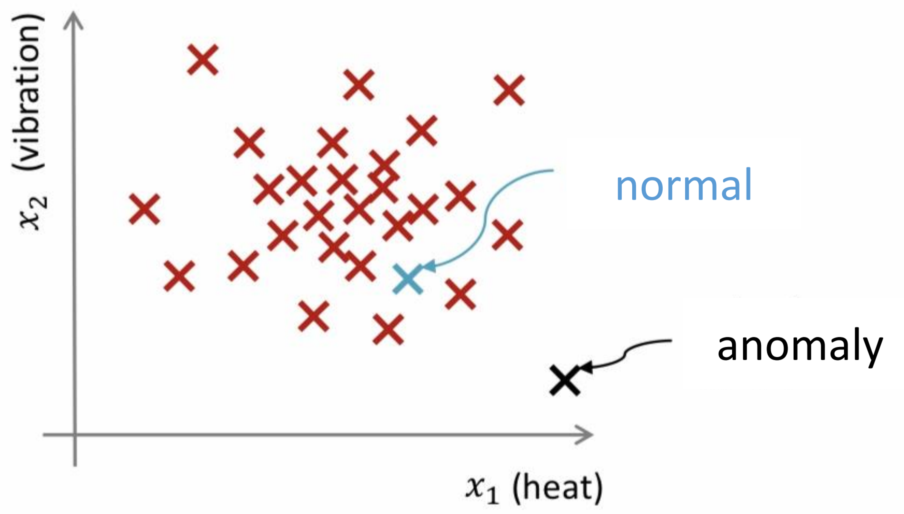
Machine Learning

By Ghazal Lalooha

Anomaly detection

Introductory example

- Features related to the aircraft engine in the safety test:
  - Heat produced
  - Vibration intensity
  - ...
- New engine: x_{test}



Training set:

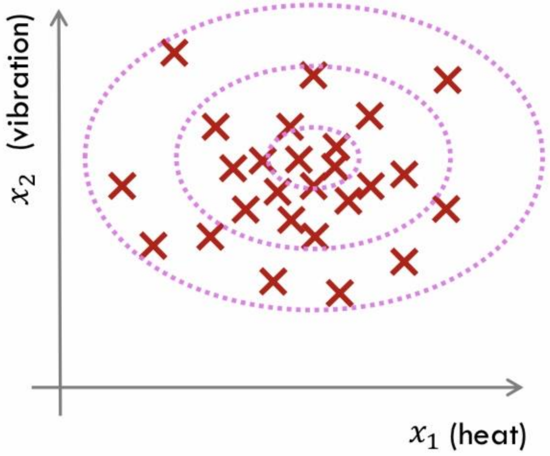
$$\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$$

Distribution estimation: classification with one class

Data set:

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$$

Question: is x_{test} anomaly?



Goal: creating a probabilistic model like p(x) so that:

If $p(x_{test}) < e$, x_{test} is anomaly

If $p(x_{test}) \geq e$, x_{test} is normal

Example applications

- Fraud detection:
  - Features: Information about user activities i
    - The number of visits to the website
    - Number of purchases made per visit
    - ...
  - Objective: To detect unusual users by checking for which user $p(x) < e$.
- Manufacturing:
  - Features: Information about manufactured products
  - Purpose: to detect abnormal products

