S21CSE471: Statistical Methods in AI **Dated: March 1, 2021**

Basics of stats and Gaussian Distribution

Prepared by: Akshat Gupta, Anish Gupta, Aritra Banerjee, and Ayush Rai

1 Mean

The mean of a set of values or measurements is the sum of all the measurements divided by the sum of all the measurements in the set:

$$\text{Mean } = \frac{\sum_{i=1}^{n} x_i}{n}$$

If we compute the mean of the population, we call it the parametric or population mean, denoted by μ (read "mu"). If we get the mean of the sample, we call it the sample mean and it is denoted by (read "x bar").

Class	Frequency (f_i)	Mid - point (x_i)	$f_i x_i$
$30 - 40$	3	35	$35 \times 3 = 105$
$40 - 50$	7	45	$45 \times 7 = 315$
$50 - 60$	12	55	$55 \times 12 = 660$
$60 - 70$	15	65	$65 \times 15 = 975$
$70 - 80$	8	75	$75 \times 8 = 600$
$80 - 90$	3	85	$85 \times 3 = 255$
$90 - 100$	2	95	$95 \times 2 = 190$
$\sum f_i x_i = 3100$	$\sum f_i = 50$		$\sum f_i x_i = 3100$

$\sum f_i \equiv 50$

$(\bar{x}) \equiv \frac{\sum f_i x_i}{\sum f_i}$

Mean $(\bar{x})$

$\equiv \frac{3100}{50}$

$\equiv 62$

2 Variance

Variance is a measure of dispersion around the mean and is statistically defined as the average squared deviation from the mean. It is noted using the symbol σ^2.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Where μ is the population mean and N is population size. The standard deviation, σ, is the square root of the variance and is commonly referred to as the volatility of the asset.Essentially it is a measure of how far on average the observations are from the mean. A population's variance is given by:

The population standard deviation equals the square root of population variance. The sample variance is given by:

$$S^2 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{n-1}$$

Finding Variance and Standard Deviation

Class	Frequency (fi)	Mid - point (x_i)	$(x_i - \bar{x})^2$	$f_i\,(x_i - \bar{x})^2$
$30 - 40$	3	35	$(35 - 62)^2 = 729$	$3 \times 729 = 2187$
$40 - 50$	7	45	$(45 - 62)^2 = 289$	$7 \times 289 = 2023$
$50 - 60$	12	55	$(55 - 62)^2 = 49$	$12 \times 49 = 588$
$60 - 70$	15	65	$(65 - 62)^2 = 9$	$15 \times 9 = 135$
$70 - 80$	8	75	$(75 - 62)^2 = 169$	$8 \times 169 = 1352$
$80 - 90$	3	85	$(85 - 62)^2 = 529$	$3 \times 529 = 1589$
$90 - 100$	2	95	$(95 - 62)^2 = 1089$	$2 \times 1089 = 2187$
$\sum f_i = 50$			Sum $= 10050$	

$\sum f_i\,(x_i - \bar{x})^2 = 10050$

$\sum f_i = 50$

3 Covariance

Covariance formula is a statistical formula which is used to assess the relationship between two variables. In simple words, covariance is one of the statistical measurement to know the relationship of the variance between the two variables.

The covariance indicates how two variables are related and also helps to know whether the two variables vary together or change together. The covariance is denoted as Cov(X,Y) and the formulas for covariance are given below.

Formulas for Covariance (Population and Sample)

Population Covariance Formula	$\mathrm{Cov}(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N}$
Sample Covariance Formula	$\mathrm{Cov}(X,Y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$

Notations in Covariance Formulas

- $\mathrm{x_i}$ = data value of x
- y_1 = data value of y

$-\bar{x}$ = mean of x

$-\bar{y}$ = mean of y

- N = number of data values.

Lets see an example problem to understand the notations and the formulas in a better way.

