Chapter #01: Generative Model

Preliminaries ----------

Note: [**Starts**]

Expected value (E): Expected value is the value that we expect to see on an average.

In probability theory, the expected value for a random variable X with possible outcomes $\{x_1, x_2, \ldots, x_n\}$ and with respective probabilities $\{p_1, p_2, \ldots, p_n\}$ can be calculated as:

$$E(X) = \sum_{i=1}^{n} x_i p_i = \mu$$

So, the above Equation is the
Expected Value,
Mean, or
Average
of the random variable X.

For example, for a fair die if all six sides are equally likely (i.e., the probability of the outcome of any one side is $1/6$), then the expected value can be computed as:

$$E = (1 \times \frac{1}{6}) + (2 \times \frac{1}{6}) + (3 \times \frac{1}{6}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) = \frac{21}{6} = 3.5$$

Therefore, if we play with the fair die for a long time and collect the outcomes and then take the average of the outcomes, theoretically, the value will be 3.5.

More generally, if $f(x)$ is any function of x, the expected value of f is defined by:

$$E[f(x)] = \sum_{x \in X} f(x) P(x)$$

The process of forming an expected value is linear, in that if α_1 and α_2 are arbitrary constants, then we can write:

$$E[\alpha_1 f_1(x) + \alpha_2 f_2(x)] = \alpha_1 E[f_1(x)] + \alpha_2 E[f_2(x)]$$

Variance, $$Var[x] = \sigma^2 = E[(x-\mu)^2] = \sum_{x \in X} (x-\mu)^2 P(x)$$

Also, this is sometimes very useful to note that:

$$\underline{Var[x] = E[x^2] - (E[x])^2}$$

How? **Ans**:

$$Var[x] = \sum_{x \in X} (x-\mu)^2 P(x)$$

$$= (x_1-\mu)^2 P(x_1) + (x_2-\mu)^2 P(x_2) + ... + (x_n-\mu)^2 P(x_n)$$

$$= [x_1^2 P(x_1) + x_2^2 P(x_2) + ... + x_n^2 P(x_n)]$$
$$- 2\mu[x_1 P(x_1) + x_2 P(x_2) + ... + x_n P(x_n)]$$
$$+ \mu^2[P(x_1) + P(x_2) + ... + P(x_n)]$$

$$= E[x^2] - 2\mu\, E[x] + \mu^2 \qquad [\because P(x_1) + P(x_2) + ... + P(x_n) = 1 \;\;]$$
$$= E[x^2] - 2\mu\,\mu + \mu^2 \qquad [\because E[x] = \mu\,]$$
$$= E[x^2] - 2\mu^2 + \mu^2$$
$$= E[x^2] - \mu^2$$
$$= E[x^2] - (E[x])^2 \qquad [\because E[x] = \mu\,]$$

Note: we can rearrange and write:

$$E[x^2] = (E[x])^2 + Var[x]$$

Similarly:

Mean Squared Error $(MSE) = bias^2 + Var$

[This is shown in 'Bias-Variance Decomposition']

Note, we have already mentioned before that:

The bias of a predictor is the difference between the predictor's expected value and the true value of the parameter. If the difference is zero then the predictor can be called unbiased.

Bias-Variance Decomposition:

Mean Squared Error (MSE): The MSE of a predictor is the mean of the square of the difference between the predicted value and the true value of the output being predicted. The MSE is also regarded as a *Risk Function*.

The MSE of a predictor *f*, when it predicts $\hat{Y} = f(X)$ for *Y*, can be written as,

$$MSE(\hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

We can also write in terms of expected value as:

$$MSE(\hat{Y}) = E[(\hat{Y} - Y)^2]$$

Bias: The bias of a predictor is the difference between the predictor's expected value and the true value of the parameter.

Therefore,

$$Bias = (E[\hat{Y}] - Y)$$

Or,
$$Bias = (\mu - Y) \qquad [\because mean, \mu = E[\hat{Y}]]$$

Now the variance (σ^2) is the square of the standard deviation σ.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \mu)^2} = \sqrt{E[(\hat{Y} - \mu)^2]}$$

$$\sigma^2 = E[(\hat{Y} - \mu)^2]$$

Now, we show, $MSE(\hat{Y}) = E[(\hat{Y} - Y)^2]$

$$= E[((\hat{Y} - \mu) + (\mu - Y))^2]$$

$$= E[(\hat{Y} - \mu)^2 + 2(\hat{Y} - \mu)(\mu - Y) + (\mu - Y)^2]$$

$$= E[(\hat{Y} - \mu)^2] + E[(\mu - Y)^2] \quad [\because E[\mu] = \mu, \mu = E[\hat{Y}], \therefore E(\hat{Y} - \mu) = 0]$$

$$= \sigma^2 + E[(\mu - Y)^2] \qquad [\because Var, \sigma^2 = E[(\hat{Y} - \mu)^2]]$$

$= \sigma^2 + (\mu - Y)^2$ [$\because$ E of a constant will remain the same]

$= \sigma^2 + (Bias)^2$

Therefore,

$$MSE = Var + bias^2$$

The variance measures the extent to which the solutions for individual data sets vary around their average, and hence this measures the extent to which the function $f(X)$ is sensitive to the particular choice of data set.

The bias represents the extent to which the average prediction over all data sets differs from the desired regression function. The MSE can be determined from *Var* and *bias* alternatively.

The following figure depicts the bias, variance relationship with the prediction error as the model complexity increases.

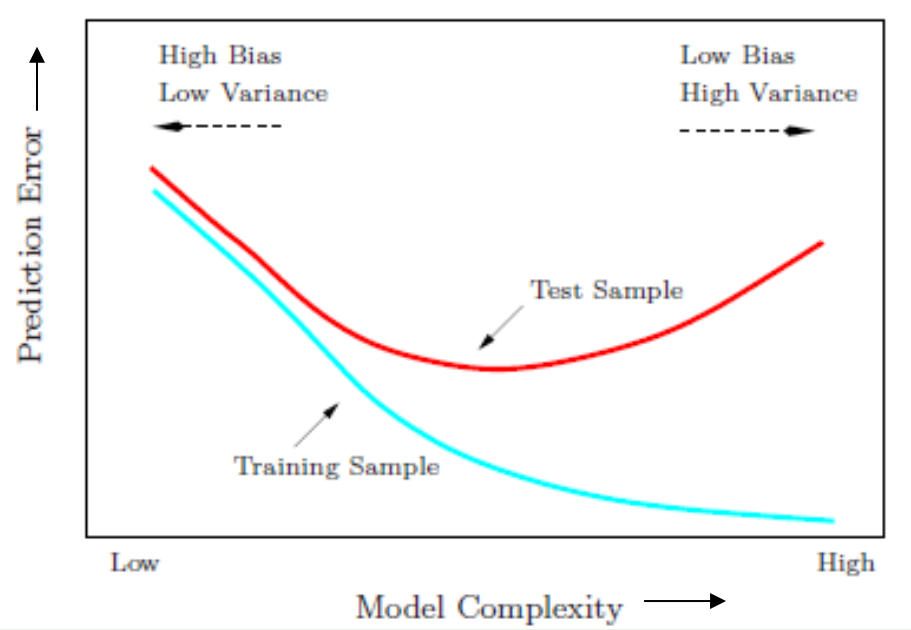


Figure: Test and training error as a function of model complexity.

Joint Probability Distribution

When probability distribution is applied to more than one random variable as a group that can give rise to a joint probability distribution. The formulations of the multivariate distribution can be similarly done as it can be done for the simple bivariate distribution.

Given two random variables *X* and *Y,* the joint distribution can be expressed as:

$$P(x, y) = P(X= x, Y= y)$$

$$P(x, y) \geq 0 \text{ and } \sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

assuming that the events *x* and *y* happened at the same time. The random variables can be discrete as well as continuous. Further, they could be independent as well as dependent.

The (*x, y*) pair can be thought of as a vector or a point in the product space of *x* and *y*.

We can write, $P(x, y) = P(x)P(y)$, when *x* and *y* are statistically independent.
We can write, $P(x, y) = P(x \mid y)P(y)$, when *x* and *y* are statistically dependent.

