Generative Learning Algorithms

Teeradaj Racharak (เอ็กซ์)
r.teeradaj@gmail.com

Generative *vs.* Discriminative

So far, the methods we have tried attempt to learn $p(y|\boldsymbol{x})$ directly.
i.e. Given an input $\boldsymbol{x}$, we are trying to map directly to the output y.

Any algorithm that does this is called a discriminative learning algorithm.

Another class of algorithms instead tried to model from $p(\boldsymbol{x}|y)$ and $p(y)$.
Such methods are called generative learning algorithms.

Using Bayes' rule

If we can build a model of $p(\boldsymbol{x} \mid y)$ and $p(y)$, then Bayes' rule tells us that:

$$p(y \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid y) \cdot p(y)}{p(\boldsymbol{x})}$$

Why don't we care about $p(\boldsymbol{x})$? Let's see …

Using Bayes' rule

If we can build a model of $p(\boldsymbol{x} \mid y)$ and $p(y)$, then Bayes' rule tells us that:

$$p(y \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid y) \cdot p(y)}{p(\boldsymbol{x})}$$

Why don't we care about $p(\boldsymbol{x})$? Let's see …

First, generative models are most often used for classification, not regression !

In classification, y is discrete, so we have:

$$p(\boldsymbol{x}) = \Sigma_i p(\boldsymbol{x} \mid y = y_i) \cdot p(y = y_i)$$

If y is continuous, we just use an integral instead of sum.

This shows that if we can model $p(\boldsymbol{x} \mid y)$ and $p(y)$, we can obtain $p(y \mid \boldsymbol{x})$ without explicitly calculating $p(\boldsymbol{x})$.

Using Bayes' rule

We could calculate $p(\boldsymbol{x})$ directly.
But usually, we want to know which y maximizes $p(y \mid \boldsymbol{x})$.

ต้องการรู้ว่า y มีความน่าจะเป็นสูงสุดที่ class ไหน

In this case, all we need to do is find:

$$
\begin{aligned}
y^* &= \arg\max_y p(y \mid \boldsymbol{x}) \\
&= \arg\max_y \frac{p(\boldsymbol{x} \mid y) \cdot p(y)}{p(\boldsymbol{x})} \\
&= \arg\max_y p(\boldsymbol{x} \mid y) \cdot p(y)
\end{aligned}
$$

So, to perform classification in the generative approach,
all we need is models for $p(\boldsymbol{x} \mid y)$ and $p(y)$.

Gaussian Discriminant Analysis (GDA) Model

Gaussian Discriminant Analysis

Let's consider the case where $p(\boldsymbol{x} \mid y)$ is a multivariate Gaussian.

A possible example would be $\mathscr{Y} = \{\textbf{male}, \textbf{female}\}$ and $\mathscr{X} = \mathbb{R}^2$, where the features are a person's height and weight.

By the way, what is multivariate Gaussian ?

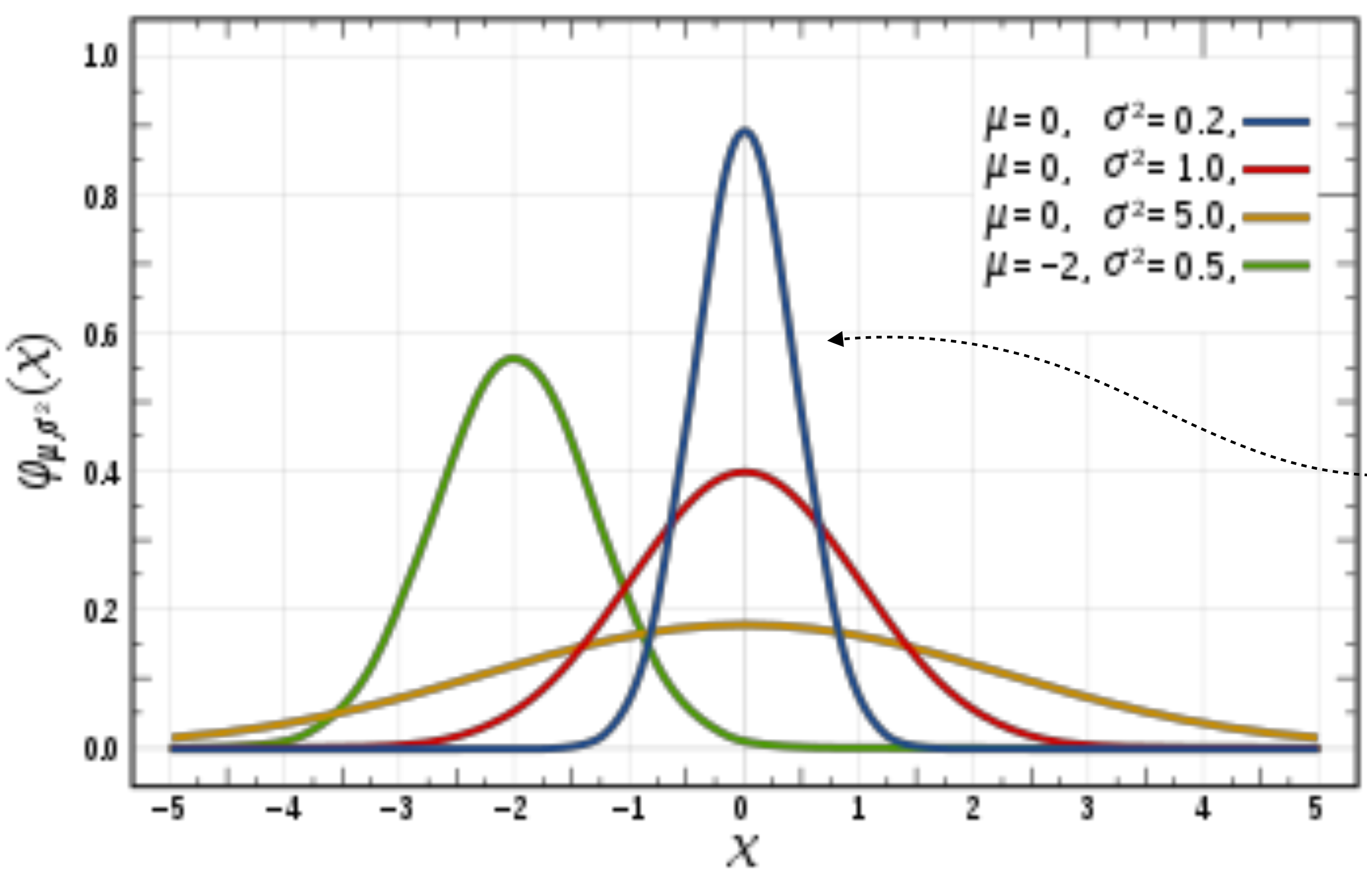
Let’s revisit Gaussian Distribution

Gaussian (Normal) Distribution

Suppose $x \in \mathbb{R}$. If x is a distributed Gaussian with mean μ and variance σ^2.

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

'distributed as'

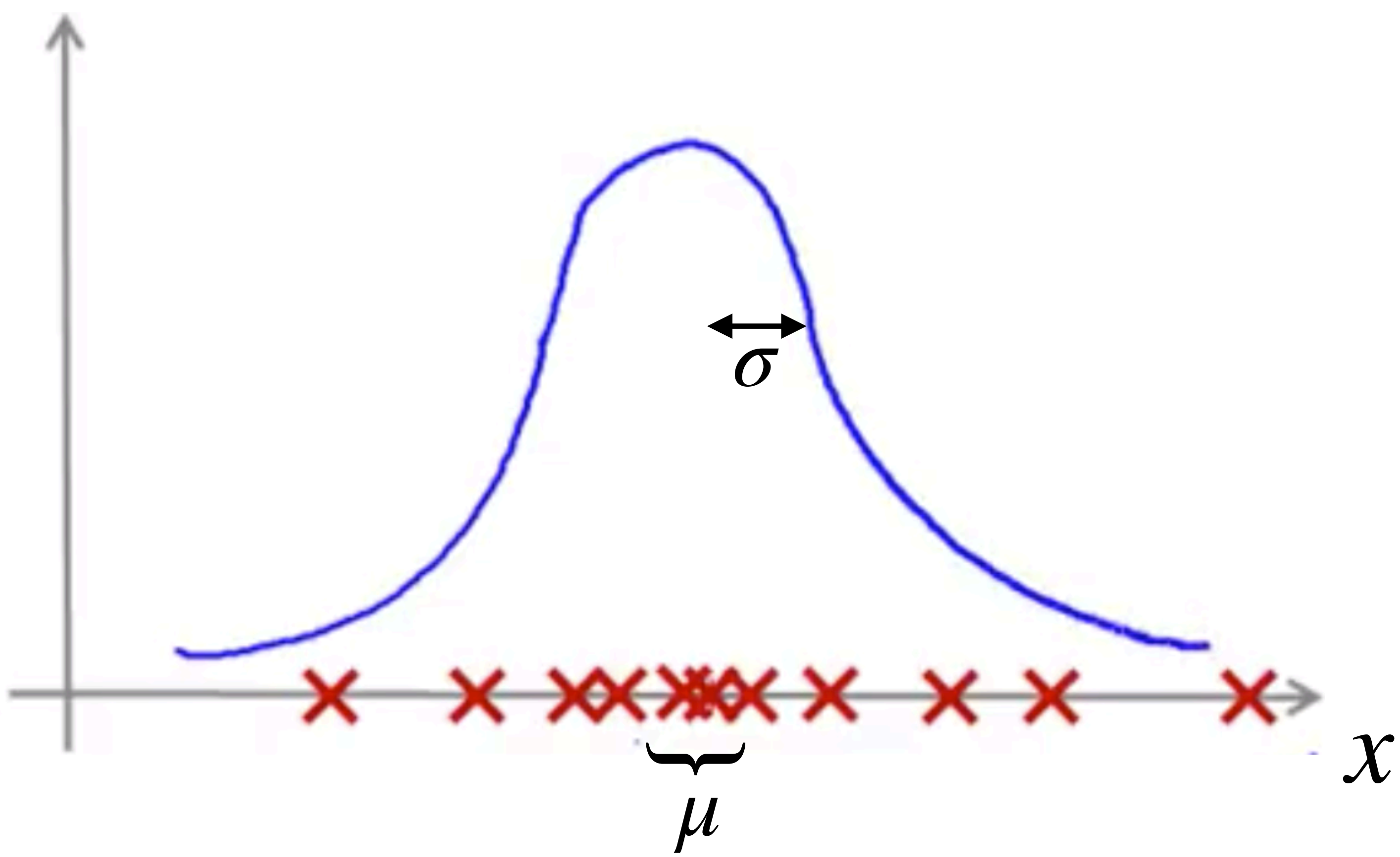


This specifies the probability of x taking on different values

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameter Estimation Problem

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ where $x^{(i)} \in \mathbb{R}$



Assume $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$

? ?