Xi	yi	Xi $-\bar{x}$	$yi - \bar{y}$
2.1	8	-1	-3
2.5	12	-0.6	1
4.0	14	0.9	3
3.6	10	0.5	-1

$C_{ov}(x,y) = \frac{(-1)(-3)+(-0.6)1+(0.9)3+(0.5)(-1)}{4-1} = \frac{3-0.6+2.7-0.5}{3} = \frac{4.6}{3} = 1.533$

Relation Between Correlation Coefficient and Covariance Formulas

Correlation $\equiv \mathrm{Cov}(x,y)\frac{}{\sigma_w * \sigma_y}$
Here, Cov (x,y) is the covariance between x and y while σ_x and σ_y are the standard deviations of x and y. Using the above formula, the correlation coefficient formula can be derived using the covariance and vice versa. Correlation ranges between +1 and -1 and is, therefore, much easier to interpret than covariance. Two variables are perfectly correlated if their correlation is equal to +1, uncorrelated if their correlation is equal to 0, and move in perfectly opposite directions if their correlation is equal to -1.

4 Expectation

If you have a collection of numbers $a_1, a_2, \ldots, a_N$, their average is a single number that describes the whole collection. Now, consider a random variable X. We would like to define its average, or as it is called in probability, its expected value or mean. The expected value is defined as the weighted average of the values in the range. Expected value (= mean=average):

Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, \ldots\}$ (finite or countably infinite). The expected value of X_1, denoted by EX is defined as

$$E(X) \equiv \sum_{x_k \in R_X} x_k P\left(X \equiv x_k\right) \equiv \sum_{x_k \in R_X} x_k P_X\left(x_k\right)$$

Lets understand the above formula with an example. The following information is the probability distribution of successes.

$No.of Success$	0	1	2
$Prob.of success$	$\frac{6}{11}$	$\frac{9}{22}$	$\frac{1}{22}$

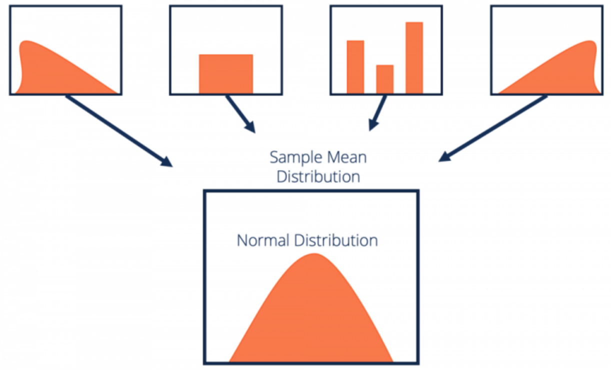
Expected number of success is $E(X) \equiv \sum_x x P_X(x)$

$$\begin{aligned}
&\equiv \left(0 \times \tfrac{6}{11}\right) + \left(1 \times \tfrac{9}{22}\right) + \left(2 \times \tfrac{1}{22}\right) \\
&\equiv \tfrac{11}{22} \\
&\equiv 0.5
\end{aligned}$$

Therefore, the expected number of success is 0.5 .

5 Central Limit Theorem

Let's put a formal definition to CLT:
CLT states that if you have a population with mean μ, sd σ, and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be normally distributed. Given a dataset with unknown distribution (it could be uniform, binomial or completely random), the sample means will approximate the normal distribution.



These samples should be sufficient in size. The distribution of sample means, calculated from repeated sampling, will tend to normality as the size of your samples gets larger.

For a population(n) if " X " has finite mean μ and sd σ, CLT is defined by,

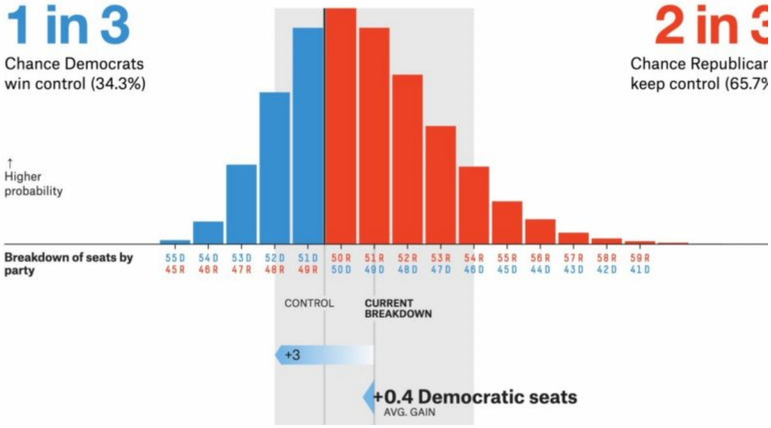
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ when } n \to +\infty$$

where the sample mean and sd is,

$$\mu_{\bar{x}} \equiv \mu$$
$$\sigma_{\bar{x}} \equiv \frac{\sigma}{\sqrt{n}}$$

So the average of the sample means will be approximate to the population mean(μ), and the sd(o) will be the average standard error.