Expected values of Function of two variables

$$E[f(x, y)] = \sum_{x \in X} \sum_{y \in Y} f(x, y)P(x, y)$$

Also, $$E[\alpha_1 f_1(x, y) + \alpha_2 f_2(x, y)] = \alpha_1 E[f_1(x, y)] + \alpha_2 E[f_2(x, y)]$$

The means (first moment) and variances (2nd moments) are:

$$\mu_x = E[x] = \sum_{x \in X} \sum_{y \in Y} x\, P(x, y)$$

$$\mu_y = E[y] = \sum_{x \in X} \sum_{y \in Y} y\, P(x, y)$$

$$\sigma_x^2 = Var[x] = E[(x - \mu_x)^2] = \sum_{x \in X} \sum_{y \in Y} (x - \mu_x)^2\, P(x, y)$$

$$\sigma_y^2 = Var[y] = E[(y-\mu_y)^2] = \sum_{x\in X}\sum_{y\in Y}(y-\mu_y)^2\, P(x,y)$$

A cross-moment, i.e., the covariance of *x* and *y* can be defined as:

$$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] = \sum_{x\in X}\sum_{y\in Y}(x-\mu_x)(y-\mu_y)\, P(x,y)$$

The covariance is one measure of the degree of statistical dependence between *x* and *y*. If *x* and *y* are statistically independent and then $\sigma_{xy} = 0$. So if $\sigma_{xy} = 0$ then the variables *x* and *y* are said to be uncorrelated.

Also, remember, the *Pearson Correlation Coefficient*, $pcc = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

By the way, using vector notation, we can rewrite the previous two sets of equations as:

Mean, $\boldsymbol{\mu} = E[\mathbf{x}] = \sum_{\mathbf{x}\in\{XY\}} \mathbf{x} P(\mathbf{x})$

Variance, $\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$, the outcome is a covariance matrix (and square matrix as well).

(Later in this chapter, we will see the vector notation)

Continuous Random variables

When the random variable *x* can take values in the continuum, it makes sense to talk about the probability that the *x* falls in some interval (*a*, *b*) instead of about its particular value. This is because the probability of any particular exact value is almost always will be zero. Here instead of using a probability mass function *P*(*x*), we use probability density function *p*(*x*).

$$\Pr[\, x \in (a,b)\,] = \int_a^b p(x)dx$$

Here, the probability density function (pdf), $p(x)$, must satisfy:

$$p(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} p(x)dx = 1$$

In general, most of the definitions and formulas for discrete random variables carry over to continuous random variables. The sums symbol is replaced by the integrals. We can have the following for example:

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$$

$$\mu = E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

$$Var[x] = \sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx$$

Also, $\sigma^2 = E[x^2] - (E[x])^2$ holds true.

Bayes Rule

The *Law of the Total Probability* states that if an event A can occur in *n* different ways: $A_1, A_2, ..., A_n$ and if these *n* subevents are mutually exclusive, i.e., they cannot occur at the same time, then the probability of A occurring is the sum of the probabilities of the subevents A_i. Specifically, the random variable y can assume the value of y in *n* different ways with *x*: $x_1, x_2, \ldots, x_n$.

Because these probabilities are mutually exclusive, it follows from the *Law of the Total Probability* that $P(y)$ is the sum of the joint probability $P(x, y)$ over all possible values of *x*. That is,

$$P(y) = P(y \mid x_1)P(x_1) + P(y \mid x_2)P(x_2) + ... + P(y \mid x_n)P(x_n)$$

Note: An Example: Both Tamiflu and Relenza are drugs that work against influenza. Let us assume, our study at UNO found that Tamiflu remains effective for 12 months in 80% of cases, and Relenza remains effective for 12 months in 90% of cases. If UNO purchases 30% Tamiflu and 70% Relenza to vaccinate its staff, how likely is it that a staff will remain influenza-free for 12 months?

Using the law of total probability we can write:

$P(y) = P(y|x_1)\, P(x_1) + P(y|x_2)\, P(x_2)$
$= (0.80)\ (0.30) + (0.90)\ (0.70)$
$= 0.87$

Therefore, there is 87% chance that a staff will remain influenza-free for 12 months.

Note [ENDs]

Therefore, we can write formally:

$$P(y) = \sum_{x \in X} P(x, y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(A)}$$

We also know from the definition of the conditional probability $P(y \mid x)$ we have:

$$P(x, y) = P(y \mid x)P(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(B)}$$

Following Equation (B), we can write:

$$P(x \mid y)P(y) = P(x, y)$$

Or,
$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

Or,
$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in X} P(x, y)}$$ [using Eqn (B) of numerator and (A) for denominator]

Or, using (B) for the denominator:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in X} P(y \mid x)P(x)}$$

The above Equation is called the *Bayes rule*, *Bayes law*, *or Bayes theorem.* The different parts of it are also known as following:

$$posterior = \frac{likelihood \times prior}{evidence}$$

Practically, $posterior \propto (likelihood \times prior)$

Example: Suppose we have invented an influenza tester. Now, say 5% of the population is sneezing due to the cold season. Running test on sneezing people, 90% test returned positive for influenza testing. Given a person is NOT sneezing, the test comes out positive 15% of the time. For a positive detection, what is the chance that the person is sneezing?

Here,

- 5% of the population is **S**neezing, $p(S)$=0.05 [*prior*]

- The test returns **T**rue given a person **S**neezing for 90% cases, $p(T \mid S)$=0.9 [*likelihood*]
- The test is True for non-Sneezing people in 15% cases, $p(T \mid \bar{S})$=0.15

We are calculating the chance that given True or positive detection of influenza, what is the change that the person is sneezing, i.e., $p(S \mid T)$=? [*posterior?*]

Following Bayes rule, we can write,

$$p(S \mid T) = \frac{p(T \mid S)p(S)}{p(T)}$$

$$= \frac{p(T \mid S)p(S)}{\sum_S p(T \mid S)p(S)}$$

$$= \frac{p(T \mid S)p(S)}{p(T \mid S)p(S) + p(T \mid \bar{S})p(\bar{S})}$$

$$= \frac{p(T \mid S)p(S)}{p(T \mid S)p(S) + p(T \mid \bar{S})[1 - p(S)]}$$

$$= \frac{0.90 \times 0.05}{0.90 \times 0.05 + 0.15 \times [1 - 0.05]}$$

$$= \frac{0.90 \times 0.05}{0.90 \times 0.05 + 0.15 \times 0.95}$$

$$= \frac{0.045}{0.1875} = 0.24 \text{ or, } 24\% \text{ chance.}$$

Finding a person positive for influenza, the change that the person is sneezing is 24%,

Or,

We can also say, the degree of belief that the person is sneezing is 24%.

Note: [**Starts**]
Here we will discuss some very important concepts:

- Gaussian or Normal Distribution,
- Probability Density Function,
- Likelihood Functions,
- Maximum Likelihood Estimation (MLE).

Gaussian Distribution:
The Gaussian is the commonly and widely used distribution for continuous variables. It is also known as the ***normal*** distribution (symbol $\mathcal{N}$).

In the case of a single variable, $x \in (-\infty, \infty)$ and the distribution is controlled by two parameters:

the mean $\mu \in (-\infty, \infty)$ and
the variance $\sigma^2 > 0$.

Expressed as: $$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

The inverse of the variance is called the ***precision*** $\tau = \sigma^{-2}$.

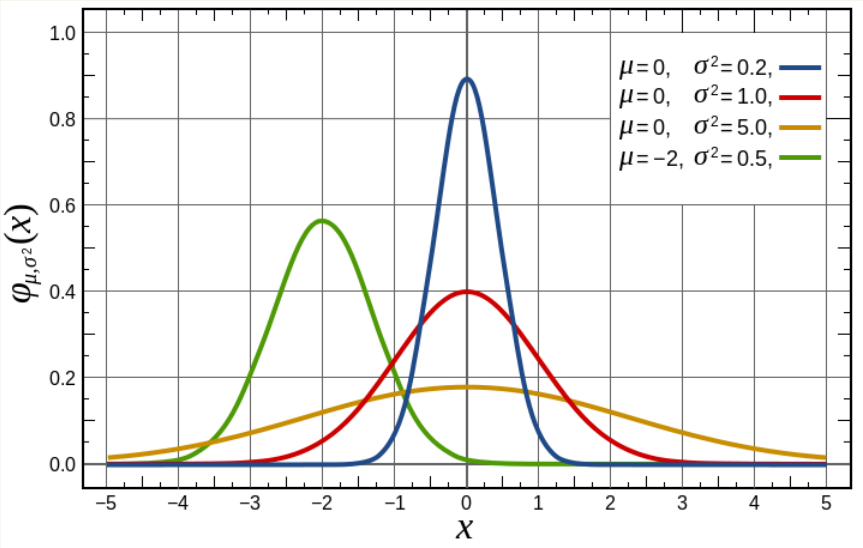


Figure: Different distribution for different values of μ and σ^2 (*ref: wiki*)

In the Gaussian distribution, *mean = mode = median*

Probability Density Function (PDF):

Pdf or ***density*** of a continuous random variable is a function, which helps compute the probability or the relative likelihood for this random variable to take on a given value. For example, the Gaussian distribution is a *pdf*. If a random variable x is following the distribution or density function f and for an interval $(a \le x \le b)$, the corresponding probability can be computed as an integral, i.e., $\mathrm{P}(a \le x \le b) = \int_a^b f(x)\,dx$.