Example applications

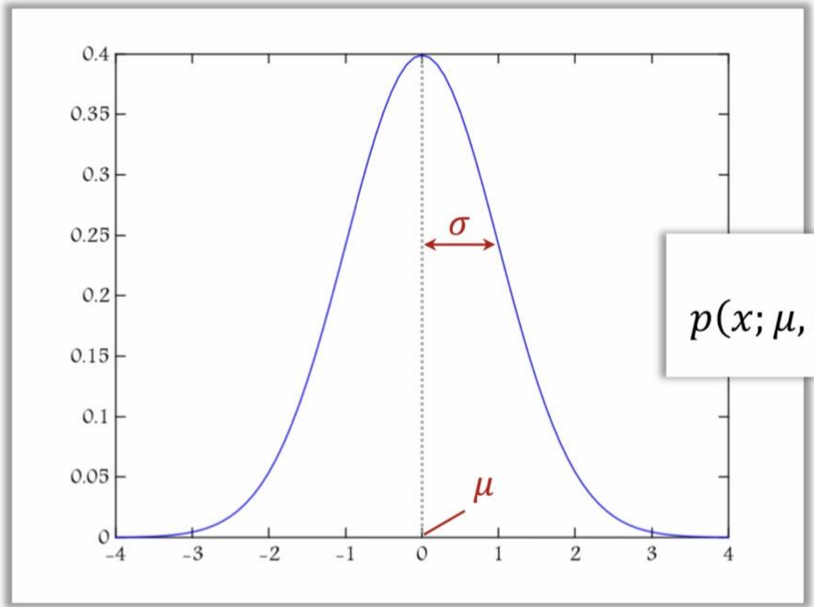
- Monitor computers in data centers:
  - Features: Information about the performance of the machine i
    - Memory consumption
    - The number of disk accesses per second
    - CPU load
    - The ratio of processor load to network traffic
    - ...
  - Purpose: to detect unusual behaviors
    - Like being stuck in an infinite loop and...

Gaussian Distribution (Normal)

Gaussian Distribution (Normal)

Gaussian distribution:
Suppose x has a Gaussian distribution, in this formula M represents the mean and the square of sigma represents the variance.

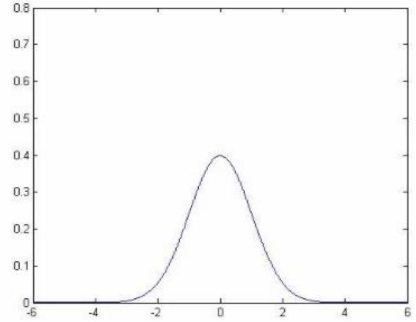
$$x \sim \mathcal{N}(\mu, \sigma^2)$$



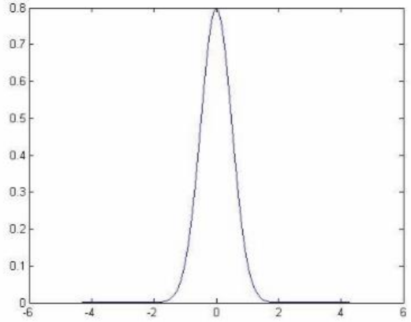
$$p(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Example

