**DS201**

**Statistical Programming**

**Assignment 5**

**Ashutosh Kumar Jha**

**12340390**

**DSAI**

**2nd Year**

**Semester 4**

**---------------------------------------------------------------------------**

**Question 1:** **Bivariate Normal Distribution Probability Computation**

**Introduction:** The objective of this study is to compute specific probability values for a bivariate normal distribution with given parameters. The random variables X and Y follow a normal distribution with means, variances, and a correlation coefficient. We aim to determine the probability of certain conditions related to these variables using statistical methods and computational tools.

**Data:** The given parameters for the bivariate normal distribution are:

* Mean of X, μx=3
* Mean of Y, μy=1
* Variance of X, σx2=16(hence, σx=4)
* Variance of Y, σy2=25(hence, σy=5)
* Correlation coefficient, ρxy=35=0.6

Using these parameters, we calculate the following probabilities:

(a) P(3<Y<8)

(b) P(3<Y<8∣X=7)

(c) P(−3<X<3)

(d) P(−3<X<3∣Y=−4)

**Methodology:** The probability calculations are performed using Python's scipy.stats library. The steps involved are:

1. **Computing Marginal Probabilities:**
   * Use the cumulative distribution function (CDF) of the normal distribution to compute the probability of Y and X within the given ranges.
2. **Computing Conditional Probabilities:**
   * For conditional probabilities, the mean and variance of the conditioned variable are adjusted using the properties of the conditional distribution in a bivariate normal distribution.
   * The conditional mean is computed as: μY∣X=μy+ρxy(σyσx)(X−μx) and μX∣Y=μx+ρxy(σxσy)(Y−μy)
   * The conditional standard deviation is computed as: σY∣X=σy1−ρxy2 and σX∣Y=σx1−ρxy2
3. **Using the Normal CDF:**
   * The computed means and standard deviations are used in the normal CDF to find the probabilities within the given range.

**Results:** The computed probabilities are:

* **P(3 < Y < 8):** 0.5178
* **P(3 < Y < 8 | X = 7):** 0.6702
* **P(-3 < X < 3):** 0.3829
* **P(-3 < X < 3 | Y = -4):** 0.1401

**Discussion:**

* The probability of Y being between 3 and 8 is around **51.78%**, indicating that Y has a moderate spread around its mean.
* Given that X=7, the probability of Y being in the range **3 to 8** increases to **67.02%**, showing the effect of conditioning on X.
* The probability that X lies between **-3 and 3** is **38.29%**, which suggests that X has a considerable spread around its mean.
* When conditioned on Y=−4, the probability of X within the range **-3 to 3** decreases to **14.01%**, indicating that extreme values of Y significantly influence X.

**Conclusion:** The results obtained from this study demonstrate the significance of conditioning in a bivariate normal distribution. The correlation between X and Y influences their conditional means and variances, leading to notable differences in probabilities when conditioning on one variable. The computed probabilities provide insights into the distribution of X and Y under different conditions, useful in various statistical and real-world applications involving correlated normal variables.

**Question 2:** Analysis of Probabilities in Bivariate and Multivariate Normal Distributions

**Introduction:** This report presents an analysis of probabilities related to bivariate and multivariate normal distributions. The objective is to compute various probability values using statistical techniques and numerical simulations. The first part of the analysis deals with a bivariate normal distribution, where probabilities related to a given set of parameters are calculated. The second part involves generating samples from a multivariate normal distribution, computing quadratic forms, and evaluating the probability of a specific event.