Initiatively, the integral over the entire range of a *pdf* is 1, i.e., $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

***Likelihood Function*:**

A likelihood function $L(\theta)$ or $L(\theta; x)$ is the <u>probability</u> or <u>probability density</u> for the occurrence of sample outcomes $(x_1, x_2, ...x_n)$ given that the probability density $f(x;\theta)$ with parameter θ. We write as:

$$L(\theta) \equiv L(\theta; x) = f(x;\theta) = p(x;\theta)$$

The likelihood is usually a synonym for probability.

If the data are *independent and identically distributed*, then the likelihood can be computed as:

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

Maximum Likelihood Estimation **(MLE):**

Now, assume we have a given data D: $\{x_1, x_2, ...x_n\}$, and we are trying to find the best (probability) distribution out of a set of them that can represent the data. We are after best P_θ, and $\theta \in \Theta$

The best can be defined as to maximize P_θ for a particular choice of θ (denoted as θ_{MLE}) out of available options from Θ. Formally,

$$\theta_{MLE} = \underset{\theta \in \Theta}{\arg\max}\, P(x;\theta)$$

Usage of *Maximum Likelihood Estimation* (Important)

In this course, more or less, we are after *parameter estimation* to find the goodness of the fit of the observed data to our models. The estimation can be done by a least-square estimation approach, as we have seen before. However, the estimation can also be done using maximum likelihood estimation.

MLE has many optimal properties[1];

(a) Sufficiency: complete information about the parameter of interest contained in its MLE estimator.
(b) Consistency: true parameter value that generated the data
(c) Efficiency: Lowest possible variance of parameter estimates achieved asymptotically.

We like to apply MLE for the following case:

For a given data D{ $x_1, x_2, ...x_N$ }, we want to ***estimate*** the mean.

<u>For a given data 'D', we want to *estimate* the mean</u>

Let us assume, the given data D = { $x_1, x_2, ...x_N$ }, iid is followed, pdf is Gaussian. We want to estimate the best mean (μ) using MLE.

Since pdf is Gaussian, we write, $P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

Since, iid, $L(\mu) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$ … (A)

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2}\right\}\right) \times \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_2-\mu)^2}{2\sigma^2}\right\}\right) \times \ldots \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_N-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times \exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(x_2-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_N-\mu)^2}{2\sigma^2}\right\}$$

[1] I. J. Myung, Journal of Mathematical Psychology, 47 (2003) 90-100

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times e^{-\left\{\frac{(x_1-\mu)^2}{2\sigma^2}+\frac{(x_2-\mu)^2}{2\sigma^2}+\ldots+\frac{(x_N-\mu)^2}{2\sigma^2}\right\}}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \times e^{-\frac{1}{2\sigma^2}\left\{(x_1-\mu)^2+(x_2-\mu)+\ldots+(x_N-\mu)^2\right\}}$$

To maximize the Equation (A), we can only do that by minimizing $\{(x_1-\mu)^2+(x_2-\mu)+\ldots+(x_N-\mu)^2\}$, right?

Now, expanding the target, we can write our goal for MLE as,

$$f(\mu)_{\min} = N\mu^2 - 2\mu(x_1+x_2+\ldots+x_N)+(x_1^2+x_2^2+\ldots+x_N^2) \qquad \text{(B)}$$

Equation (B) is a quadratic function, and the coefficient of μ^2, i.e., N is positive. Thus the minimum exists.

Therefore to get minimum, we can differentiate $f(\mu)$ and then set 0, i.e., $f'(\mu)=0$

$$\Rightarrow 2.N.\mu - 2(x_1+x_2+\ldots+x_N)=0$$

$$\Rightarrow \mu = \frac{(x_1+x_2+\ldots+x_N)}{N}$$

Therefore, our estimated best mean using MLE is actually the arithematic average of the given data.

It is to be noted that the expression of (A) [i.e., the heavy expression] is often not convenient, thus we usually take ***log-likelihood***.

$$l(\theta) = Log\ L(\theta)$$

Since $l(\theta)$ is proportional to $L(\theta)$, we can reach the same goal if we instead use $l(\theta)$.

Thus for Equation (A), we could use:

$$
\begin{aligned}
l(\mu) &= \log\left(\prod_{i=1}^{N} P(x_i)\right) \\
&= \log\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right) \\
&= \log\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right) \\
&= -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \qquad \text{(C)}
\end{aligned}
$$

Now we can say to maximize $l(\mu)$, we need to minimize the term $\sum_{i=1}^{N}(x_i-\mu)^2$ because it has a negative sign in front of it in Equation (C). The other terms in Equation (C) are constants with respect to μ. Now, as we have previously done, here we can set our next goal to minimize the following,

$$f(\mu) = N\mu^2 - 2\mu(x_1 + x_2 + \ldots + x_N) + (x_1^2 + x_2^2 + \ldots + x_N^2).$$

Note [ENDs]

Generative Model **(definition)**: A generative model describes how a dataset is generated in terms of a probabilistic model. By sampling from this model, we are able to generate new data [1].

For example, suppose we have a dataset containing images of horses. We may wish to build a model that can generate a new image of a horse that has never existed but still looks real because the model has learned the general rules that govern the appearance of a horse. This is the kind of problem that can be solved using generative modeling.

A generative model must also be probabilistic rather than deterministic. If our model is merely a fixed calculation, such as taking the average value of each pixel in the dataset, it is not generative because the model produces the same output every time. The model must include a stochastic (random) element that influences the individual samples generated by the model.

There has been increased media attention on generative modeling projects such as StyleGAN[2] from NVIDIA, which is able to create hyper-realistic images of human faces, and the GPT-2 language model from OpenAI[3], which is able to complete a passage of text given a short introductory paragraph.

The Generate Model Framework ----------

We have a dataset of observations **X**.

- We assume that the observations have been generated according to some unknown distribution, p_{data}.
- A generative model p_{model} tries to mimic p_{data}. If we achieve this goal, we can sample from p_{model} to generate observations that appear to have been drawn from p_{data}.
- We are impressed by p_{model} if:
  - Rule 1: It can generate examples that appear to have been drawn from p_{data}.
  - Rule 2: It can generate examples that are suitably different from the observations in **X**. In other words, the model shouldn't simply reproduce things it has already seen.

Generative Model from Probability Domain: Let us start with our hypothetical cancer detection example. Assume we predict whether a tumor is benign or malignant based on two different features: *Size* and *Age*. Let us indicate Benign as class C_1 and Malignant as class C_2.

Also, assume we are a popular diagnostic center, and the cases we receive can be considered to be the representative sample of the large population. Thus, we can assume that the prior probabilities $P(C_1)$ and $P(C_2)$ for the classes are available. We

[2] GAN stands for Generative Adversarial Networks – see StyleGAN
[3] Language Models Are Unsupervised Multitask Learners, see paper.

also assume there are no other classes except C_1 and C_2. Thus, we can also write $P(C_1) + P(C_2) = 1$ in this case.

Assume, in our center, we have a total of 15 cases. Out of 15, 5 are in C1 (Benign), and the rest are in C_2 (Malignant) class. This information will help us have prior knowledge. We can say, the next new patient coming will fall under class C_1 with probability (5/15) and C_2 with (10/15) without having any knowledge about the properties of the tumor of that patient.

