use appropriate libraries such as numpy, pandas, matplotlib, scipy.stats, and seaborn.

4. Clearly explain your approach in markdown cells where required.

Questions

Section 1: Understanding Multivariate Distributions

1. Load the dataset into a pandas DataFrame and display the first five rows.

2. Select any five continuous variables (e.g., Temperature, Glucose, BP, Hb, Uric Acid). Compute and display their covariance matrix.

3. Compute and interpret the correlation coefficients among the selected variables. Visualize the correlation matrix using a heatmap.

4. Explain in one paragraph the difference between covariance and correlation.

Section 2: Multivariate Normal Distribution and Sampling

5. Fit a multivariate normal distribution to the selected five continuous variables. Compute and print the estimated mean vector and covariance matrix.

6. Generate 1000 random samples from the estimated multivariate normal distribution and visualize the distribution using pair plots.

7. Compare the generated samples with the original data distribution. Comment on any noticeable patterns or discrepancies.

Section 3: Probability Computation and Inference

8. Select a random patient from the dataset and calculate the probability of observing their feature values given the estimated multivariate normal model.

9. Assume a new patient has temperature = 98.6, glucose = 120, BP = 130/85, Hb = 14, and uric acid = 5. Compute the probability of this patient's feature vector under the fitted model.

10. Discuss the significance of computing such probabilities in medical diagnosis.

Section 4: Application in Disease Classification

11. Use the computed multivariate normal distribution to analyze whether patients with heart disease or cancer show different distributions for the selected variables. Visualize the comparison using boxplots.

12. Comment on how the difference in distributions could be used in predictive modeling.