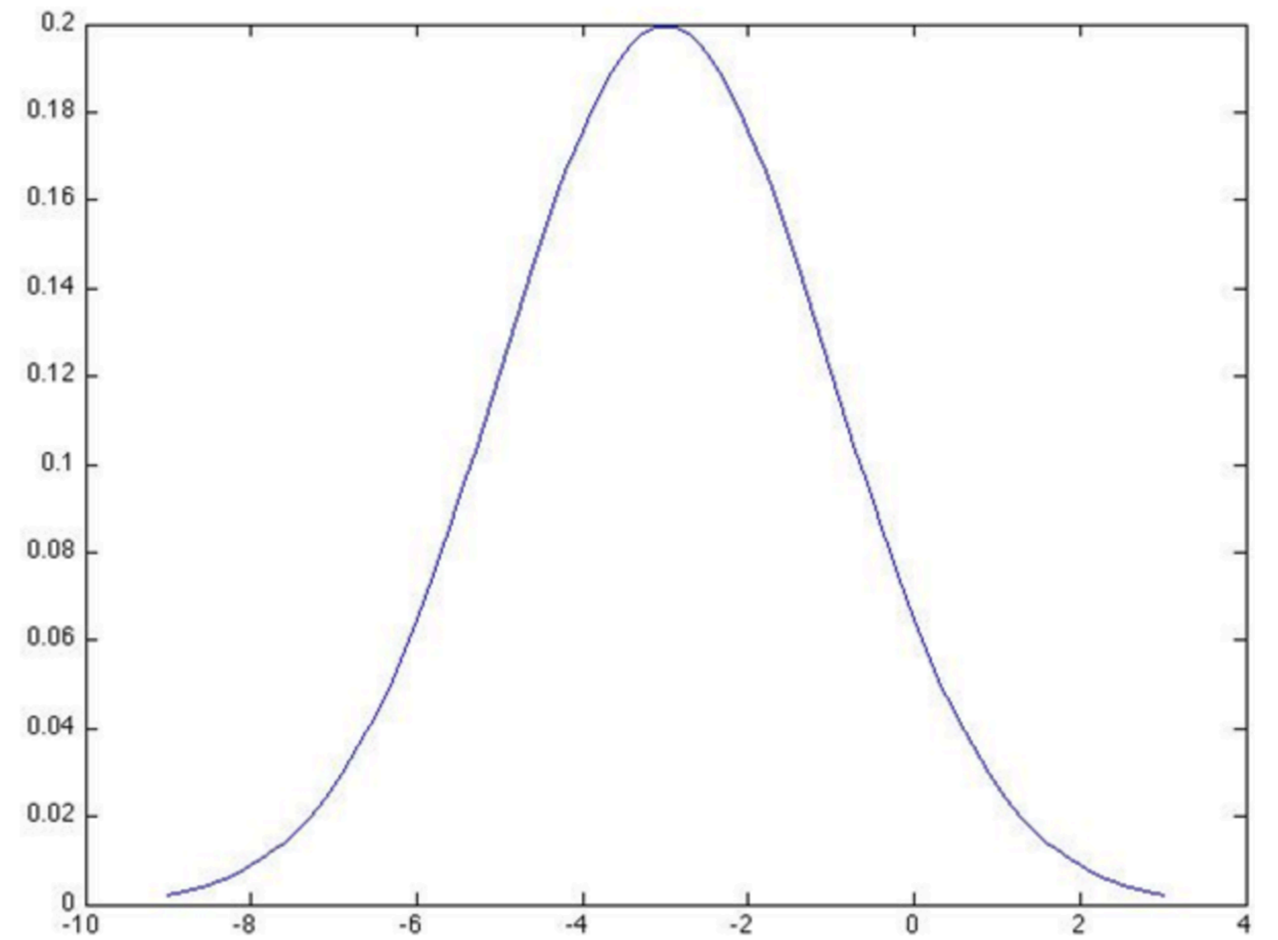
$$\mu = \frac{1}{m}\Sigma_{i=1}^{m} x^{(i)} \qquad \sigma^2 = \frac{1}{m}\Sigma_{i=1}^{m}(x^{(i)} - \mu)^2$$

Question

The formula for the Gaussian density is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Which of the following is the formula for the density to the right?



(i) $p(x) = \frac{1}{\sqrt{2\pi} \times 2} \exp\left(-\frac{(x-3)^2}{2 \times 4} \right)$

(ii) $p(x) = \frac{1}{\sqrt{2\pi} \times 4} \exp\left(-\frac{(x-3)^2}{2 \times 2} \right)$

(iii) $p(x) = \frac{1}{\sqrt{2\pi} \times 2} \exp\left(-\frac{(x+3)^2}{2 \times 4} \right)$

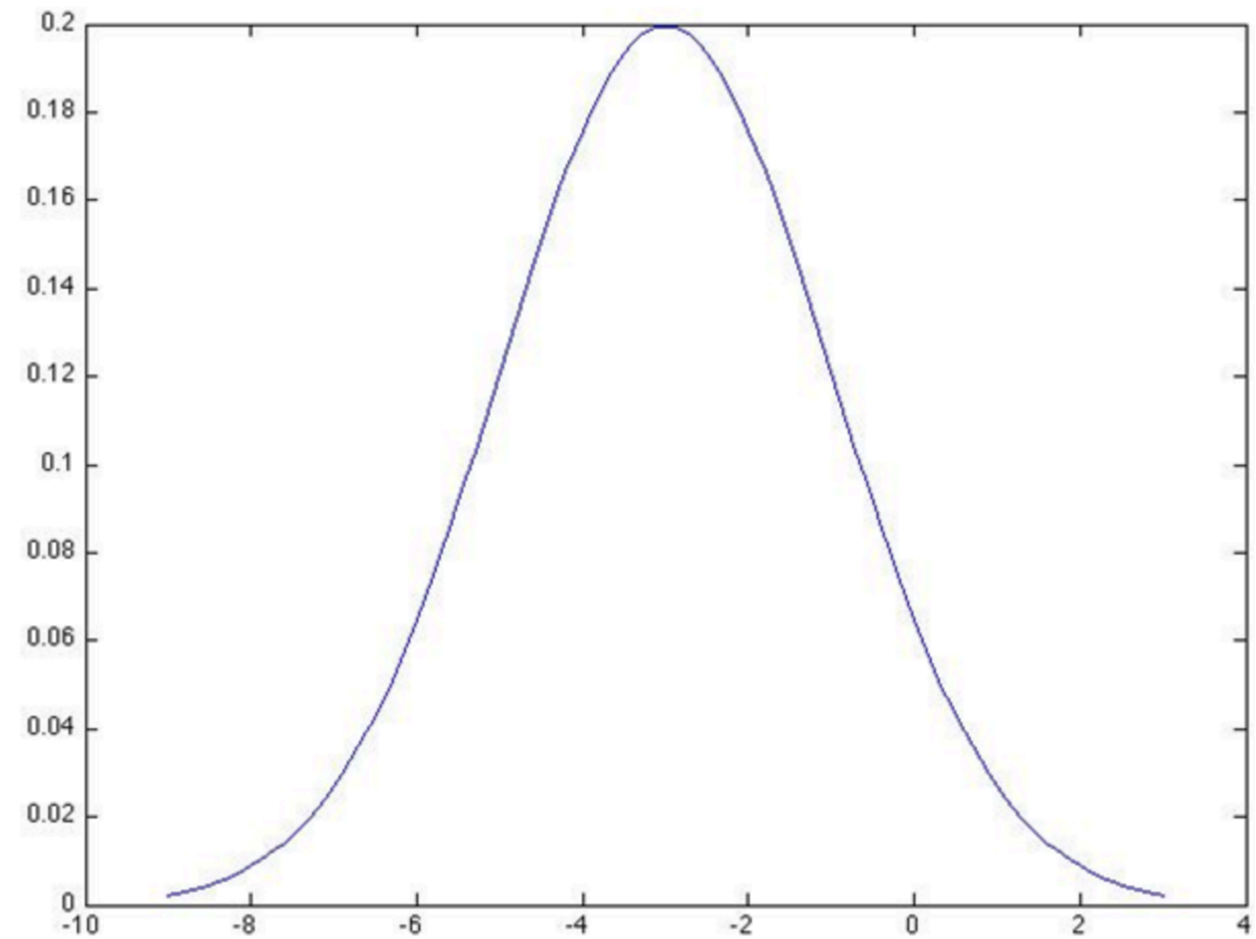
(iv) $p(x) = \frac{1}{\sqrt{2\pi} \times 4} \exp\left(-\frac{(x+3)^2}{2 \times 2} \right)$

Question

The formula for the Gaussian density is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Which of the following is the formula for the density to the right?



(i) $p(x) = \frac{1}{\sqrt{2\pi} \times 2} \exp\left(-\frac{(x-3)^2}{2 \times 4} \right)$

(ii) $p(x) = \frac{1}{\sqrt{2\pi} \times 4} \exp\left(-\frac{(x-3)^2}{2 \times 2} \right)$

(iii) $p(x) = \frac{1}{\sqrt{2\pi} \times 2} \exp\left(-\frac{(x+3)^2}{2 \times 4} \right)$

(iv) $p(x) = \frac{1}{\sqrt{2\pi} \times 4} \exp\left(-\frac{(x+3)^2}{2 \times 2} \right)$

Multivariate Gaussian Distribution

Multivariate Gaussian Distribution

The n-dimensional multivariate Gaussian distribution has:

- a mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$
- a covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$

We require that $\Sigma \geq 0$ is symmetric and positive semidefinite.