6 Practical Applications of CLT -



1. Political/election polls are prime CLT applications. These polls estimate the percentage of people who support a particular candidate. You might have seen these results on news channels that come with confidence intervals.

2. The central limit theorem helps calculate that Confidence interval, an application of CLT, is used to calculate the mean family income for a particular region

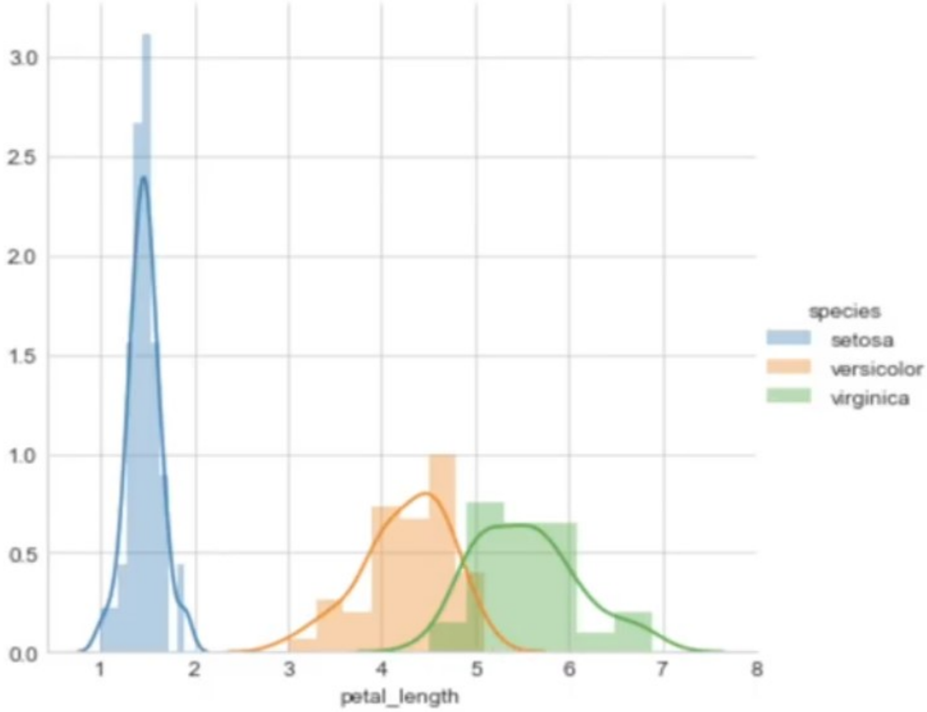
7 Gaussian Distribution

7.1 Introduction

Before understanding the Gaussian Distribution, we need to understand a few basic concepts, which will be used while we delve into Gaussian Distribution

7.1.1 Probability Density Function

In the simplest of words, probability density function (PDF) is a smoothed out histogram.One may remember that a histogram is a plot that displays the values of the variable that one is interested in, on x axis, and on y axis, it displays, how often that data point occurs in the data.



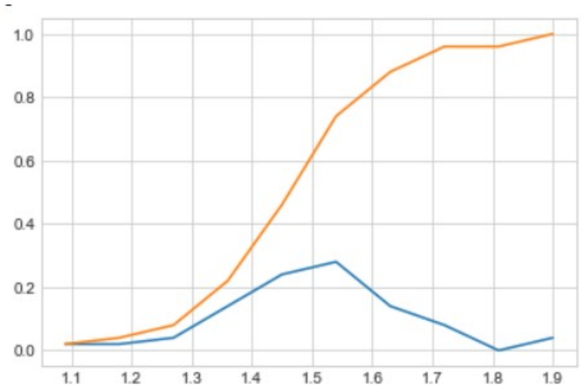
Above figure is a representation between histogram and probability density function. The dataset used is the famous IRIS dataset. The rectangular boxes on the plot are the histogram, and the smooth curve overlapping it, is the probability density function.
The PDF is the density of probability. It tells us how dense is the probability for a particular region, i.e., higher the PDF curve at a region, more the probability of occuring of that datapoint. The concept is very similar to mass density in physics: its unit is probability per unit length.