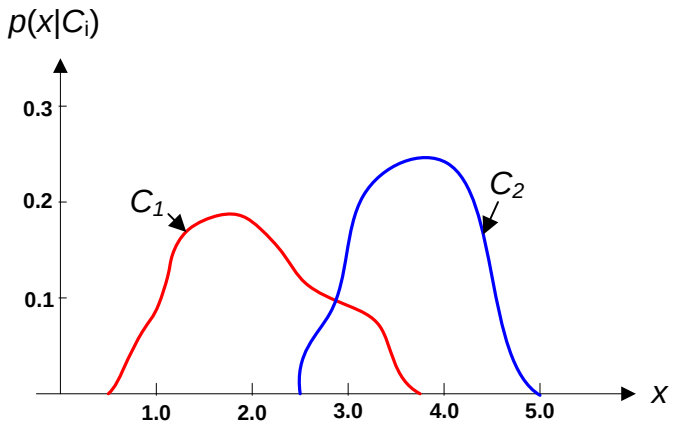


Figure 1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category C_i. If x represents the tumor-*size* (or the *age*), the difference in tumor size (or *age*) of populations of two types of tumors, density functions are normalized, and the area under each curve is 1.0.

In general, if $P(C_2) >> P(C_1)$, then without investigating further, we can often estimate the type of all new patients as type C_2 if the minor error rate is acceptable. However, these may not always be acceptable cases. We may not want someone to have unnecessary treatment when s/he does not have cancer (malignant type tumors). We need to consider other features to try our best for accurate prediction.

We consider feature x to be a continuous random variable. Its distribution can be presented as $p(x|C_i)$, namely the class-conditional probability density function. The difference between $p(x|C_1)$ and $p(x|C_2)$ can describe the tumor-*size* or the tumor-*age* between Benign versus Malignant types (see Figure 1).

Assume, we know both the prior probabilities $P(C_i)$ and the conditional dependencies $p(x|C_i)$. Now from Bayes formula [2] (see page 8), for each of classes we can write:

$$P(C_1 \mid x) = \frac{p(x \mid C_1)P(C_1)}{p(x)} \quad (1)$$

And,

$$P(C_2 \mid x) = \frac{p(x \mid C_2)P(C_2)}{p(x)} \quad (2)$$

By the names of the parts of the Equation, it can be viewed as:

$$posterior = \frac{likelihood \times prior}{evidence} \quad (3)$$

Bayes formula shows that by observing the value of x, we can convert the prior probability $P(C_i)$ to a posterior probability $P(C_i \mid x)$. We call $p(x \mid C_i)$ the likelihood of C_i with respect to x, a term chosen to indicate that, other things being equal, the category C_i for which $p(x \mid C_i)$ is large is more "likely" to be the true category. It is to be noted that the product of prior probability and likelihood is determining the posterior probability, and the evidence (the denominator) is playing as a scaling factor. Practically, from Equation (1) and (2) we can obtain the decision rule:

If $p(x \mid C_1)P(C_1) > p(x \mid C_2)P(C_2)$ THEN (4)
    we Decide C_1
ELSE
    we Decide C_2

Special Cases:

(a) For some x, if we have $p(x \mid C_1) = p(x \mid C_2)$, then this particular observation gives us no information about the state of nature. In such a case, prior probability helps us decide.

(b) For some x, if we have $p(C_1) = p(C_2)$, then the state of nature is equally probable. So, we rely on likelihoods in such cases.

In general, both of these factors are important in making a decision, and the Bayes decision rule combines them to achieve the minimum probability of error.

Naïve Bayes Classifier: The Naïve Bayes classifier is based on the Bayes rule. Naïve Bayes is popular being simple, intuitive and works well in the higher dimensional space. Let us redefine the rule (Equation (4)) more generally. Assume, given a set of variables, $X = \{x_1, x_2, ..., x_d\}$, and class sets $C_i = \{C_1, C_2, ..., C_k\}$ following Bayes rules we can write:

$$P(C_i \mid x_1, x_2, ..., x_d) \equiv p(x_1, x_2, ..., x_d \mid C_i) P(C_i) \tag{5}$$

Further, naïve Bayes assumes that the conditional probabilities of the variables are statistically independent (this is a "naïve" assumption, thus known as "Idiot's Bayes, however, practically naïve Bayes classifiers often outperform far more sophisticated alternatives). This likelihood is computed as products:

$$P(X \mid C_i) \equiv \prod_{j=1}^{d} p(x_j \mid C_i) \tag{6}$$

Note: Without the naïve assumption or, in other words, without the resulting Equation (6), the computation would have been expensive. This could have been expanded using the chain rule as:

$$\begin{aligned} P(C_i \mid x_1, x_2, ..., x_d) &\equiv p(x_1, x_2, ..., x_d \mid C_i) P(C_i) \\ &= p(x_2, ..., x_d \mid C_i, x_1) P(x_1 \mid C_i) P(C_i) \\ &= p(x_3, ..., x_d \mid C_i, x_1, x_2) P(x_2 \mid C_i, x_1) P(x_1 \mid C_i) P(C_i) \\ &= \ldots \end{aligned}$$

Finally, we can summarize from Equation (5) and (6):

$$P(C_i \mid X) \equiv P(C_i) \prod_{j=1}^{d} p(x_j \mid C_i) \tag{7}$$

We can use Equation (7) for classification problem and can pick the class as:

$$C_i = \arg\max_{c_i} P(c_i) \prod_{j=1}^{d} p(x_j \mid c_i) \tag{8}$$

Basically, we pick the class for which the value is found maximum.

Example: Let us discuss a popular example to demonstrate the naïve Bayes classification algorithm. The example file, "weather.nominal.arff" is available under '../weka/data/', and the data is given here in Table 1.

Table 1: The weather data (in the nominal form) is given to decide whether to play outside or not.

outlook	temperature	humidity	Windy	play
sunny	**hot**	**high**	FALSE	no
sunny	**hot**	**high**	**TRUE**	no
overcast	**hot**	**high**	FALSE	**yes**
rainy	mild	**high**	FALSE	**yes**
rainy	**cool**	normal	FALSE	**yes**
rainy	**cool**	normal	**TRUE**	no
overcast	**cool**	normal	**TRUE**	**yes**
sunny	mild	**high**	FALSE	no
sunny	**cool**	normal	FALSE	**yes**
rainy	mild	normal	FALSE	**yes**
sunny	mild	normal	**TRUE**	**yes**
overcast	mild	**high**	**TRUE**	**yes**
overcast	**hot**	normal	FALSE	**yes**
rainy	mild	**high**	**TRUE**	no

Data Set: There are five nominal or categorical data. From the dataset, we want to decide whether we should play outside (i.e., play= 'yes') or we should not play outside (i.e., play= 'no') based on the values of the other 4 input columns {outlook, temperature, humidity, windy}.

There are 14 rows given in Table (#1), N = 14. The dimension of the input features d = 4, since X={outlook, temperature, humidity, windy}. There are two possible values of 'play'; therefore, the classification is a binary classification problem. Let us label value 'yes' as class C_1 and 'no' as class C_2.

We can compute the prior probabilities as:

$$P(C_1) = \frac{9}{14} \text{ [There are 9 'yes' out of 14 rows]}$$

$$P(C_2) = \frac{5}{14}$$

Now let us compute the likelihood by computing the frequencies from Table 1. The results are placed in Table 2.

Table 2: Likelihood computation

Outlook			temperature			humidity			Windy		
play →	yes	no	play →	yes	no	play →	yes	no	play →	yes	no
sunny	$\frac{2}{9}$	$\frac{3}{5}$	hot	$\frac{2}{9}$	$\frac{2}{5}$	high	$\frac{3}{9}$	$\frac{4}{5}$	TRUE	$\frac{3}{9}$	$\frac{3}{5}$
overcast	$\frac{4}{9}$	$\frac{0}{5}$	mild	$\frac{4}{9}$	$\frac{2}{5}$	normal	$\frac{6}{9}$	$\frac{1}{5}$	FALSE	$\frac{6}{9}$	$\frac{2}{5}$
Rainy	$\frac{3}{9}$	$\frac{2}{5}$	cool	$\frac{3}{9}$	$\frac{1}{5}$						

Let us assume a day (rainy, hot, high, TRUE, ?), which is not an existing combination in Table #1. Thus we need to predict the answer, whether we should play or not. Following Equation (7), to compute the posterior values for the classes, we can write:

For *yes* class, $P(C_1 \mid x_j^T) \equiv P(C_1) \prod_{j=1}^{d} p(x_j \mid C_1)$

$$= P(C_1) \times p(rainy \mid C_1) p(hot \mid C_1) p(high \mid C_1) p(TRUE \mid C_1)$$

$$= \frac{9}{14} \times \left(\frac{3}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \right)$$

$$= 0.005291.$$

For *no* class, $P(C_2 \mid x_j^T) \equiv P(C_2) \prod_{j=1}^{d} p(x_j \mid C_2)$

$$= P(C_2) \times p(rainy \mid C_2) p(hot \mid C_2) p(high \mid C_2) p(TRUE \mid C_2)$$

$$= \frac{5}{14} \times \left(\frac{2}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} \right)$$

$$= 0.027428.$$

Since $P(C_2 \mid x_j^T) > P(C_1 \mid x_j^T)$, the prediction suggests that we should not play outside when the day is *rainy*, the temperature is *hot*, the humidity is *high*, and it is windy.