The distribution is written $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and the density is:

$$\boxed{p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}}$$

where $|\cdot|$ is the determinant.

Multivariate Gaussian Distribution

If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}; \Sigma)$, then we can write:

$$E[\boldsymbol{X}] = \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) d\boldsymbol{x} = \boldsymbol{\mu}$$

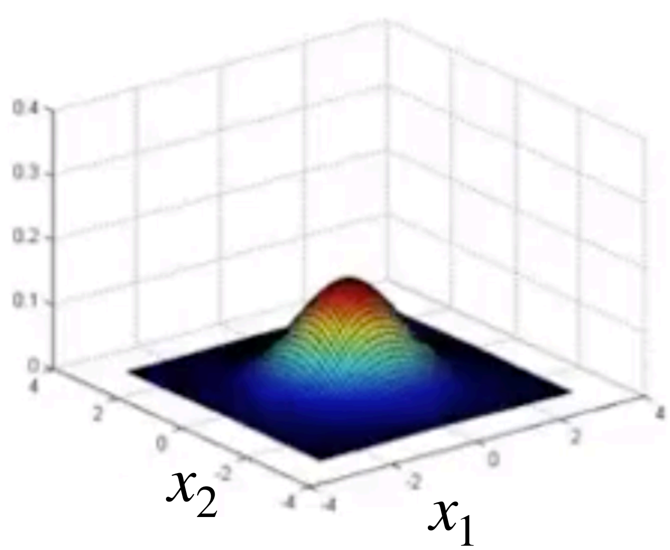
and

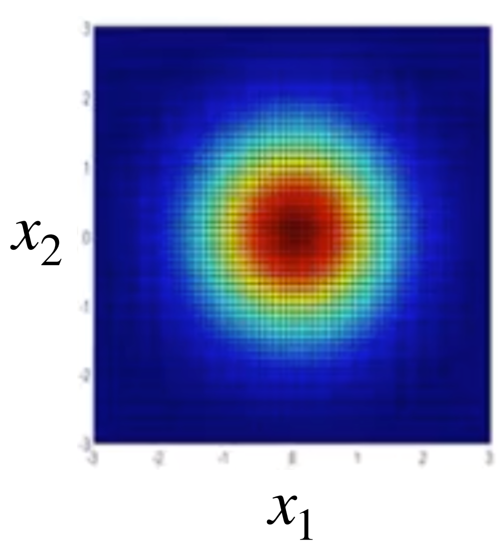
$$\textbf{Cov}(\boldsymbol{X}) = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T] = \Sigma$$

We will practice about the multivariate Gaussian in the lab to get an intuition about interpreting the covariance matrix.

Multivariate Gaussian (Normal) Distribution Example

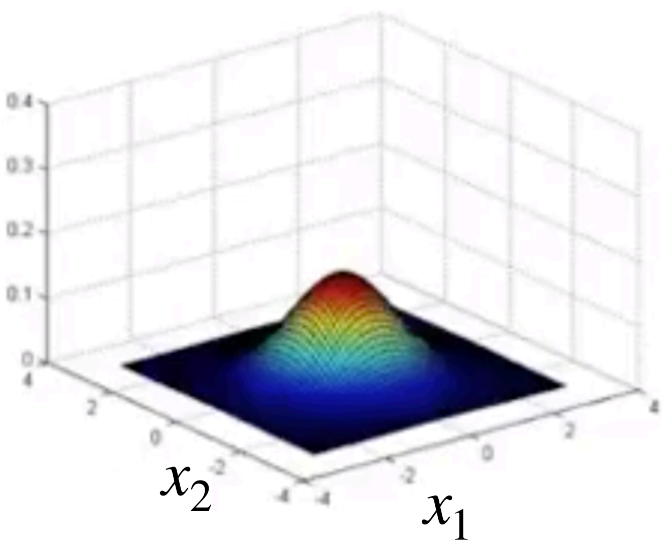
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{(Identity matrix)}$$

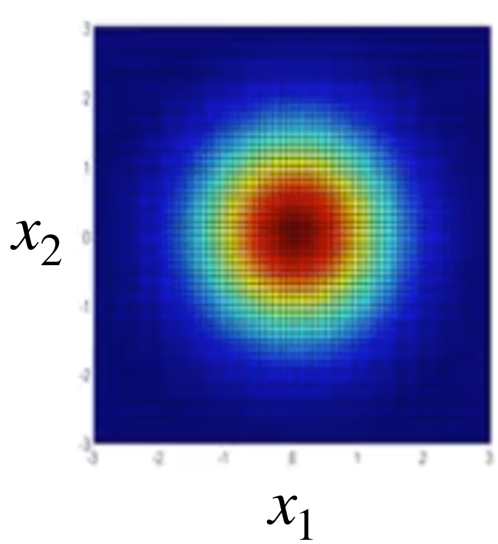




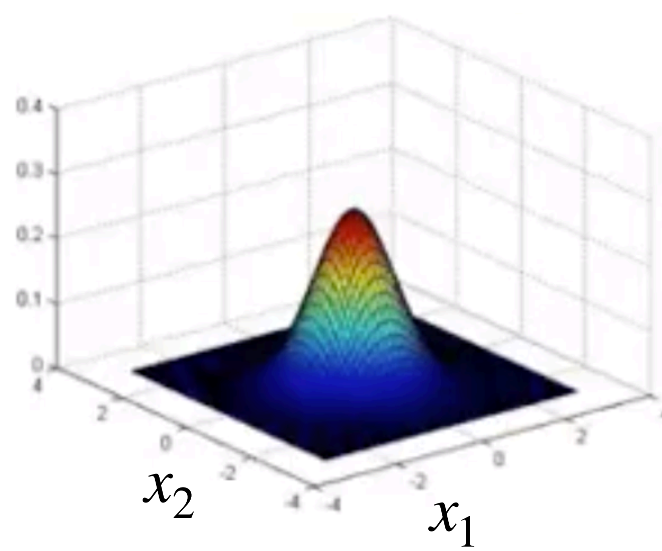
Multivariate Gaussian Distribution Example

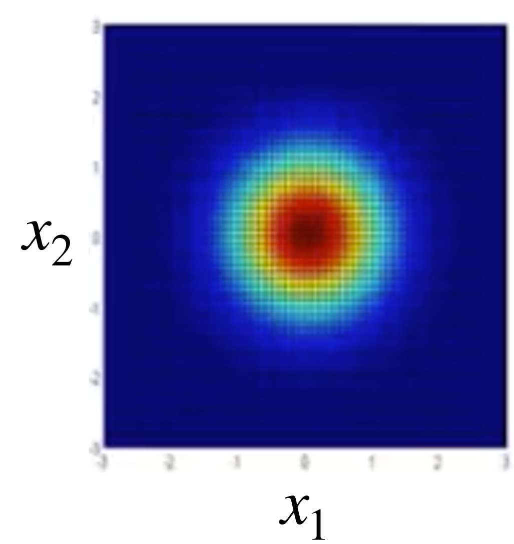
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



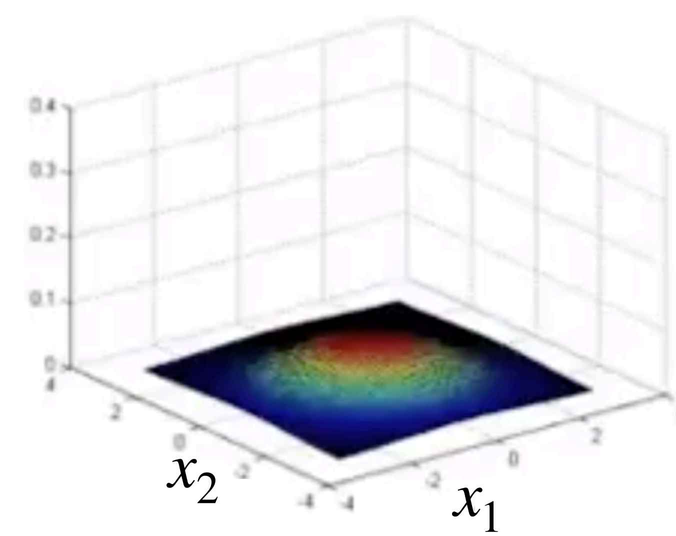


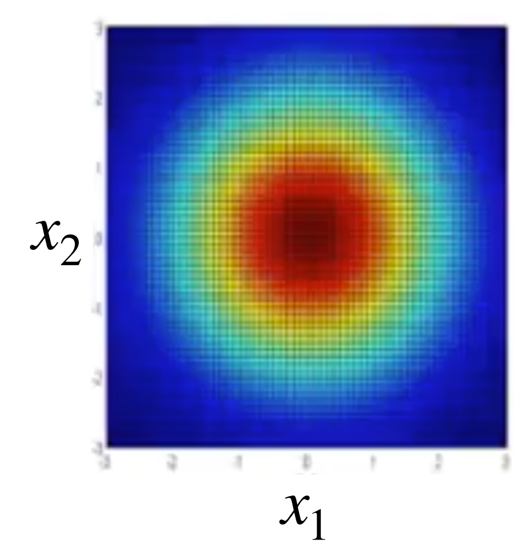
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$