7.1.2 Cumulative Distribution Function

The cumulative distribution function of a random variable X is defined as

$$F_X(x) = P(X \leq x), \text{for all x} \in R$$

The cumulative distribution function (CDF) $F_X(x)$ describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x. That is, for a given value x, $F_X(x)$ is the probability that the observed value of X is less than or equal to x. Take for example the below figure.



In the above figure, the blue curve represents the probability density function, and the curve in orange represents its cumulative distribution function. This has been taken from the actual dataset IRIS. Let us assume that the variable on x-axis, for which we are plotting the curve is V. Take for example, the value of V = 1.6.
The value corresponding to V = 1.6 in the PDF is roughly 0.2, and the corresponding value in the CDF is roughly 0.82. This tells us that in the dataset, 82% of the datapoints in the dataset, have the value V which is less than equal to 1.6.
Basically, CDF is the area under PDF curve till that point.

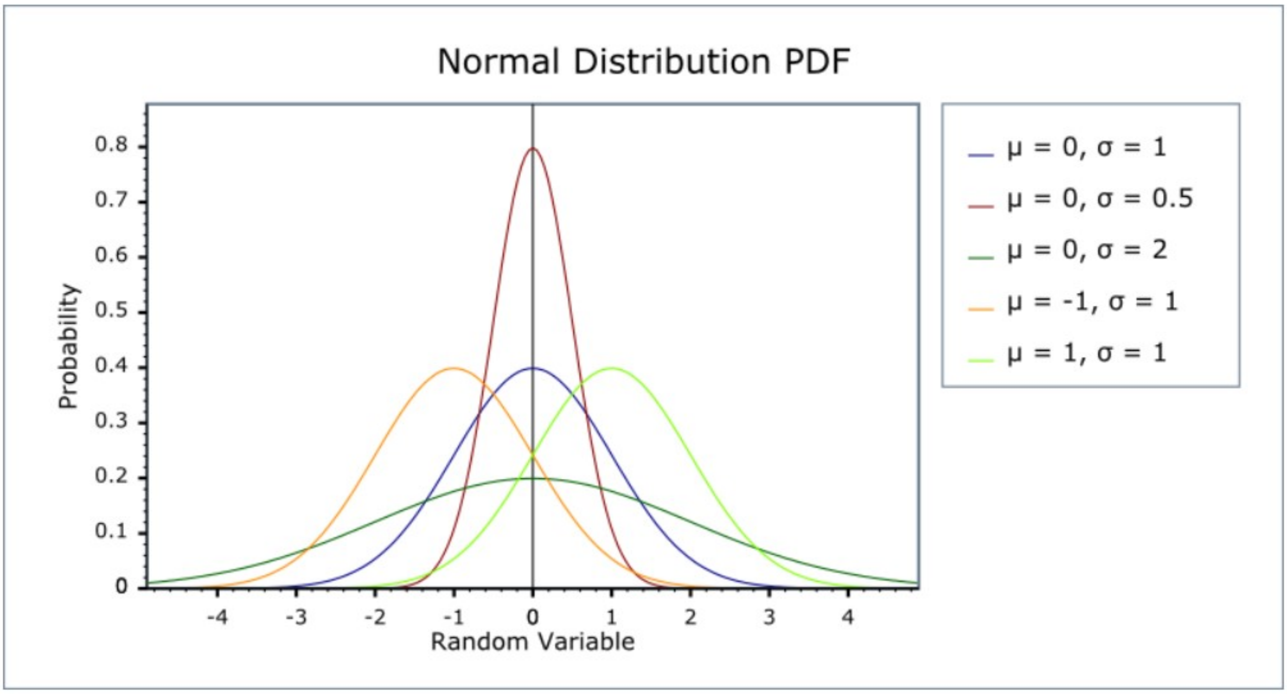
7.2 Normal (Gaussian) Distribution

The normal distribution is by far the most important probability distribution. One of the main reasons for that is the Central Limit Theorem (CLT). CLT states that if you add a large number of random variables, the distribution of the sum will be approximately normal under certain conditions. The importance of this result comes from the fact that many random variables in real life can be expressed as the sum of a large number of random variables and, by the CLT, we can argue that distribution of the sum should be normal.
A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, if its PDF is given by

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

where μ and σ have usual meanings.

Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the central limit theorem which states that, under some conditions, the average of many samples (observations) of a random variable with finite mean and variance is itself a random variable—whose distribution converges to a normal distribution as the number of samples increases. Therefore, physical quantities that are expected to be the sum of many independent processes, such as measurement errors, often have distributions that are nearly normal.
The PDF of a Normal distribution is a bell shaped curve.

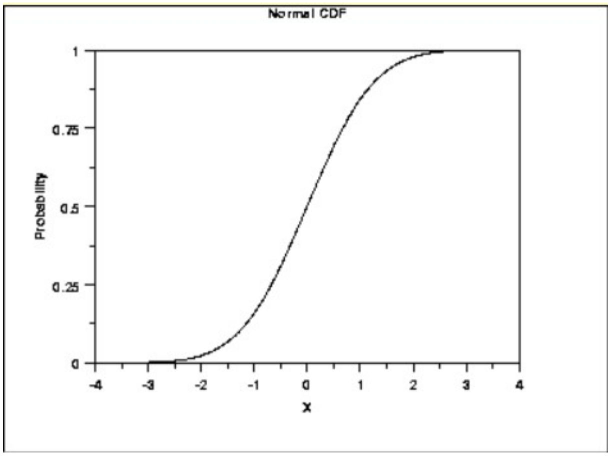


7.2.1 CDF of Normal distribution

The CDF of the standard normal distribution is denoted by the ϕ function:

$$\phi(x) = P(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du$$

This integral does not have a closed form solution. Nevertheless, because of the importance of the normal distribution, the values of $F_Z(z)$ have been tabulated and many calculators and software packages have this function.



8 Anomaly Detection

This kind of problems deal with the identification of outliers(anomalies) in the data relative to some standard or general outcome. It has its application across industries. This kind of problems mainly arise when there is a **large imbalance** between *postive* and *negative* data for a situation. Some common examples of anomaly detection are:

1. Fraud detection in an online transaction
2. Unexpected growth in the number of users in a website which looks like a spike
3. Sensors detecting faulty components at a manufacturing plant

8.1 Basic Approach

Anomaly Detection tried to define a boundary around the *normal data points* by fitting a Gaussian distriution to it. Then based on this distribution, we can compute the probability of whether a certain data point(situation) can be described as normal or not. However, several factors affect this approach.
Anomaly detection problem uses is implemented through two kinds of Gaussian distributions depending on the situation in which they perform better.

1. Univariate Gaussian normal distribution model
2. Multivariate Gaussian normal distribution model

8.2 Univariate Gaussian Distribution Model

This Gaussian model is described through parameters x, μ, and σ. x is the feature matrix, μ is the mean and σ is the covariance matrix. These are the steps in the algorithm:

1. x_i are the data points (some of them should be anomalous examples)
2. Fit parameters μ_1,.....,μ_n and σ_1^2,.....,σ_n^2

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_i^j \tag{1}$$

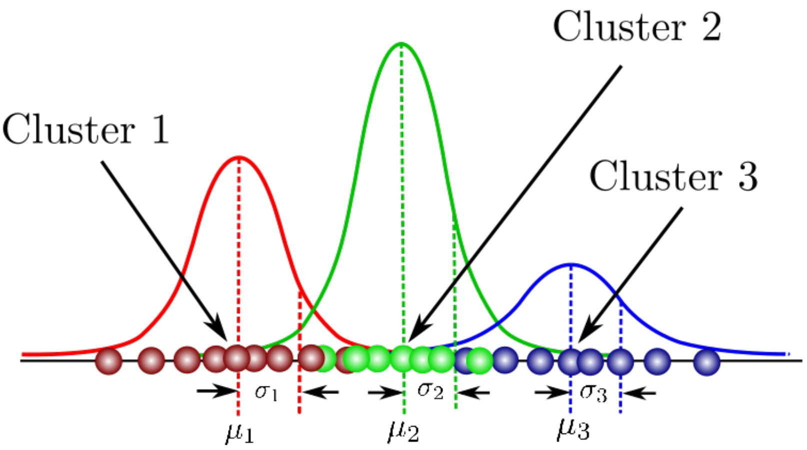
$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^i - \mu_j)^2 \tag{2}$$

3. Given new example x, compute $p(x)$

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j}} \exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2} \tag{3}$$