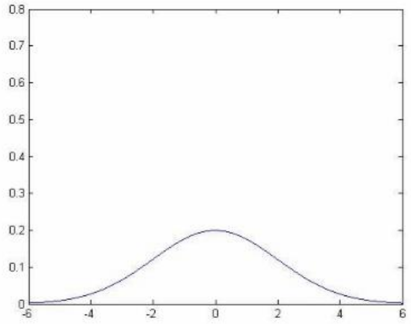
$\mu = 0, \sigma = 1$



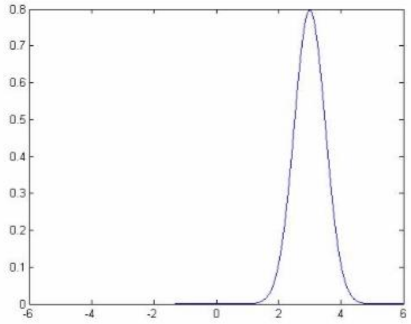
$\mu = 0, \sigma = 0.5$



$\mu = 0, \sigma = 2$



$\mu = 3, \sigma = 0.5$

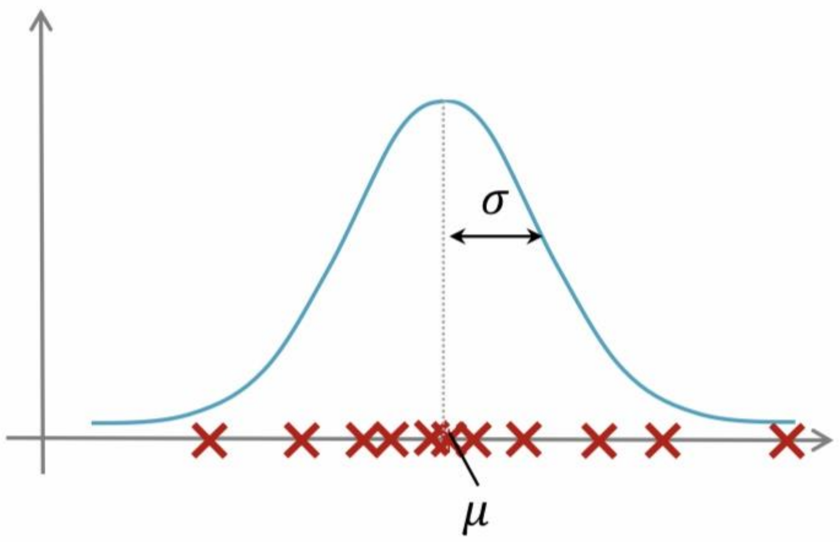


Parameter estimation

Data set:

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$$

Goal: Estimating the values of M and sigma:



$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu\right)^2$$

Anomaly Detection Algorithm

Distribution estimation algorithm

teaching institution:

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}, \qquad x^{(i)} \in \mathbb{R}^n$$

$$x_j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right)$$

assumptions:
The characteristics follow a normal distribution.
There is no correlation between features.

$$p(\boldsymbol{x}) = p(x_1;\mu_1,\sigma_1^2)p(x_2;\mu_2,\sigma_2^2)p(x_3;\mu_3,\sigma_3^2)\cdots p(x_n;\mu_n,\sigma_n^2)$$

$$= \prod_{j=1}^{n} p\left(x_j;\mu_j,\sigma_j^2\right)$$

Anomaly Detection Algorithm

- Determining features that can be useful in diagnosing anomalies.
- Estimation of parameters (for 1 ≤ j ≤ n)

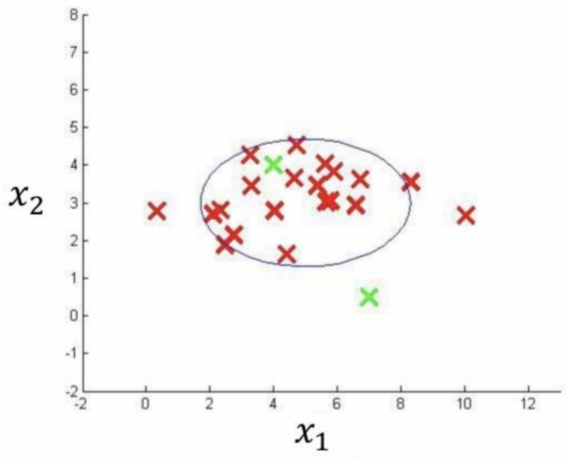
$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \qquad \sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$