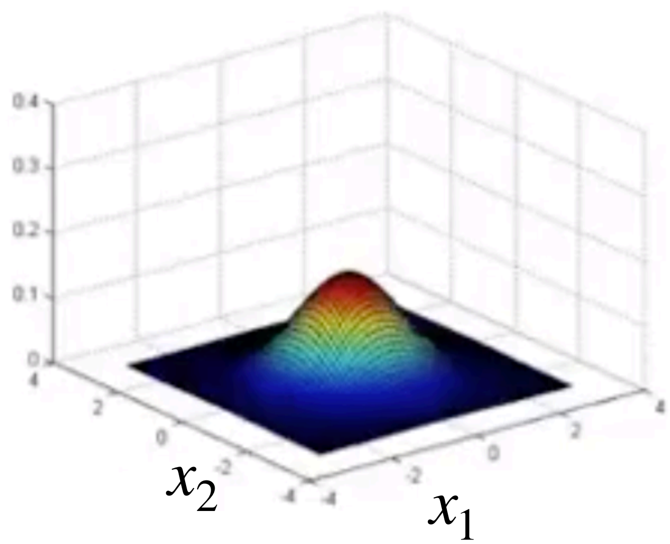
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

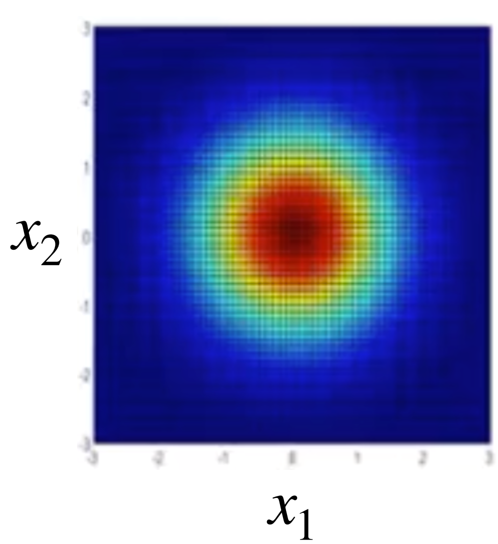




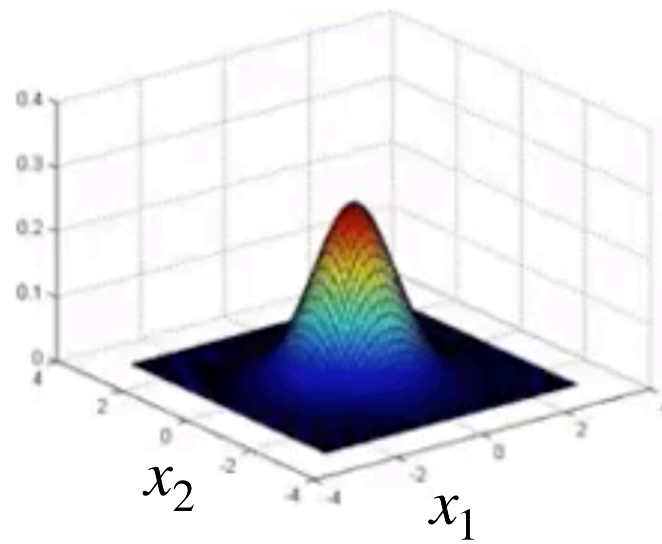
Multivariate Gaussian Distribution Example

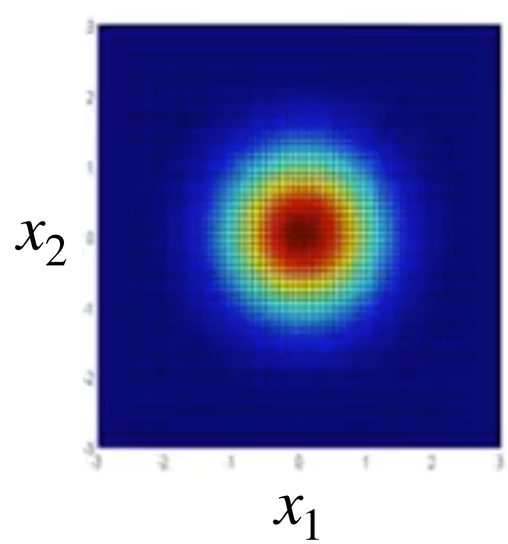
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



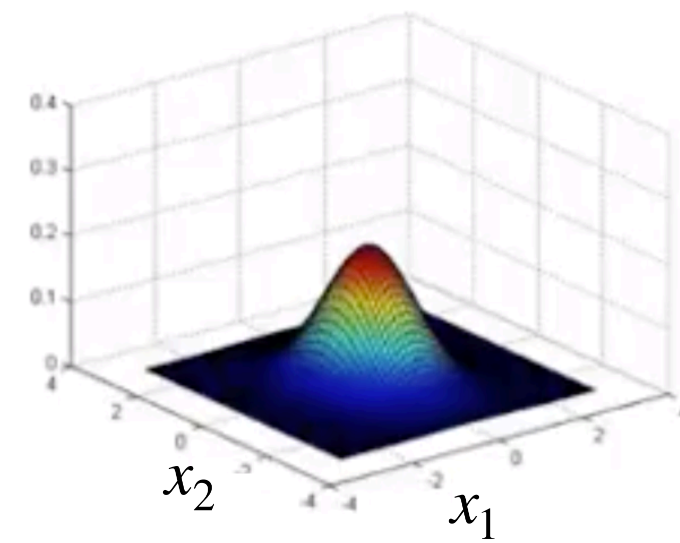


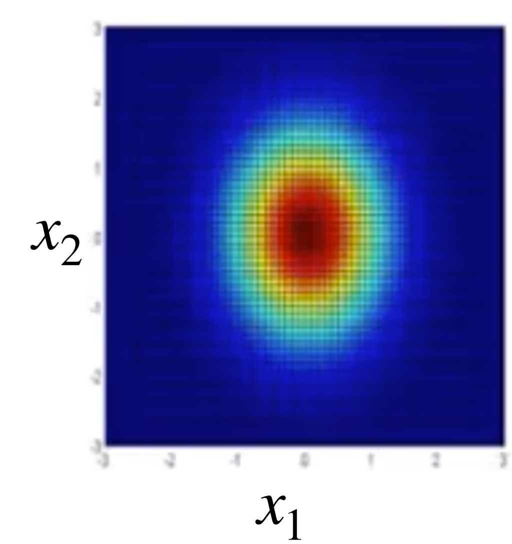
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$





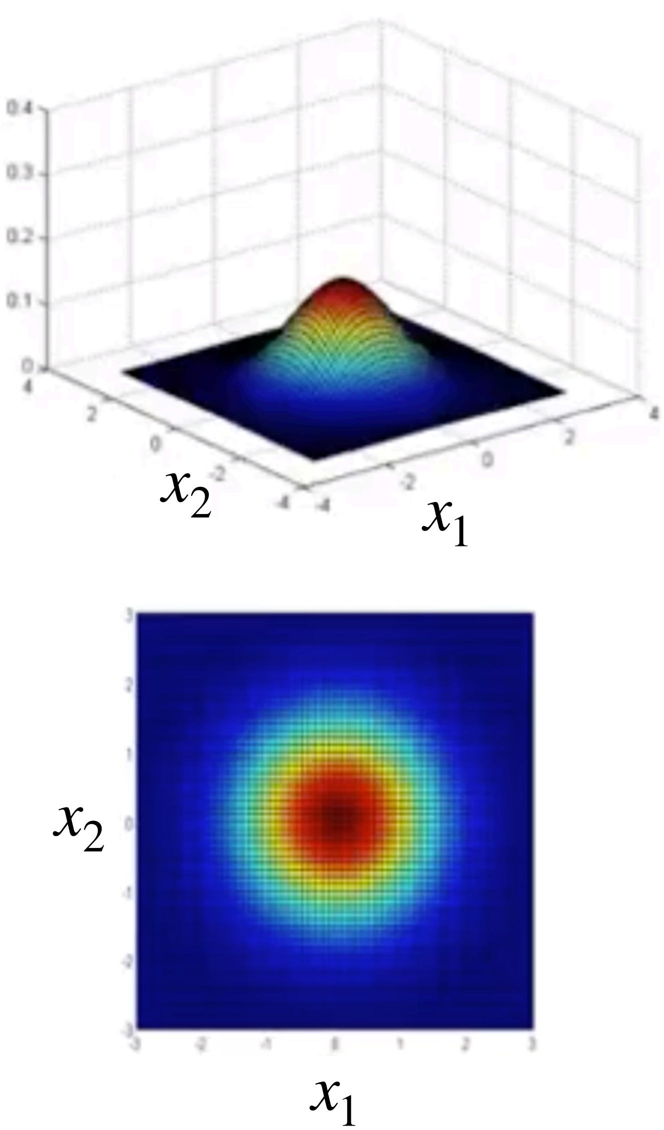
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{array}{cc} x_1 & \\ \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix} & x_2 \end{array}$$