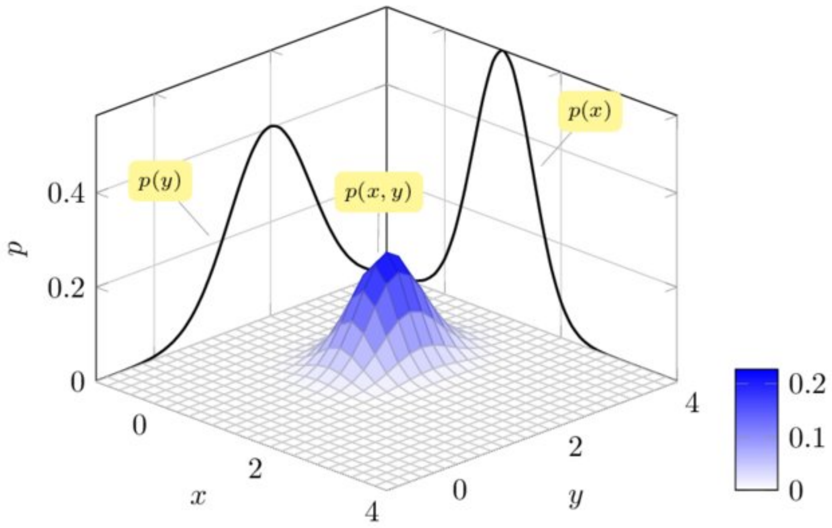
4. Anomaly if $p(x) < \epsilon$

If the number of factors/features is 1, then they can be simply modeled by a normal gaussian distribution:

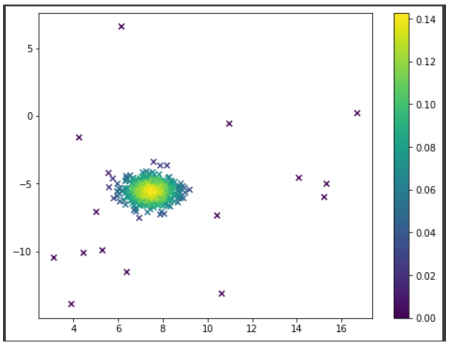


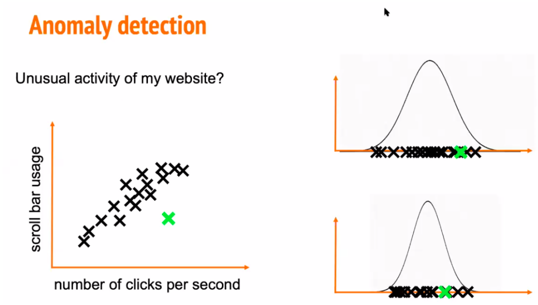
Suppose if the number of features are 2(or more), in that case, we can use the **INDEPENDENCE ASSUMPTION** and convert them into two separate individual Gaussian distributions. In that way, we can better find out distributions to fit the data and anomalies can be found out by multiplying the probabilities from the individual distributions. In short:

$$p(x, y) = p(x) \times p(y) < \epsilon \implies \textbf{anomaly} \tag{4}$$



However, this method does not work in all cases. Whether the **independence assumption** is applicable or not, can be understood from the shape of the cluster. If the features were independent then the plot would look more like a circle (to show value of one feature is completey independent of the other. For only one cluster, this assumption holds holds for a few cases:

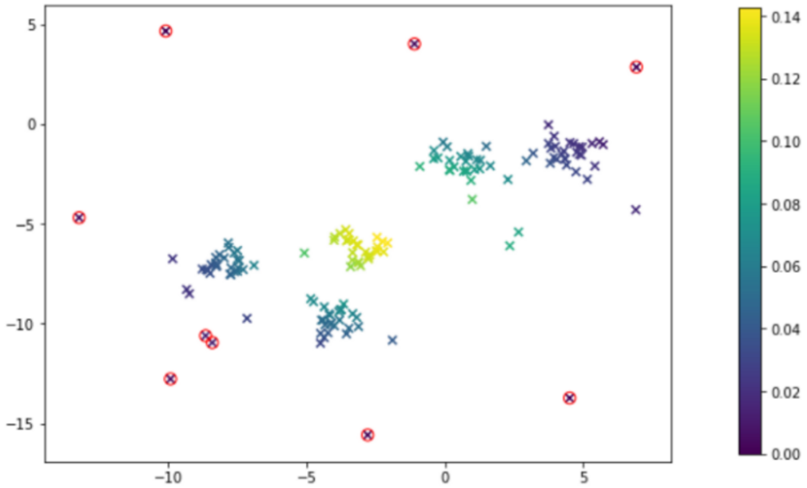




Anomaly detection

Unusual activity of my website?