The outcomes are relative values proportional to the actual probabilities since we omitted the 'evidence' or the original Equation's denominator. However, if it is needed, we can compute the class probabilities as:

$$P(C_1 \mid x_j^T) = \frac{0.005291}{0.005291 + 0.027428} = 0.1617,$$

And, $$P(C_2 \mid x_j^T) = \frac{0.027428}{0.005291 + 0.027428} = 0.8382.$$

The example demonstrates how simple it is to apply the naïve Bayes classifier. The training phase is very fast too. In practice, it is reported to outperform many sophisticated state-of-art classifiers, and it is found useful when the feature dimension increases and other approaches start to perform poorly. Naïve Bayes classifier is well known for its effective applications such as spam-filtering, document-classification, etc.

Laplace Smoothing / Additive Smoothing:

From Table #2, we see an entry, P(outlook=*overcast* | play=*no*) $= \frac{0}{5}$. This is practically a problem because:

(a) for involved computation, it can turn the results into 0, which is not desirable. Say, for a day (overcast, hot, high, TRUE, ?), and we need to compute:

$$P(C_2 \mid x_j^T) \equiv P(C_2) \times p(overcast \mid C_2)\, p(hot \mid C_2)\, p(high \mid C_2)\, p(TRUE \mid C_2)$$

$$= \frac{5}{14} \times \left(\frac{0}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} \right)$$

= 0, and this is just because of P(*overcast* | *no*) $= \frac{0}{5}$.

(b) Since we have no occurrence for an event in the sample, it does not necessarily mean that the chance of that event to occur is absolutely zero.

We can modify such problematic occurrence(s) using *Laplace Smoothing or Additive Smoothing*. The idea is to add '1' to the numerator and add *n* to the

denominator, where *n* = number of possible values for that involved features in that classification problem. Therefore, we modify the following entries using Laplace smoothing:

$$P(\text{outlook}=overcast \mid \text{play}=no) = \frac{0+1}{5+(n=3)} = \frac{1}{8}, \text{ (previously was } \frac{0}{5}).$$

$$P(\text{outlook}=sunny \mid \text{play}=no) = \frac{3+1}{5+3} = \frac{4}{8}, \text{ (previously was } \frac{3}{5}).$$

$$P(\text{outlook}=rainy \mid \text{play}=no) = \frac{2+1}{5+3} = \frac{3}{8}, \text{ (previously was } \frac{2}{5}).$$

Further Preliminaries on Normal/Gaussian Distribution:

For further extension of the theme (Bayes theorem), we will discuss some preliminaries first. We have discussed Normal/Gaussian distribution so far. We had just talked about the univariate case, and here we will discuss bivariate and multivariate Normal/Gaussian distributions.

First of all, why are we concerned about the distribution in particular normal/Gaussian distribution? The ***Central Limit Theorem*** (CLT), which is one of the most important outcomes of the probability theory, tells us that, under various conditions, the distribution for the sum (or mean) of *n* independent random variables approaches a particular limiting form known as the normal distribution. Therefore, the normal or Gaussian probability density can help us build the model and approximate our data distribution reasonably well.

In one dimension or for univariate normal distribution, we have the following Eq^{n}:

$$\mathcal{N}(x \mid \mu, \sigma^2) = p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad (9)$$

Further, one-dimensional Gaussian distribution with 0 mean and variance 1 (also called *standard normal distribution*), is denoted as $p(x) \sim \mathcal{N}(0, 1)$ is shown in Figure 2.

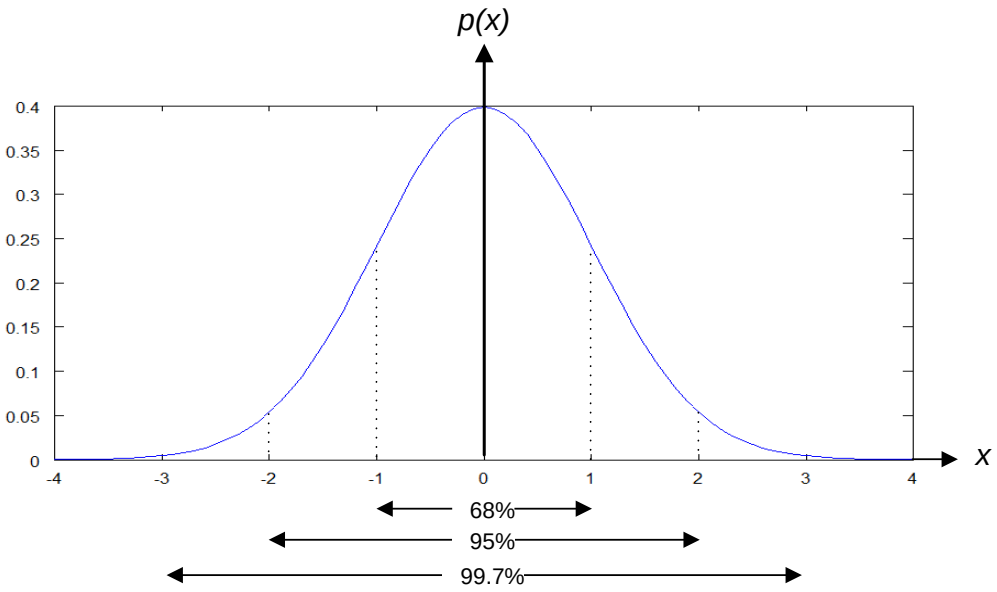


Figure 2: $p(x) \sim \mathcal{N}(0, 1)$ is shown. It has 68% of its probability mass in the range of $|x| \leq 1$ or, and we can write $\Pr[|x - \mu| \leq \sigma] \approx 0.68$, 95% in the range $|x| \leq 2$ or, $\Pr[|x - \mu| \leq 2\sigma] \approx 0.95$ and 99.7% in the range $|x| \leq 3$ or, $\Pr[|x - \mu| \leq 3\sigma] \approx 0.997$.

A natural measure of the distance from x to the mean μ is the distance $|x- \mu|$ measured in units of standard deviations:

$$r = \frac{|x - \mu|}{\sigma}$$

In one dimension, this is the z-score.

Multivariate Normal Densities

It is found that the sum of two independent normal variables is again normal. In fact, sums of dependent normal variables also have normal distributions.

Assume that each of the n random variables x_i is normally distributed, each with its own *mean* and *variance*, as:

$$p_{x_i}(x_i) \sim N(\mu_i, \sigma_i^2)$$

If these variables are independent, then their joint density can be computed as:

$$
\begin{aligned}
p(x) &= \prod_{i=1}^{n} p(x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\} \\
&= \frac{1}{(2\pi)^{n/2} \prod_{i=1}^{n} \sigma_i} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right\}
\end{aligned} \tag{10}
$$

We want to express Equation (10) in compact form using Matrix.