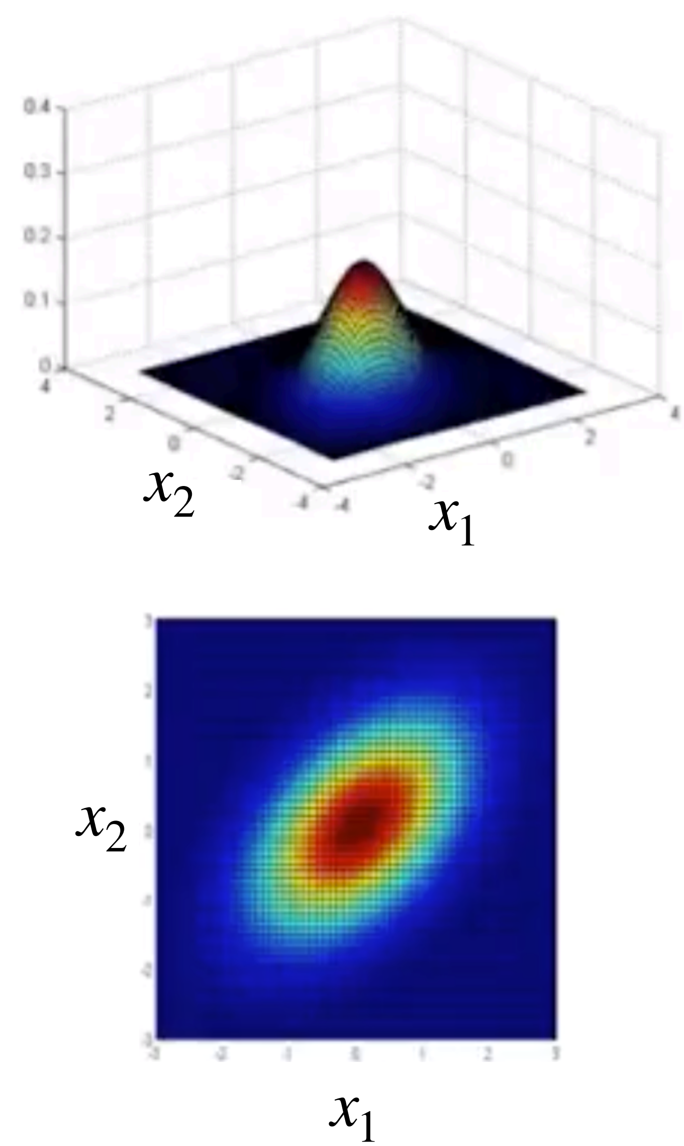


Multivariate Gaussian Distribution Example

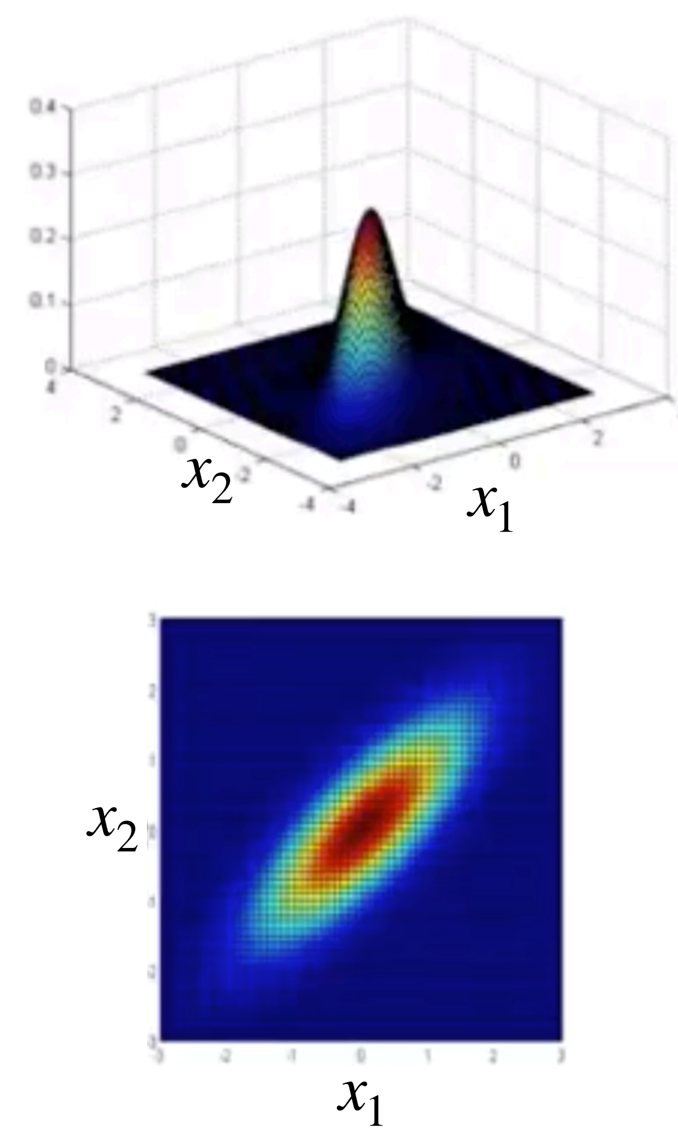
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

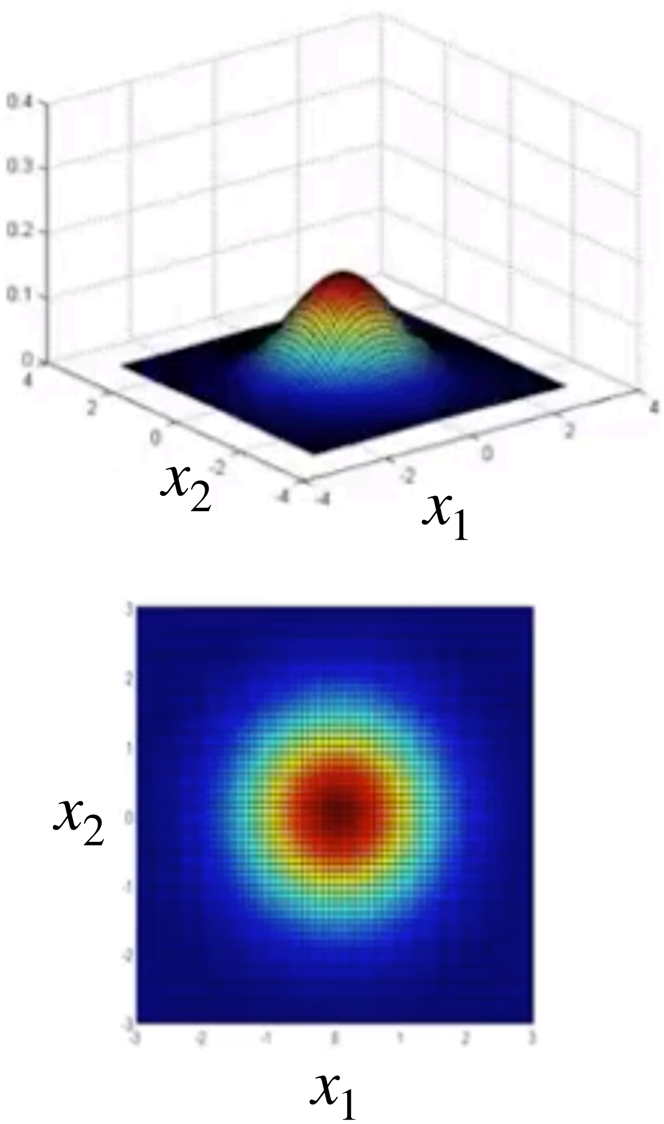


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

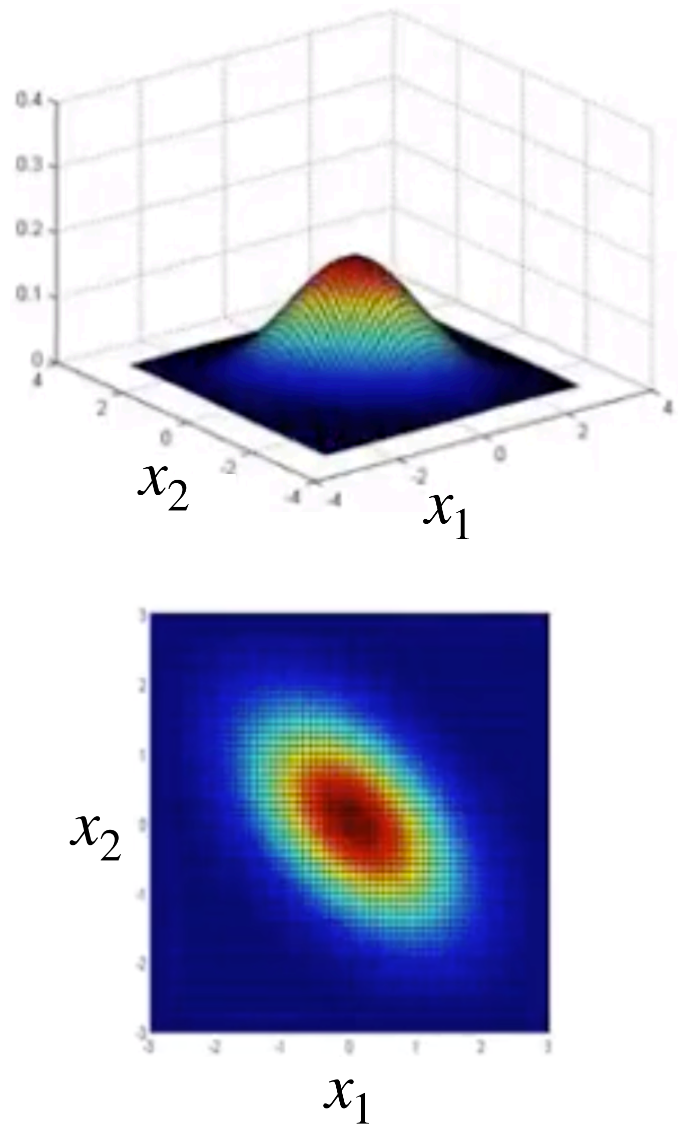


Multivariate Gaussian Distribution Example

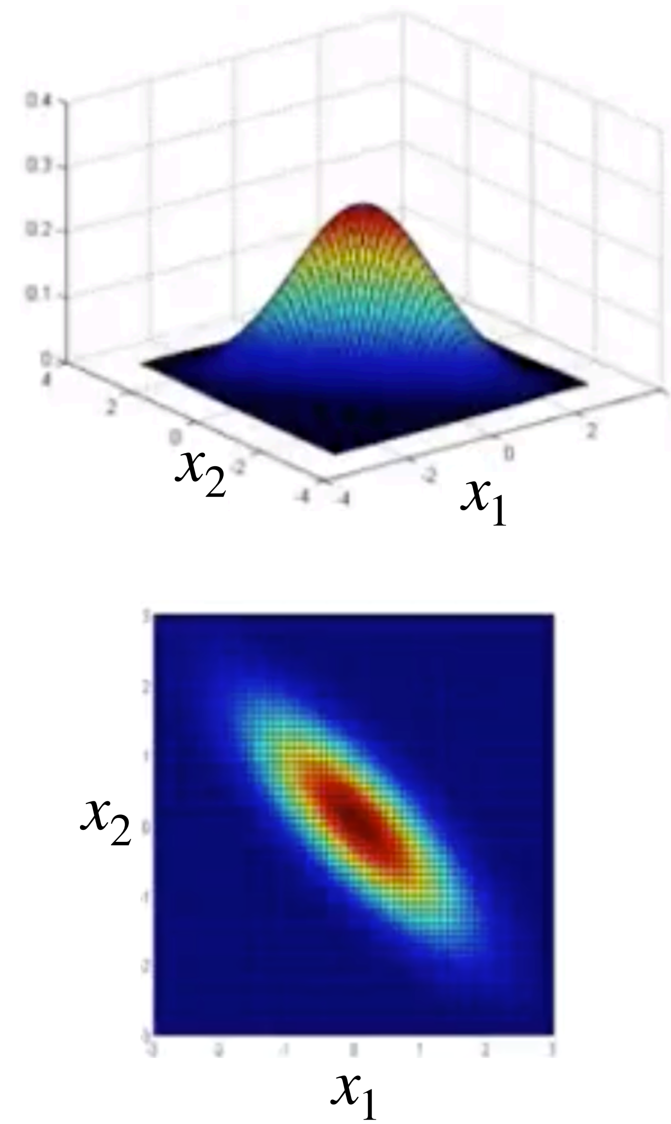
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