**Data:** For the bivariate normal distribution, the given parameters are:

* Mean of X: μx=3
* Mean of Y: μy=1
* Variance of X: σx2=16
* Variance of Y: σy2=25
* Correlation coefficient: ρxy=35

For the multivariate normal distribution:

* Dimension: n
* Number of samples: P
* Mean vector: μ
* Covariance matrix: Σ

**Methodology:**

1. **Bivariate Normal Distribution Analysis:**
   * Compute the probability of 3<Y<8 using the cumulative distribution function (CDF) of the normal distribution.
   * Compute the conditional probability P(3<Y<8∣X=7) using the conditional expectation and variance formula.
   * Compute P(−3<X<3) using the normal CDF.
   * Compute P(−3<X<3∣Y=−4) using conditional distribution properties.
2. **Multivariate Normal Distribution Analysis:**
   * Generate P samples from a multivariate normal distribution with given μ and Σ.
   * Compute quadratic form Y=(X−μ)TΣ−1(X−μ).
   * Analyze the distribution of Y for varying n and P.
   * Compute the probability Prob[(x−μ)TΣ−1(x−μ)≤c2] for a given c.

**Results:**

* The probability values for the bivariate normal distribution were successfully computed using normal CDF properties.
* The distribution of Y was visualized using a histogram.
* The probability of the quadratic form satisfying the given inequality was estimated numerically.

**Discussion:**

* The results demonstrate the effectiveness of normal distribution properties in probability computation.
* The quadratic form Y follows a chi-squared distribution, which aligns with theoretical expectations.
* The probability estimation provides insights into confidence regions for multivariate normal distributions.
* Future extensions could involve varying covariance structures and analyzing their effects on probability calculations.

**Conclusion:** This study successfully implemented computational methods to evaluate probabilities related to bivariate and multivariate normal distributions. The obtained results align with theoretical expectations, providing valuable insights into statistical analysis techniques for normal distributions. These methods are applicable in various domains, including data science, machine learning, and financial modeling.

**Question 3: Bayesian Classification of Data Using Multivariate Normal Distributions**

**Introduction:** In this study, we classify data points into two distinct categories using Bayesian classification. The probability distributions of these classes follow a normal distribution, with given mean vectors and covariance matrices. Using Bayes' Theorem, we assign class labels to the data points based on their likelihood under each distribution. This report outlines the implementation and results of this classification approach.

**Data:** The dataset consists of two-dimensional data points stored in "File\_Datapoints.txt." Each data point belongs to one of two classes, whose distributions are defined as follows:

* **Class 1:**
  + Mean: μ1=[2, 3]
  + Covariance Matrix: Σ1=[1, 0.5, 0.5, 2] (2x2, shown in the compressed form)
* **Class 2:**
  + Mean: μ2=[−2, −3]
  + Covariance Matrix: Σ2=[2, −0.3, −0.3, 1] (2x2, shown in the compressed form)

**Methodology:**

1. **Data Loading:** The dataset is read from the text file, extracting two-dimensional feature values.
2. **Probability Computation:** Using the given mean and covariance matrices, we compute the probability density function (PDF) of each class using the multivariate normal distribution.
3. **Bayesian Classification:** We assign a class label to each data point based on the maximum posterior probability. Assuming equal priors for both classes (0.5 each), we classify each point according to: P(Classi∣X)∝P(X∣Classi)P(Classi)/P(X)
4. **Visualization:** The classified data points are plotted in a 2D scatter plot, color-coded based on their assigned class.

**Results:** The classification results are visualized in a scatter plot, where each point is assigned to either Class 1 or Class 2 based on the Bayesian classification approach. The decision boundary is implicitly formed by comparing posterior probabilities.

**Discussion:**

* The classification depends on the spread and overlap of the two distributions.
* If the covariance matrices are similar, the decision boundary will be linear; otherwise, it may be non-linear.
* The method assumes equal priors, but modifying them could change the classification results, favoring one class over the other.
* Misclassified points are likely due to overlapping probability distributions.

**Conclusion:** This report demonstrates Bayesian classification using multivariate normal distributions. The classifier effectively distinguishes data points into two classes based on probability density functions. The approach is useful for real-world applications such as pattern recognition, speech processing, and medical diagnostics, where decision-making relies on probability distributions.

Code: [12340390 Ashutosh Asg5.ipynb](https://colab.research.google.com/drive/1j-dfrgxGJ8y0btjbT6DXz4TqE2d9SiEc?usp=sharing)