We can write for one of the terms in Equation (10) as:

$$
\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{11}
$$

Note [starts]:

Previously, we briefly mentioned:
Mean, $\boldsymbol{\mu} = E[\mathbf{x}] = \sum_{\mathbf{x} \in \{XY\}} \mathbf{x} P(\mathbf{x})$
Variance, $\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$

In detail, the *n*-dimensional mean vector $\boldsymbol{\mu}$ is defined by:

$$
\boldsymbol{\mu} = E[\mathbf{X}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}
$$

The covariance matrix $\boldsymbol{\Sigma}$ is defined as the square matrix whose $(i,j)^{\text{th}}$ element σ_{ij} (or, σ_{ji}) is the covariance of x_i and x_j, that is:

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

Therefore, we can write;

$$\boldsymbol{\Sigma} = \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & E[(x_1-\mu_1)(x_2-\mu_2)] & \cdots & E[(x_1-\mu_1)(x_n-\mu_n)] \\ E[(x_2-\mu_2)(x_1-\mu_1)] & E[(x_2-\mu_2)(x_2-\mu_2)] & \cdots & E[(x_2-\mu_2)(x_n-\mu_n)] \\ \cdots & \cdots & \ddots & \cdots \\ E[(x_n-\mu_n)(x_1-\mu_1)] & E[(x_n-\mu_n)(x_2-\mu_2)] & \cdots & E[(x_n-\mu_n)(x_n-\mu_n)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]$$

- $\boldsymbol{\Sigma}$ is symmetric (i.e., a square matrix that is equal to its transpose).
- diagonal elements are the variances of the individual elements of **x**, which is always positive.
- The off-diagonal elements are the covariances, can be +ve or –ve. Note: σ_{ij} and σ_{ji} will have the same sign.
- If the variables are statistically independent, the covariances are zero, and the matrix becomes a diagonal matrix.
- The $\boldsymbol{\Sigma}$ must be **positive semidefinite** (PSD). This is equivalent to the requirement that none of the eigenvalues of $\boldsymbol{\Sigma}$ can be negative. Alternatively, we can say $\boldsymbol{\Sigma}$ being a symmetric matrix, is said to be PSD (expressed as $\boldsymbol{\Sigma} \succeq 0$) iff the associated quadratic form $(x^T \boldsymbol{\Sigma}\, x) \geq 0$ for all x.
- Or, $\boldsymbol{\Sigma}$ is **positive semidefinite** (PSD) if and only if all its leading **principal minors** are nonnegative. For example, for the following matrix, M:

$$M = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

$$\Delta_1 = 3 \geq 0, \quad \Delta_2 = \det \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix} = 12 \geq 0 \quad \text{and} \quad \Delta_3 = \det \mathrm{M} = 20 \geq 0.$$

  - Therefore, M is a positive semidefinite matrix. Note that M is also symmetric, as $\mathrm{M} = \mathrm{M}^{\mathrm{T}}$.

How can we write the RHS of Equation (11), i.e., $(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})$?

Let us explain it by analogy and starting from an expression similar to the RHS.

Assume, $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, therefore, $X^T = [x_1 \quad x_2 \quad x_3]$

Also, assume $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$, therefore, $\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 \\ 0 & 0 & 1/\sigma_3^2 \end{bmatrix}$

……

We can easily verify the above inverse relationship by analogy again. We how to compute the inverse of a matrix A, i.e.,

$$A^{-1} = \frac{1}{\det(A)}(\text{Cofactor matrix of } A)^T$$

For example, if $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$, then det(A)=8,

and, **A_{11}= 8**, A_{12}=0, A_{13}=0, A_{21}=0, **A_{22}=4**, A_{23}=0, A_{31}=0, A_{32}=0, **A_{33}=2**.

Then, $A^{-1} = \frac{1}{8}\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^T = \frac{1}{8}\begin{bmatrix} 8 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$

So, when $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$, we get $A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \end{bmatrix}$

……

Now, let us expand the expression $X^T\Sigma^{-1}X$.

$$X^T\Sigma^{-1}X = [x_1 \quad x_2 \quad x_3]_{1\times 3}\begin{bmatrix} 1/\sigma_1^2 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 \\ 0 & 0 & 1/\sigma_3^2 \end{bmatrix}_{3\times 3}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{3\times 1}$$

$$= \begin{bmatrix} \frac{x_1}{\sigma_1^2} & \frac{x_2}{\sigma_2^2} & \frac{x_3}{\sigma_3^2} \end{bmatrix}_{1\times 3} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{3\times 1}$$

$$= \begin{bmatrix} \frac{x_1.x_1}{\sigma_1^2} + & \frac{x_2.x_2}{\sigma_2^2} + & \frac{x_3.x_3}{\sigma_3^2} \end{bmatrix}_{1\times 1}$$

$$= \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}$$

Thus we have:

$$\sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} = X^T \Sigma^{-1} X$$

We similarly wrote Equation (11), which was: $\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

Note [**Ends**]

Here we had the covariance matrix (covariances are all zero, is assumed):

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \qquad (12)$$

and therefore the inverse of the covariance matrix is:

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{bmatrix} \qquad (13)$$

We can also write, $\prod_{i=1}^{n} \sigma_i = |\Sigma|^{1/2}$ to make the component of Equation (10) compact further.

Note [**Starts**]

How can we write $\prod_{i=1}^{n}\sigma_i = |\Sigma|^{1/2}$?

Again by analogy, we can proceed. Say we have,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}, \text{ then}$$

Determinant => $|\Sigma| = \sigma_1^2(\sigma_2^2.\sigma_3^2 - 0)$

$$|\Sigma| = (\sigma_1\sigma_2\sigma_3)^2$$

$$|\Sigma|^{1/2} = (\sigma_1\sigma_2\sigma_3)$$

$$|\Sigma|^{1/2} = \prod_{i=1}^{3}\sigma_i$$

Note [**Ends**]

Finally, we can write Equation (10) in compact form:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \qquad (14)$$

The above Equation is the general form of a *multivariate density function*. **It is important to remember that it is not necessary for the (covariance matrix) $\mathbf{\Sigma}$ to be a diagonal matrix**.

<Check the Matlab/Octave code provided at the end of this lecture note to conceptualize the Gaussian distribution>

Bernoulli distribution:

Bernoulli distribution is a discrete distribution, which results in two possible outcomes ($y \in \{0,1\}$), and one (say, $y = 1$) of them occurs with probability p, whereas

the other (y = 0, for example) occurs with probability (1-p). Such a probability density function can be written as:

$$P(y) = \begin{cases} (1-p), & \text{when } y = 0 \\ p, & \text{when } y = 1 \end{cases} \quad (15)$$

In compact form, we can write it:

$$P(y) = p^{y}(1-p)^{(1-y)} \quad (16)$$

We see from Equation (16) that,

when y = 1, we get P(y) = p, and
when y = 0, we get $P(y)$ = (1-p)

For a binary classification problem $C_i \in \{0,1\}$, we can also express the distribution as:

$$C_i \sim Bernoulli(p)$$

Gaussian Discriminant Analysis based Model

Let us consider our cancer detection (Benign or Malignant) problem) (see the exercise):

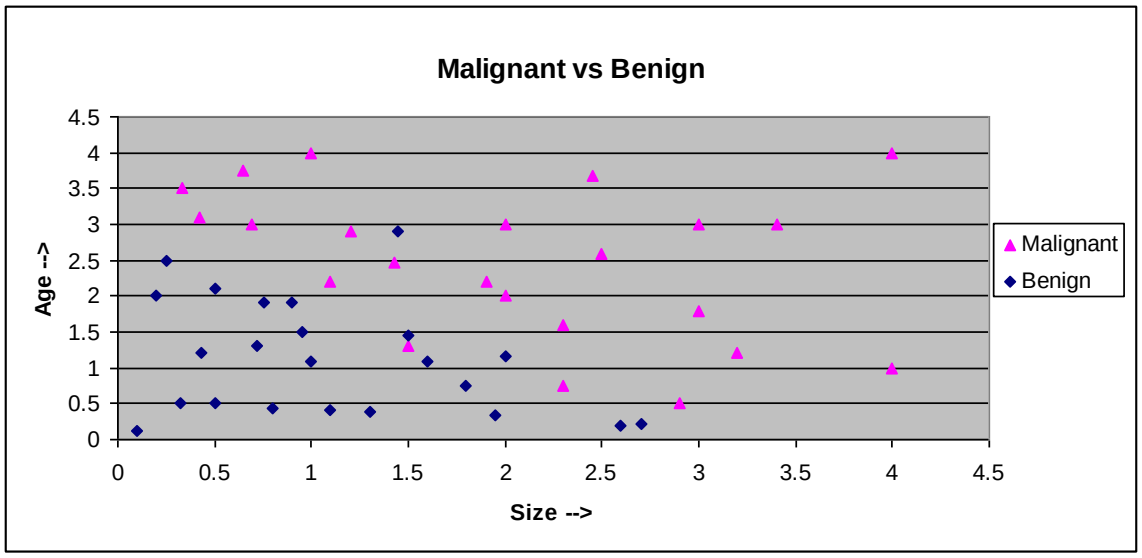


Figure 3: The (hypothetical) dataset is presenting the *Size* and *Age* of tumors, and the color label is indicating either it is 'Benign' or 'Malignant'.