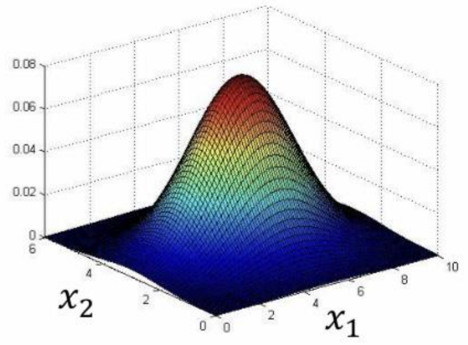
- Compute p(x) for new data x

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

- produce the output "yes" provided that p(x) < e

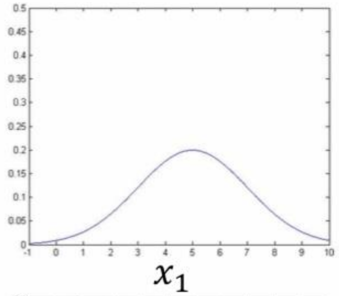
Example



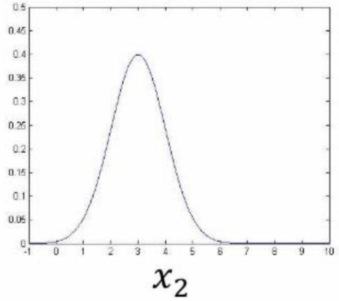


$\mu_1 = 5, \sigma_1 = 2$