In the above images, we see that the image on the left is a case where the univariate distribution works and the image on the right is where this does not work.
One more case where this univariate analysis fails is when there are multiple clusters.



Code for the this example can be found in this **link** Courtesy: Reference link 4.

Due to these reasons, for **multiple features** and for **multiple clusters**, we should be using **Multivariate Normal Disribution**. Which will be discussed in the next section.

9 Multivariate Density Estimation)

9.1 Density Estimation

Standard Defination In probability and statistics, density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function. The unobservable density function is thought of as the density according to which a large population is distributed; the data are usually thought of as a random sample from that population.

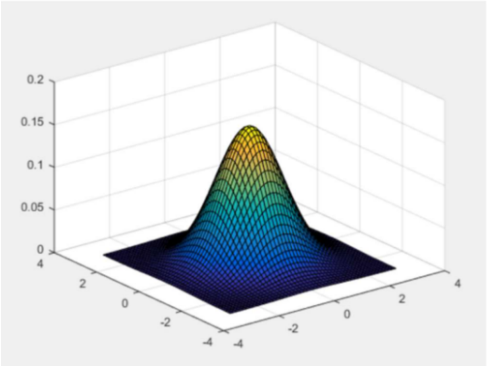
A variety of approaches to density estimation are used, including Parzen windows and a range of data clustering techniques, including vector quantization. The most basic form of density estimation is a rescaled histogram.

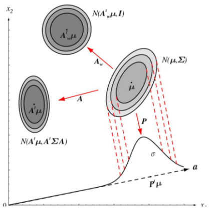
9.2 Multivariate Normal Density Estimation

In case of a normal distribution, we can use the following formula for density estimation:

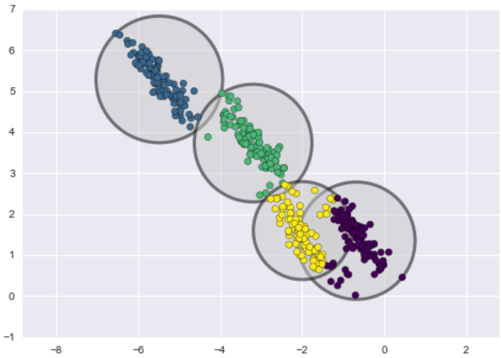
$$p(x) = \frac{1}{(2\pi)^{n/2}|\sum|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T(\sum^{-1})(x-\mu))$$

(5)

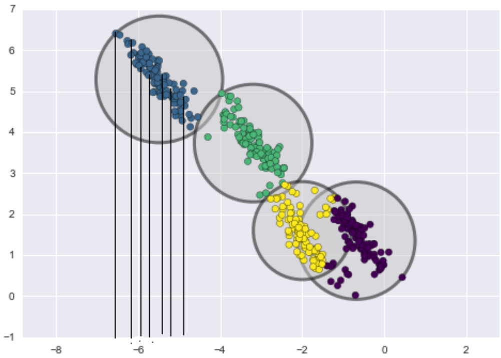




The countours formed represents the induvidual gaussians which are present the figure. The can be estimated by taking some samples and clustering. After that, we calculate their individual parameters by calculating the mean and other statistical paramteres of the individual clusters.



In case of the above image, we project all the samples of a given cluster on the x axis.



As we are approximating individual clusters to be gaussian, we know $p(\frac{x}{y_i)})$. Using the baysian Baysian decision rule, we can caluclate: $p(\frac{y_i}{x)})$ as:

$$p(\frac{y_i}{x)}) = \frac{p(\frac{x}{y_i)p(y_i)})}{p(x)}$$

(6)

10 References

1. https://medium.com/analytics-vidhya/central-limit-theorem-and-machine-learning-part-1-af3b65

2. www.probabilitycourse.com

3. https://udohsolomon.github.io/_posts/2017-09-12-Anomaly-detection/

4. https://towardsdatascience.com/understanding-anomaly-detection-in-python-using-gaussian-mixt

5. https://www.boost.org/doc/libs/1_49_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/normal_dist.html