To model this classification problem based on Gaussian Discriminant Analysis (GDA), we can write:

Assume, $C_i \in \{0,1\}$, where '0' = Benign, '1'=Malignant.

Since this is a binary classification, we can model the output as $C_i \sim Bernoulli(p)$. Here, we consider the input feature(s) (i.e., x_1=*Size* and x_2=*Age*) X is continuous random variable. We can thus use GDA to model likelihood for each of the individuals as $p(X|y = 0) \sim N(\mu_0, \Sigma)$ and $p(X|y = 1) \sim N(\mu_1, \Sigma)$. To model by GDA, we assume that the two classes have different means (i.e., μ_0, μ_1) and have the same covariance matrix Σ.

Following Equation(14) and Equation(16), we can write the complete equation for the 3 possible distributions as:

$$P(C_i) = p^{C_i}(1-p)^{(1-C_i)} \tag{17}$$

$$p(X \mid C_i = 0) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\} \tag{18}$$

$$p(X \mid C_i = 1) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} \qquad (19)$$

Let us use our data for cancer detection. We have a total N = 46 data points, and 23 of them are benign, and 23 of them are malignant. Therefore, the following Equation (17) we can write:

$$\mathrm{p}(\mathrm{C}_0) = \frac{23}{46} = 0.5 \quad [\text{prior probability of benign class}]$$

$$\mathrm{p}(\mathrm{C}_1) = \frac{23}{46} = 0.5 \quad [\text{prior probability of malignant class}]$$

Formally, we could also write the above computations as following:

$$p(C_0) = \frac{\sum_{j=1}^{N} I_j\{i = 0, in\ C_i\}}{N} \quad \text{and} \quad p(C_1) = \frac{\sum_{j=1}^{N} I_j\{i = 1, in\ C_i\}}{N}$$

Equation (18) (as well as (19)) would be the bivariate case for the given an example. For the benign class, we can compute

$$\boldsymbol{\mu}_0 = \begin{bmatrix} E[x_{1,C_0}] \\ E[x_{2,C_0}] \end{bmatrix} = \begin{bmatrix} \mu_{x_1,C_0} \\ \mu_{x_2,C_0} \end{bmatrix} = \begin{bmatrix} \frac{25.42}{23} \\ \frac{25.96}{23} \end{bmatrix} = \begin{bmatrix} 1.1052 \\ 1.1286 \end{bmatrix}$$

Similarly, $$\boldsymbol{\mu}_1 = \begin{bmatrix} E[x_{1,C_1}] \\ E[x_{2,C_1}] \end{bmatrix} = \begin{bmatrix} \mu_{x_1,C_1} \\ \mu_{x_2,C_2} \end{bmatrix} = \begin{bmatrix} 2.0552 \\ 2.4578 \end{bmatrix}$$

We can visualize the location of μ_0 and μ_1 in Figure 4 (indicated as Mean_B and Mean_M respectively – basically, these are the centroids for the two classes).

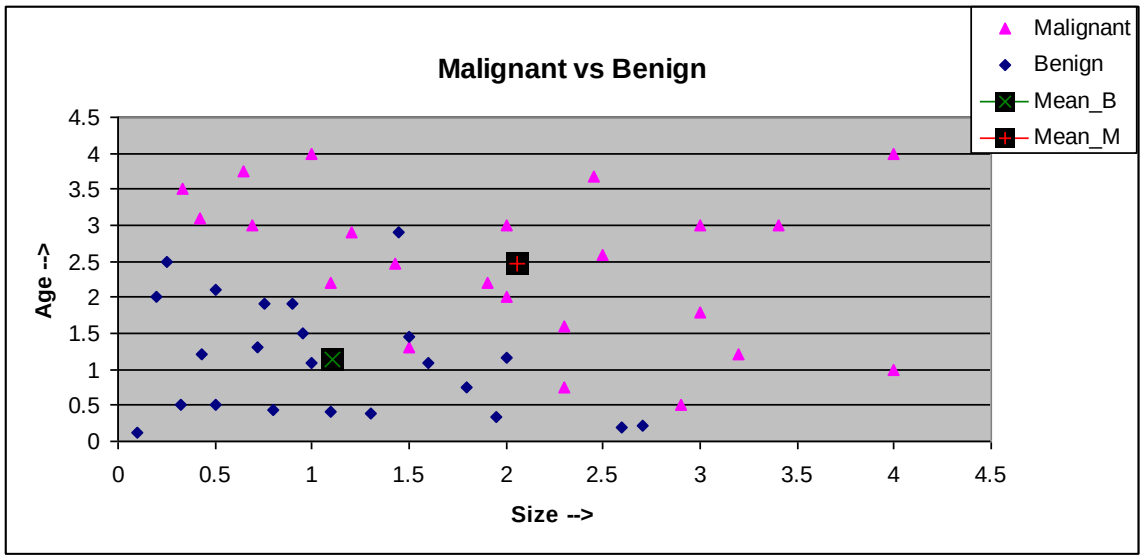


Figure 4: Mean value of Benign data (Mean_B) and mean value of Malignant data (Mean_M) have been indicated in the (hypothetical) dataset.

Now we need to compute the (common) covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1.09229995 & 0.03073048 \\ 0.03073048 & 1.29917801 \end{bmatrix}$$

Note:

It is handy to remember the following covariance σ_{xy} computation formula for the variable *x* and *y*:

$$\begin{aligned} \sigma_{xy} &= E[(x-\mu_x)\,(y-\mu_y)] \\ &= E[(xy - x\mu_y - y\mu_x + \mu_x\mu_y)] \\ &= E[xy] - E[x]\mu_y - E[y]\mu_x + \mu_x\mu_y \\ &= E[xy] - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y \\ &= E[xy] - 2\mu_x\mu_y + \mu_x\mu_y \\ &= E[xy] - \mu_x\mu_y \end{aligned}$$

Therefore, $\sigma_{xy} = E[xy] - \mu_x\mu_y$

Or, $\sigma_{xy} = E[xy] - E[x]E[y]$

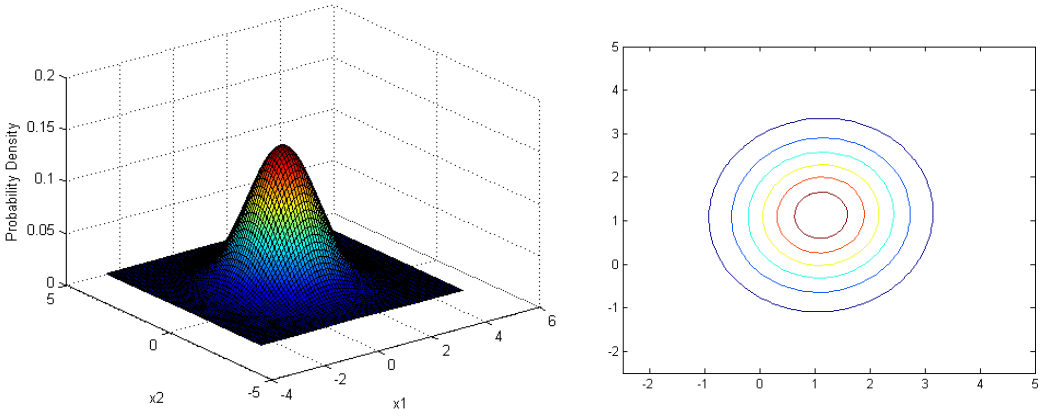
Thus, $\sigma_{12} = \sigma_{21} = E[x_1x_2] - \mu_1\mu_2$

