LAB MANUAL
UCS654: Predictive Analytics Using Statistics

Experiment–1: Probability Distribution and PDF Fitting

Objective: To generate data and fit a probability density function to understanding data assumptions.

Theory: A Probability Density Function (PDF) represents the likelihood of a continuous random variable. In Machine Learning, many algorithms assume that input data follows a Gaussian (normal) distribution because it improves numerical stability and learning efficiency. PDF fitting helps verify whether such assumptions are valid before applying ML models.

Stepwise Procedure

1. Import required Python libraries.
2. Generate synthetic numerical data using a Gaussian distribution.
3. Compute basic statistics such as mean and standard deviation.
4. Plot histogram to visualize empirical data distribution.
5. Fit Gaussian PDF using statistical parameters.
6. Overlay PDF on histogram and analyze the fit.

Example Python Code [Please do it yourself]

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

data = np.random.normal(loc=50, scale=10, size=1000)

print("Mean:", np.mean(data))
print("Standard Deviation:", np.std(data))

plt.hist(data, bins=30, density=True, alpha=0.6)

x = np.linspace(min(data), max(data), 100)
pdf = norm.pdf(x, np.mean(data), np.std(data))

plt.plot(x, pdf, 'r', linewidth=2)
plt.title("Gaussian PDF Fitting")
plt.show()
```

Observation Table

Parameter	Observation
Number of Samples	
Mean (μ)	
Standard Deviation (σ)	
Shape of Histogram	
Observed Distribution	

Conclusion: The dataset approximately follows a Gaussian distribution as observed from the histogram and fitted PDF. This validates the assumption of normally distributed input features in Machine Learning.

Applications in Machine Learning

- Data normalization and standardization
- Feature preprocessing
- Data drift detection
- Probabilistic modeling