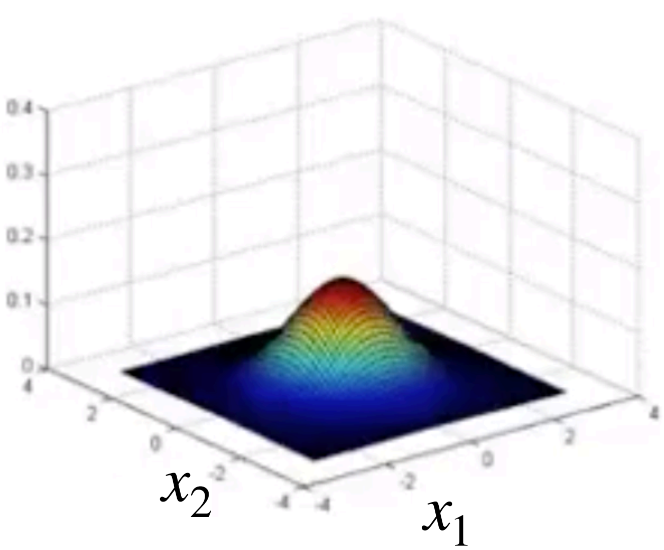


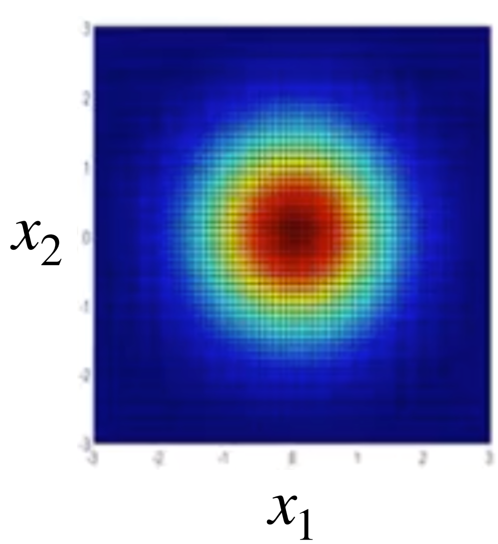
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



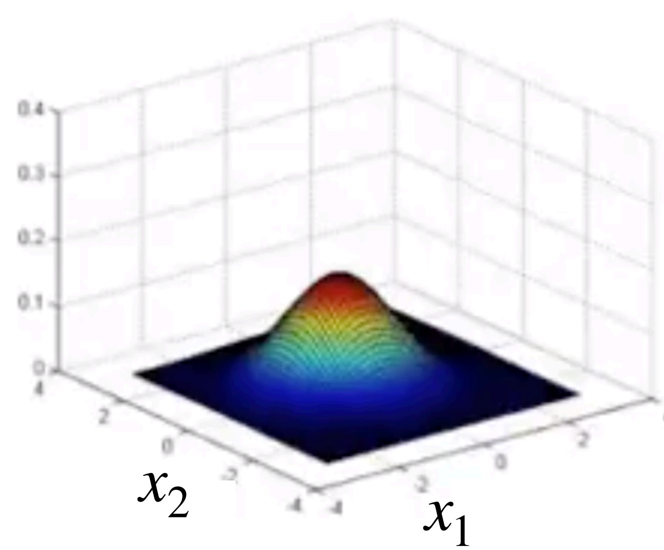
Multivariate Gaussian Distribution Example

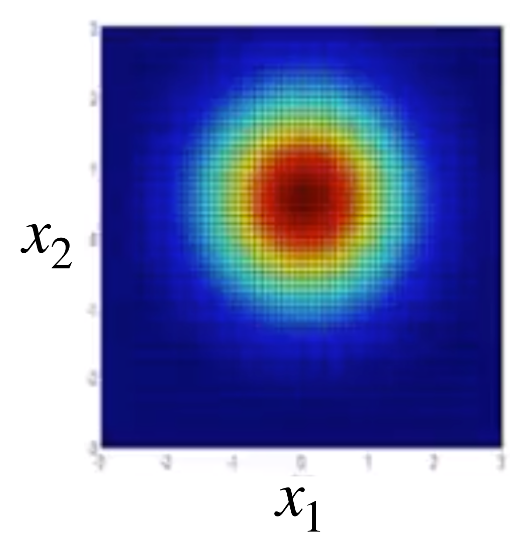
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



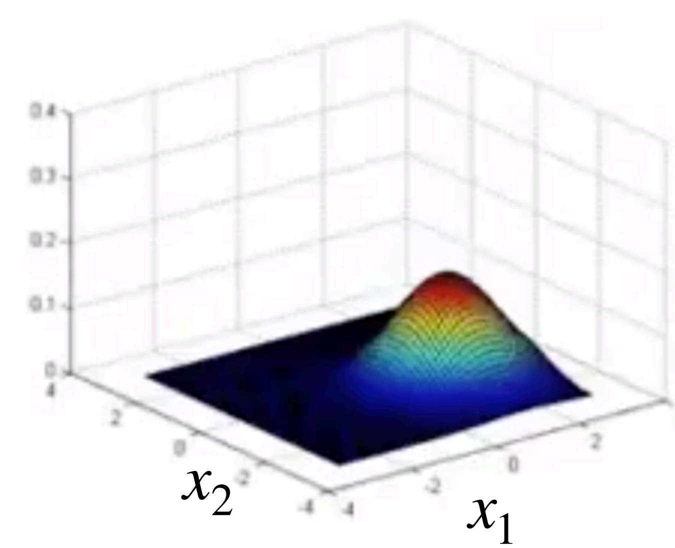


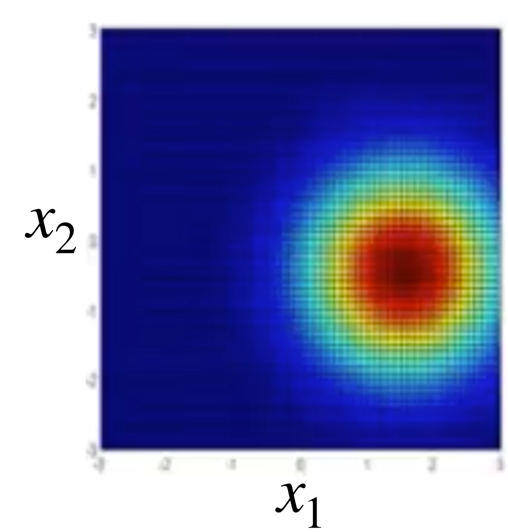
$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$





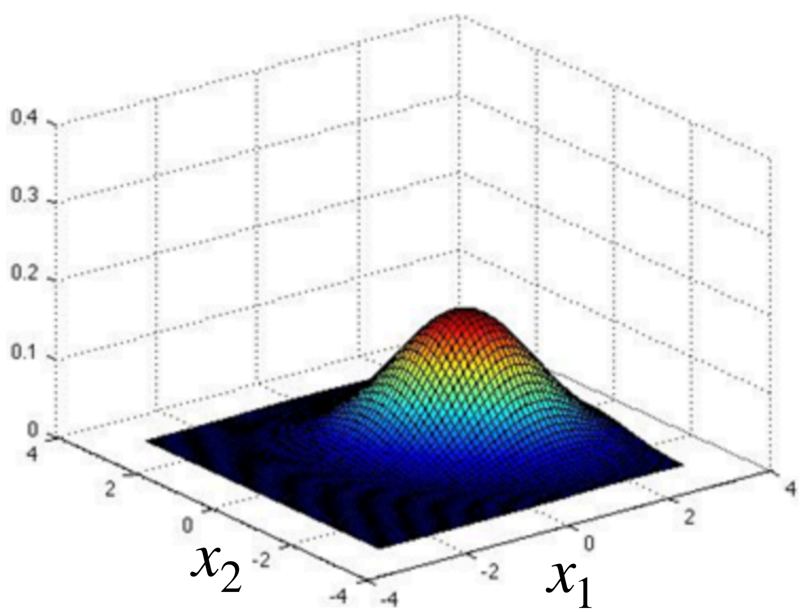
$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

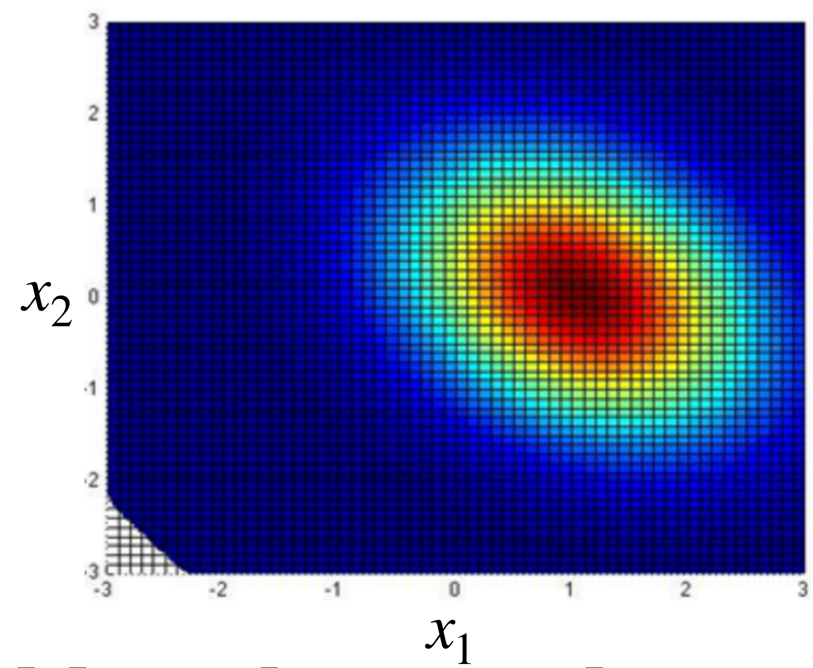




Question

Consider the following multivariate Gaussian, which of the following are the values of μ and Σ for this distribution?