For our example, $E[x_1x_2] = 2.864472$,

$$\mu_1 = 1.580217\ ,$$
$$\mu_2 = 1.793261\ .$$

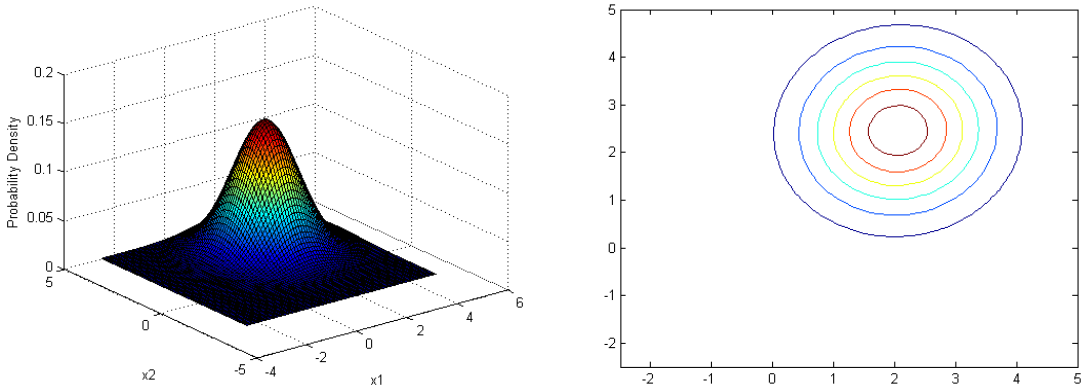
So, $\sigma_{12} = \sigma_{21} = E[x_1 x_2] - \mu_1\mu_2 = (2.864472) - (1.580217 \times 1.793261)$
$= 0.030730482363.$

With $\boldsymbol{\mu}_0 = \begin{bmatrix} 1.1052 \\ 1.1286 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1.09229995 & 0.03073048 \\ 0.03073048 & 1.29917801 \end{bmatrix}$, the bivariate normal distribution for the Benign class would look like the Figures of Figure 5.



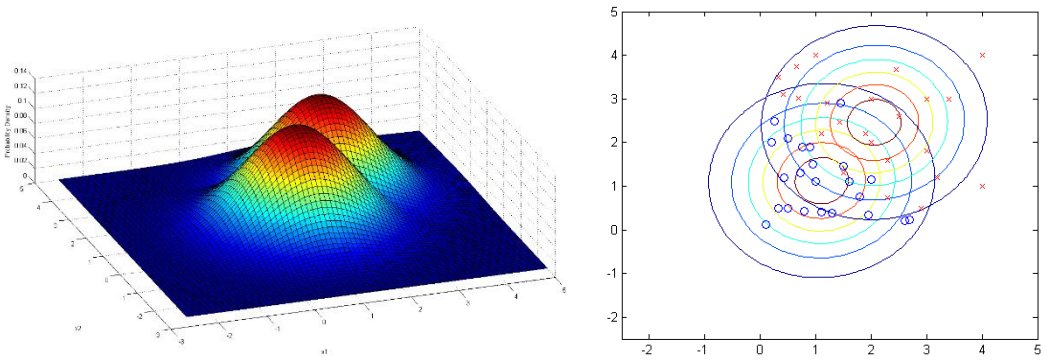
(a) (b)

Figure 5: For benign class (a) pdf and (b) contour are shown.



(a) (b)

Figure 6: For malignant class (a) pdf and (b) contour are shown.



(a) (b)

Figure 7: (a) Pdfs of both the classes: benign and malignant. (b) Given the datasets, the Contours of both classes are superimposed. '×' indicates data that belongs to the malignant class, and 'o' indicates data that belongs to the benign class.

With $\boldsymbol{\mu}_1 = \begin{bmatrix} 2.0552 \\ 2.4578 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1.09229995 & 0.03073048 \\ 0.03073048 & 1.29917801 \end{bmatrix}$, the bivariate normal distribution for the Malignant class would look like the Figures of Figure 6. Figure 7 shows the two contours of the two classes superimposed on the datasets.

Now for any given new point, we can check which class the new point belongs to based on the Bayes law that we have learned.
Let us find a new point X =[*Size Age*] = [0.5 0.5] using *MATLAB* code**:**

For ***benign*** class:

```
Mu = [1.1052  1.1286];
Sigma= [1.09229995    0.03073048;    0.03073048    1.29917801];
X = [0.5 0.5];
Px = mvnpdf(X,Mu,Sigma);
```

Px → 0.0979 (likelihood)

$P(C_0 \mid X) \equiv$ *likelihood* (see Eqn (18)) × *Prior-probability* ($p(C_0)$ =0.5)

= 0.0979 × 0.5

= 0.04895

For ***malignant*** class:

```
Mu = [2.0552  2.4578];
Sigma= [1.09229995    0.03073048;    0.03073048    1.29917801];
X = [0.5 0.5];
Px = mvnpdf(X,Mu,Sigma);
```

Px → 0.0108 (likelihood)

$P(C_1 \mid X) \equiv$ *likelihood* (see Eqn (19)) × *Prior-probability* ($p(C_1)$ =0.5)

$= 0.0108 \times 0.5$
$= 0.0054$

In summary, for a point X = [0.5 0.5], we computed,

$P(C_0 \mid X) \equiv 0.04895$ and

$P(C_1 \mid X) \equiv 0.0054$.

Therefore, the new point belongs to C_0 or *benign* class.

GDA approach makes stronger modeling assumptions and utilizing the data relatively more efficiently. And if the modeling assumption fits, the data sets well, and the prediction would be satisfactory. Remember that for our worked out example, the same covariance matrix Σ is used for both the distribution, but the means were different. To form the decision boundary, we can simply pick up the points where both the posterior probabilities for the two classes remain equal.

We have basically extended the Bayes classifier using Gaussian distribution for modeling the likelihood. Other potential models can be plugged in, such as *Poisson*, *gamma*, *lognormal* density function, and so on.

Appendix

Drawing Pdfs

Matlab: pdf :Y = pdf(NAME, X, A, B, …),

Or, ***Univariate Normal Distribution*** => **normpdf** (*x*), **normpdf** (*x, mu*), **normpdf** (*x, mu, sigma*)

Example:

```
X= -6:0.1:6;
Mu=0;        % means
Sigma=1;     % Standard deviation
Y=pdf ('normal',X, Mu, Sigma); %or, Y = normpdf (X, Mu, Sigma)
plot(X,Y);
```

Octave:

Univariate Normal Distribution => **normpdf** (*x*), **normpdf** (*x, mu, sigma*)

```
X= -6:0.1:6;
Mu=0;        % means
Sigma=1;     % Standard deviation
Y= normpdf (X, Mu, Sigma);
```

```
plot(X,Y);
```

Matlab:

mvnpdf %Multivariate normal probability density function

```
% Testing bivariate (such as x1, x2) case.
Mu = [0 0];
Sigma= [1 0; 0 1];
x1=-5:0.1:5;
x2=-5:0.1:5;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
Px = mvnpdf(X,Mu,Sigma);
Px = reshape(Px,length(x2),length(x1));
surf(x1,x2,Px);
xlabel('x1'); ylabel('x2'); zlabel('Probability Density');

hold on;                    % skip the hold on to have the contour separately
contour(x1,x2,Px);

%  Sigma must be square, symmetric, positive definite matrix
% pdf and contour for benign class
Mu = [1.1052  1.1286];
Sigma= [1.09229995 0.03073048; 0.03073048 1.29917801];
x1=-2.5:0.1:5;
x2=-2.5:0.1:5;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
Px = mvnpdf(X,Mu,Sigma);
Px = reshape(Px,length(x2),length(x1));
surf(x1,x2,Px);
xlabel('x1'); ylabel('x2'); zlabel('Probability Density');

hold on;                    % skip the hold on to have the contour separately
contour(x1,x2,Px);

% pdf and contour for Malignant class
Mu = [2.0552  2.4578];
Sigma= [1.09229995 0.03073048; 0.03073048 1.29917801];
```

```
x1=-2.5:0.1:5;
x2=-2.5:0.1:5;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
Px = mvnpdf(X,Mu,Sigma);
Px = reshape(Px,length(x2),length(x1));
surf(x1,x2,Px);
xlabel('x1'); ylabel('x2'); zlabel('Probability Density');

hold on;                    % skip the hold on to have the contour separately
contour(x1,x2,Px);
```

plot the data points (***Matlab code***):

```
load B.txt;
plot(B(:,1), B(:,2),  'o');
hold on;
load M.txt
plot(M(:,1), M(:,2),  'rx');
```

Reference:

[1] D. Foster, *Generative Deep Learning* O'Reilly 2019.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*: Wiley, 2000.

--- x ---