Lecture 3

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Continuing from last class, discussing some examples for maximum likelihood estimation.

Example 1: Exponential Distribution

The ML estimate for the parameter λ turns out to be -

$$\lambda_{ML} = \frac{N}{\sum_i x_i} = \text{Inverse of the sample mean}$$

Sufficient Statistics

The smallest set of statistics that would give us the MLE of a parameter. Test Statistic is simply a function of the samples.

Some examples:

- Sample mean and variance for Gaussian distribution
- Sample mean for exponential distribution
- Max and min for uniform distribution

Non-parametric density estimation

Where a simple distribution doesn't fit the samples.
Can consider a mixture of multiple distributions.

$$p_X(x) = \frac{1}{N}\sum_{i=1}^{N} k(x - x_i),$$

where, $k(x) \geq 0, \forall x$, and $\int_{-\infty}^{\infty} k(x)dx = 1$
The function $k(\cdot)$ is called the kernel function. For all samples we consider their kernel functions and estimate density using the described method.

Assume a Gaussian Kernel

$$k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

Select σ based on a rule of thumb that takes into account the given data (Avoid extreme cases).

1. MLE for parametric distribution

2. Bayesian estimation for parametric distribution

It takes into account a prior belief over parameters contrary to MLE. Many times, it is necessary to take into account previous beliefs instead of relying completely on the data.

Here, Assume a prior belief (distribution) $p_\Theta(\theta)$.
Now, instead of maximizing $\int \theta L_\Theta(X) d\theta$, maximize -

$$\frac{\int \theta p_\Theta(\theta) L_\Theta(X) d\theta}{\int p_\Theta(\theta) L_\Theta(X) d\theta}$$

The denominator term is there to keep the integral to 1.

Multivariate PDF

For a multivariable probability distribution, considering $2D$, $p(x_1, x_2) \geq 0$ and

$$\int \int p(x_1, x_2) dx_1 dx_2 = 1.$$

Also, marginal distribution over a reduced dimension would be -

$$p(x_2) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_1$$

Conditional probability would be calculated as -

$$p(x_2 | x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

where,

$$p(x_1 = a) = \int_{-\infty}^{\infty} p(x_1 = a, x_2) dx_2$$

Multivariate Gaussian

In this multi-dimensional representation (here, considering $2D$), the contours would be elliptical for the random variables covering the two axes. Also, in-place of variance, we would have a covariance matrix, which would help establish relationship between the two r.v.'s,

whether they are independent (correlated), or otherwise.

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix},$$

where $\sigma_{ij} = \sigma_{ji}$.
The σ_{ij}'s will be calculated as -

$$\sigma_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

The Gaussian distribution in this scenario has the following formula -

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k det(\Sigma)}} \exp\left(-\frac{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}{2}\right)$$

where, k is the data dimension, Σ is the covariance matrix and μ is the mean vector.

Exploratory Data Analysis

```
EDA is about taking stock of data
```

Things that we will check in data

- The entire data
- Each variable
- Pair of variables

Type of questions about each variable

- Type and coding
  - Nominal
  - Ordinal
  - True numerical
- Distribution
  - Descriptive statistics
  - Histograms
- Utility and ethics
  - Variability
  - Availability

Integers can be used to code nominal, categorical, ordinal, numerical and temporal data. In EDA, try to identify the description of discrete variables, such as the list of unique values, or order of values, etc.

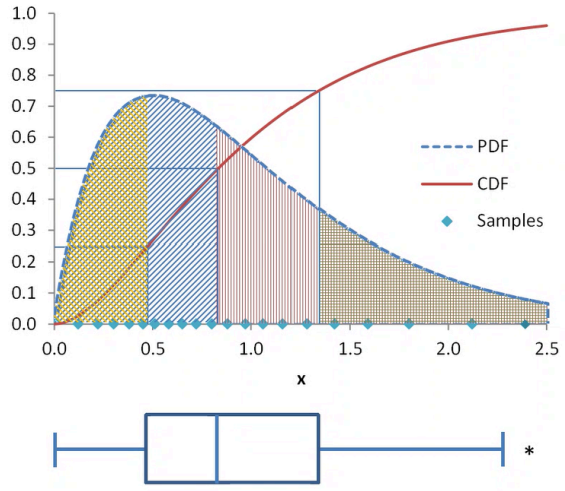
Data representations, such as histograms can help indicate anomalies, problems in the data.

Now, continuous variables are represented using PDFs, then the histogram divides the range into discrete **bins** for counting samples. It is similar to a stepwise approximation of a continuous distribution.

Mean is the Center-of-Gravity and Median divides the PDF into equal areas. Quartiles divides the PDF into four equal areas.

Right-Skewed (i.e. tail is longer on the right side of the peak) and *Left-Skewed* data.

Box and Whiskers plot summarizes the PDF. The three vertical lines in the box represent the 25th (***Q1***), 50th (***Q2*** `a.k.a` median) and 75th (***Q3***) percentile of the data. For the following image, it clearly shows that the data is right-skewed.



The Inter-Quartile Range (***IQR***) is defined as -

$$IQR = Q_3 - Q_1$$

> From above the upper quartile (***Q3***), a distance of 1.5 times the IQR is measured out and a whisker is drawn up to the largest observed data point from the dataset that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured out below the lower quartile (***Q1***) and a whisker is drawn down to the lowest observed data point from the dataset that falls within this distance.

Box plot - Wikipedia

```
Correlation and scatter plots are between pairs of continuous variables
```

Correlation matrix can be computed for all variables together.

EDA using Python

```
import _libraries_
from _libary_ import __module__
```

Some important libraries for EDA (and ML in general)

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy as sp
from sklearn import __module__
```