(i) $\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$

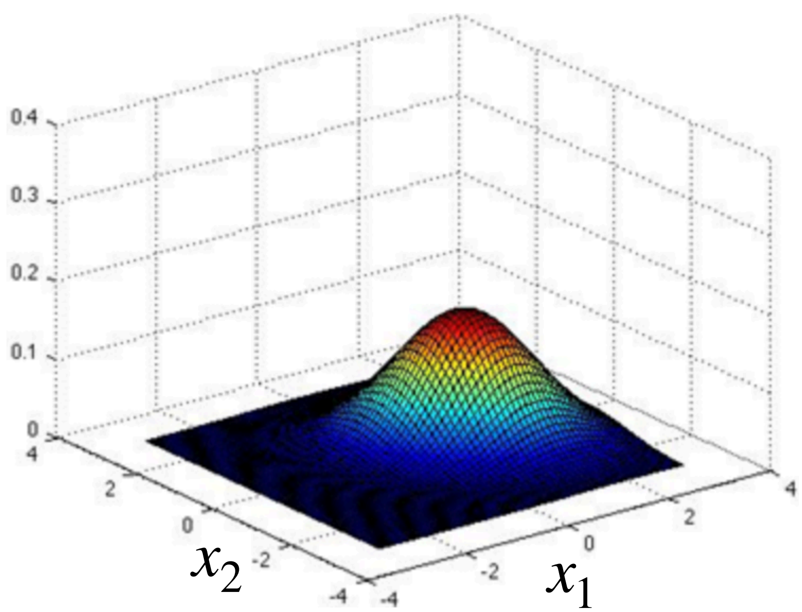
(ii) $\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$

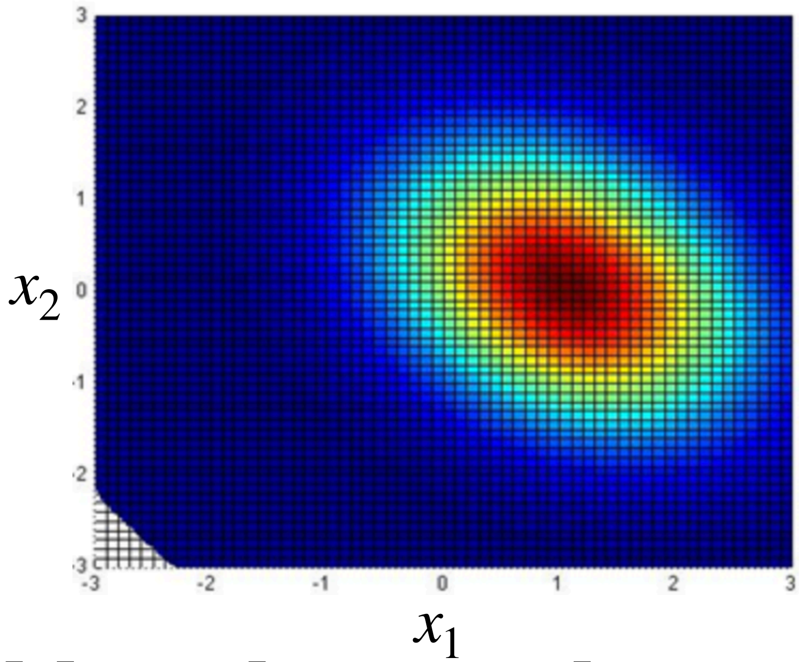
(iii) $\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}$

(iv) $\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}$

Question

Consider the following multivariate Gaussian, which of the following are the values of μ and Σ for this distribution?





(i) $\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$

(ii) $\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$

(iii) $\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}$

(iv) $\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}$

Now, we’re ready to talk about GDA

Gaussian Discriminant Analysis

Now, the GDA model for a 2-class problem is:

$$y \sim \textbf{Bernoulli}(\phi)$$
$$(\boldsymbol{x} \mid y = 0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$$
$$(\boldsymbol{x} \mid y = 1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$$

We can then write the distributions as follows:

$$p(y) = \phi^y (1 - \phi)^{1-y}$$
$$p(x \mid y = 0) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \Sigma_0^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_0)}$$
$$p(x \mid y = 1) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_1)}$$

Gaussian Discriminant Analysis

Now, it's time to optimize !

$$
\begin{aligned}
l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1) &= \log \Pi_{i=1}^{m} p(\boldsymbol{x}^{(i)}, y^{(i)}; \phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1) \\
&= \log \Pi_{i=1}^{m} p(\boldsymbol{x}^{(i)} \,|\, y^{(i)}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1) p(y^{(i)}; \phi)
\end{aligned}
$$

Gaussian Discriminant Analysis

It turns out that if we assume $\Sigma_0 = \Sigma_1 = \Sigma$, we obtain the maximum likelihood estimates as follows:

$$\phi = \frac{1}{m}\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\}$$

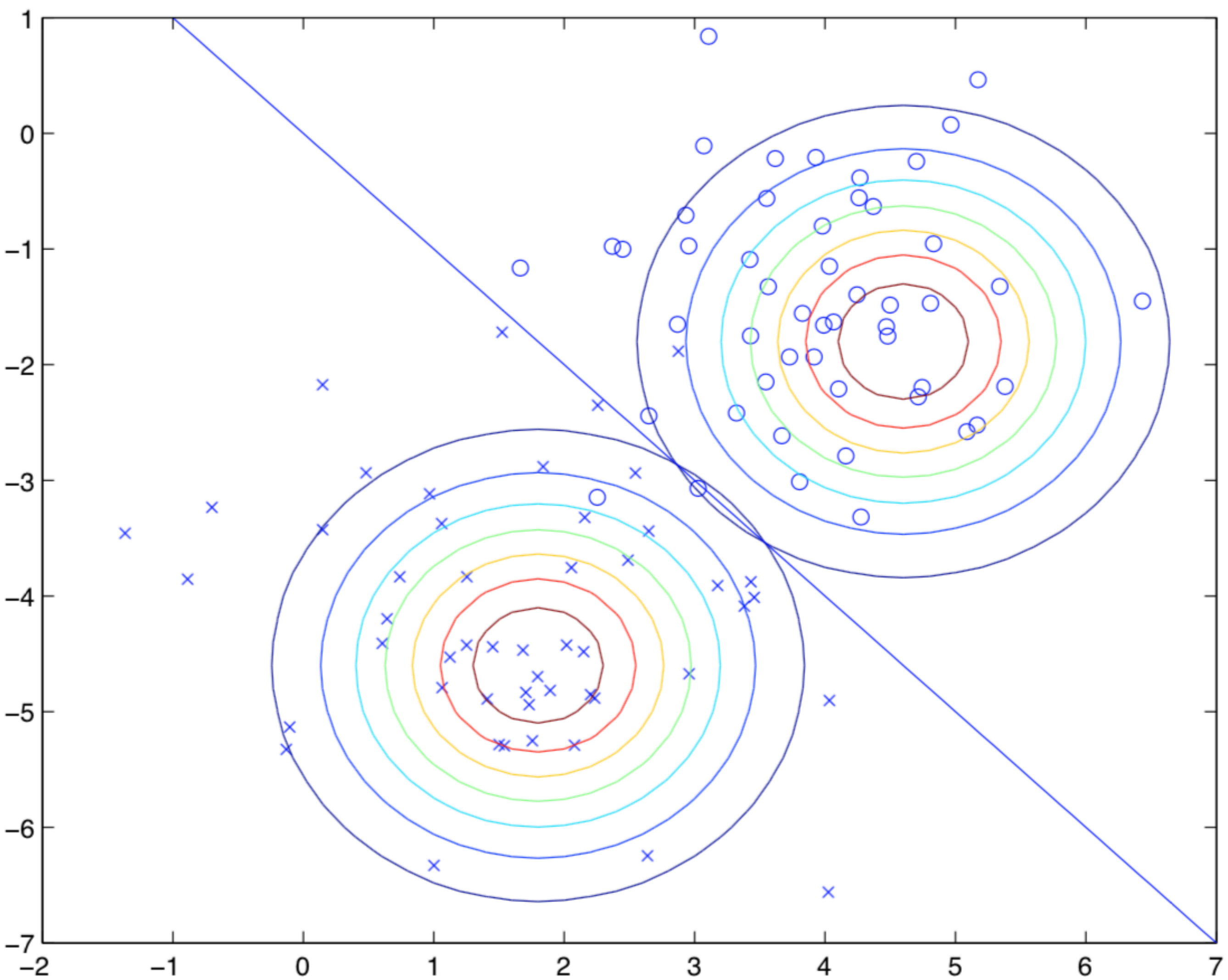
$$\boldsymbol{\mu}_0 = \frac{\Sigma_{i=1}^{m} 1\{y^{(i)} = 0\} x^{(i)}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

$$\boldsymbol{\mu}_1 = \frac{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

$$\Sigma = \frac{1}{m}\Sigma_{i=1}^{m} (\boldsymbol{x}^{(i)} - \mu_{y^{(i)}})(\boldsymbol{x}^{(i)} - \mu_{y^{(i)}})^T$$

Gaussian Discriminant Analysis

Pictorially, what the algorithm is doing can be seen in as follows:



- Two Gaussians have the same shape and orientation since they share Σ; they have different μ.
- The straight line is the decision boundary at which $p(y = 1 \mid x) = 0.5$. On one side of the boundary, we'll predict $y = 1$; on the other side, we'll predict $y = 0$.

Gaussian Discriminant Analysis

The assumption $\Sigma_0 = \Sigma_1 = \Sigma$ turns out to be very useful
i.e. it is possible to show that:

$$p(y = 1 \,|\, \boldsymbol{x}; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

with $\boldsymbol{\theta}$ as a function of $\phi, \Sigma, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1$.
This shows that the GDA model is related to the logistic regression.

Gaussian Discriminant Analysis

The assumption $\Sigma_0 = \Sigma_1 = \Sigma$ turns out to be very useful
i.e. it is possible to show that:

$$p(y = 1 \,|\, \boldsymbol{x}; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

with $\boldsymbol{\theta}$ as a function of $\phi, \Sigma, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1$.
This shows that the GDA model is related to the logistic regression.

GDA and logistic regression have the same form but give different decision boundaries. We'll later practice about it in the lab !

GDA *vs.* Logistic Regression

- GDA will be better (*i.e.* will require less training data to provide accurate predictions) if $p(\boldsymbol{x} \mid y)$ is in fact multivariate Gaussian or almost multivariate Gaussian.

- Logistic regression will probably be better if $p(\boldsymbol{x} \mid y)$ is definitely non-Gaussian or unknown.

Summary

Let's again modify our check-list reminder !

- If you have continuous $\mathcal{X}$ and continuous $\mathcal{Y}$, your first go-to model should be linear regression. Also, consider non-linear transformation of the inputs.

- If you have continuous $\mathcal{X}$ and discrete $\mathcal{Y}$ but don't know much about $p(\boldsymbol{x} \mid y)$, your first go-to model should be logistic or softmax regression, or may come up with a new GLM from scratch.

- If you have continuous $\mathcal{X}$ and discrete $\mathcal{Y}$ and know something about $p(\boldsymbol{x} \mid y)$, you should model the distribution accurately, as a Gaussian (GDA) or build a new generative model from scratch.

Naive Bayes Classifier

Naive Bayes

GDA was an example of a generative classification method for problems in which $\mathscr{X} = \mathbb{R}^n$.

What if our features have discrete values? *e.g.*

- Car buying behavior:
  - ซื้อ, ไม่ซื้อ
    $\mathscr{Y} := \{\textbf{buy}, \neg\textbf{buy}\}$ and $\mathscr{X} := [\textbf{Size}, \textbf{Color}, \textbf{Gender}, \textbf{Age}]$, where $\textbf{Size} := \{\textbf{small}, \textbf{medium}, \textbf{large}\}$, $\textbf{Color} := \{\textbf{red}, \textbf{green}, \textbf{blue}\}$, $\textbf{Gender} := \{\textbf{male}, \textbf{female}\}$, $\textbf{Age} := \{\textbf{0 - 18}, \textbf{19 - 39}, \textbf{30 - 39}, \textbf{40+}\}$
- Email spam filter:
  - $\mathscr{Y} := \{\textbf{spam}, \neg\textbf{spam}\}$ and $\mathscr{X} := \{0,1\}^K$, where each variable $x_1, \ldots, x_k$ representing presence/absence of a particular word in a vocabulary.

one hot vactor

Naive Bayes

Whatever $\mathscr{Y}$ and $\mathscr{X}$, a generative model needs a form for $p(y|\boldsymbol{x})$.

The simplest approach to these problems is to use the multinomial distribution over the set of possible outcomes for $\boldsymbol{x}$.

Exercise: How many outcomes for $\boldsymbol{x}$ are there for the spam filter examples?

Suppose a vocabulary contains 50000 words, then $\boldsymbol{x} \in \{0,1\}^{50000}$ *i.e.* $\boldsymbol{x}$ is a 50000-dimensional vector of 0's and 1's.

If we were to model x explicitly with a multinomial distribution over the 2^{50000} possible outcomes, then we'd end up with a $(2^{50000} - 1)$ dimensional parameter vector. This is clearly too many parameters !

Naive Bayes

Naive Bayes attempts to reduce the number of parameters required using the (very strong but very useful) assumption that the features are conditionally independent given y:

$$
\begin{aligned}
p(\boldsymbol{x} \mid y) &= p(x_1, x_2, \ldots, x_n \mid y) \quad \text{xi = คำใน mail} \\
&= p(x_1 \mid y) \cdot p(x_2 \mid x_1, y) \cdot \ldots \cdot p(x_n \mid x_1, x_2, \ldots, x_{n-1}, y) \\
&\approx \Pi_{i=1}^{n} p(x_i \mid y)
\end{aligned}
$$

This model requires a set of parameters $\phi_{ij|y=1} = p(x_i = j \mid y = 1)$ and a set of parameters $\phi_{ij|y=0} = p(x_i = j \mid y = 0)$.

Noted that if the variables x_i are binary, we only need two parameters $\phi_{i|y=1} = p(x_i = 1 \mid y = 1)$ and $\phi_{i|y=0} = p(x_i = 1 \mid y = 0)$.

Naive Bayes

For the case where x_i are binary, we get the joint likelihood:

$$\mathscr{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \Pi_{i=1}^{m} p(\boldsymbol{x}^{(i)}, y^{(i)})$$

Maximizing this w.r.t. $\phi_y, \phi_{i|y=0}, \phi_{i|y=1}$ gives the following maximum likelihood estimates:

$$\phi_{j|y=1} = \frac{\Sigma_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = \frac{\Sigma_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

จำนวนของ spam email = $\phi_y = \dfrac{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$

1{.} = ถ้าค่าใน {} ถูกให้ตอบ 1

Naive Bayes

If n is large, there may be some features that do not appear in all combinations with every class.

Example:

- The term 'viagra' might never occur before.
- In a given period of time, there might not be any customers 18 years or younger who bought a car.

Suppose we are then faced with an email $\boldsymbol{x}$ with the term 'viagra' or a customer $\boldsymbol{x}$ whose age is 0-18.

Do you predict $p(\textbf{spam} \,|\, \boldsymbol{x}) = 1$ and $p(\textbf{spam} \,|\, \boldsymbol{x}) = 0$?

Naive Bayes

Assuming that 'viagra' was the 35000th word in the dictionary, your Naive Bayes spam filter therefore had picked its maximum likelihood estimates of the parameters $\phi_{35000|y}$ to be:

$$\phi_{35000|y=1} = \frac{\Sigma_{i=1}^{m} 1\{x_{35000}^{(i)} = 1 \wedge y^{(i)} = 1\}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\}} = 0$$

$$\phi_{35000|y=0} = \frac{\Sigma_{i=1}^{m} 1\{x_{35000}^{(i)} = 1 \wedge y^{(i)} = 0\}}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 0\}} = 0$$

Hence, the class posterior probabilities yield:

$$p(y = 1 \,|\, x) = \frac{\Pi_{i=1}^{n} p(x_i \,|\, y) p(y = 1)}{\Pi_{i=1}^{n} p(x_i \,|\, y) p(y = 1) + \Pi_{i=1}^{n} p(x_i \,|\, y) p(y = 0)} = \frac{0}{0}$$

Naive Bayes

Laplace smoothing avoids 0-probability issues by adding one pseudo-example to the dataset:

$$\phi_{j|y=1} = \frac{\Sigma_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\} + 1}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$

$$\phi_{j|y=0} = \frac{\Sigma_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\} + 1}{\Sigma_{i=1}^{m} 1\{y^{(i)} = 0\} + 2}$$

This is for binary features.

While Naive Bayes works well for many classification problems, for text classification, there is a related model that does even better.

Event Models

Naive Bayes as presented uses the multi-variate Bernoulli event model.

E.g. first it is randomly determined (according to the class priors $p(y)$) whether a spammer or non-spammer will send you your next message. Then, the person sending the email runs through the dictionary, deciding whether to include each word i in that email independently and according to the probabilities $p(x_i = 1 \mid y) = \phi_{i|y}$.

Thus, the probability of a message was given by $p(y)\Pi_{i=1}^{n} p(x_i \mid y)$.

The multinomial event model can be defined in a similar manner *i.e.*

$$p(y)\Pi_{i=1}^{n} p(x_i \mid y)$$

2 สมการ เหมือนกัน แต่คนละเรื่อง

for the probability of a message. Note that this is a different formula from the above since $x_i \mid y$ is now a multinomial.

Event Models

เมื่อได้ model มาแล้ว จะหา estimates ที่ดีที่สุด

The likelihood of the data is given by:

$$
\begin{aligned}
\mathcal{L}(\phi, \phi_{k|y=0}, \phi_{k|y=1}) &= \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \prod_{i=1}^{m} \left(\prod_{j=1}^{n_i} p(x_j^{(i)} | y; \phi_{k|y=0}, \phi_{k|y=1}) \right) p(y^{(i)}; \phi_y)
\end{aligned}
$$

Event Models

Maximizing yields the best likelihood estimates as follows:

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i + |V|}$$

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i + |V|}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}.$$

Here, n_i is the number of words in the i-training example and $|V|$ is the size of our vocabulary (we use the blue font to indicate an application of Laplace smoothing).

Summary

Let's again modify our check-list reminder !

- If you have continuous $\mathscr{X}$ and continuous $\mathscr{Y}$, your first go-to model should be linear regression. Also, consider non-linear transformation of the inputs.

ไม่มีข้อมูลเพียงพอ

- If you have continuous $\mathscr{X}$ and discrete $\mathscr{Y}$ but don't know much about $p(\boldsymbol{x} \mid y)$, your first go-to model should be logistic or softmax regression, or may come up with a new GLM from scratch.

มีข้อมูลเพียงพอ

- If you have continuous $\mathscr{X}$ and discrete $\mathscr{Y}$ and know something about $p(\boldsymbol{x} \mid y)$, you should model the distribution accurately, as a Gaussian (GDA) or build a new generative model from scratch.
- If you have discrete $\mathscr{X}$ and $\mathscr{Y}$, you should probably start with naive Bayes and build up from there.

That’s it for Generative Learning !

We’ll come back to Discriminative Learning again when we discuss about kernel methods and SVMs.