$\mu_2 = 3, \sigma_2 = 1$



$p(x_1; \mu_1, \sigma_1^2)$



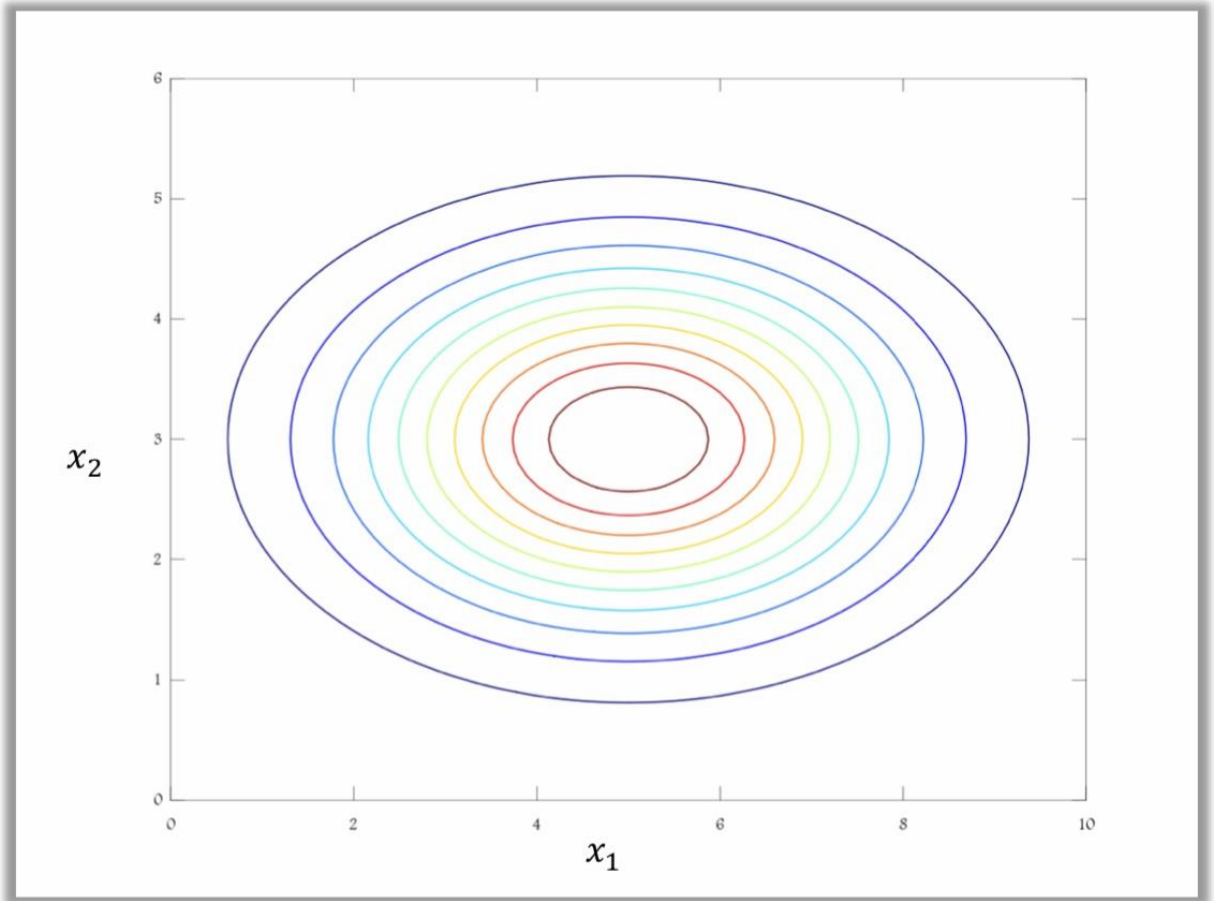
$p(x_2; \mu_2, \sigma_2^2)$

$\varepsilon = 0.02$

$p(x_{test}^{(1)}) = 0.0426$

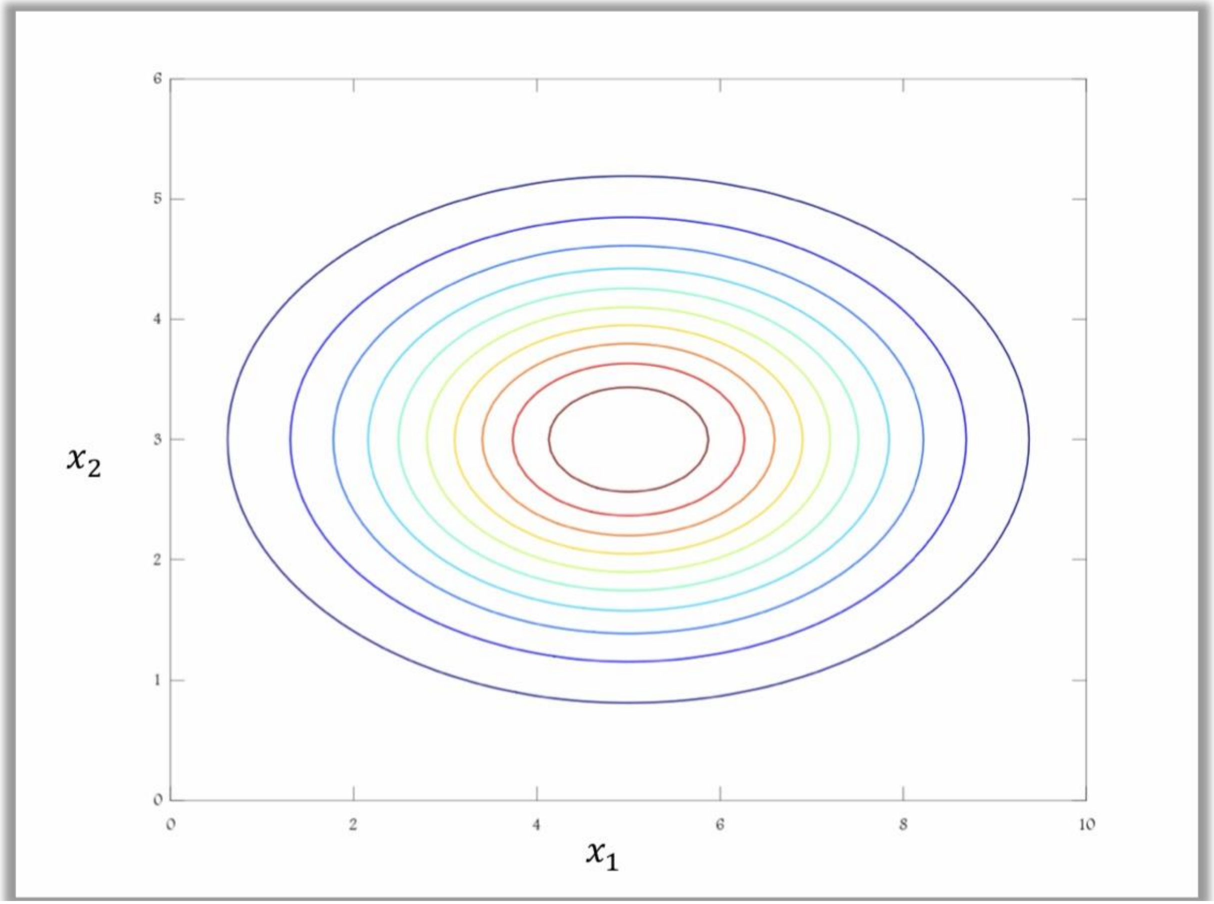
$p(x_{test}^{(2)}) = 0.0021$

Draw the simultaneous probability distribution



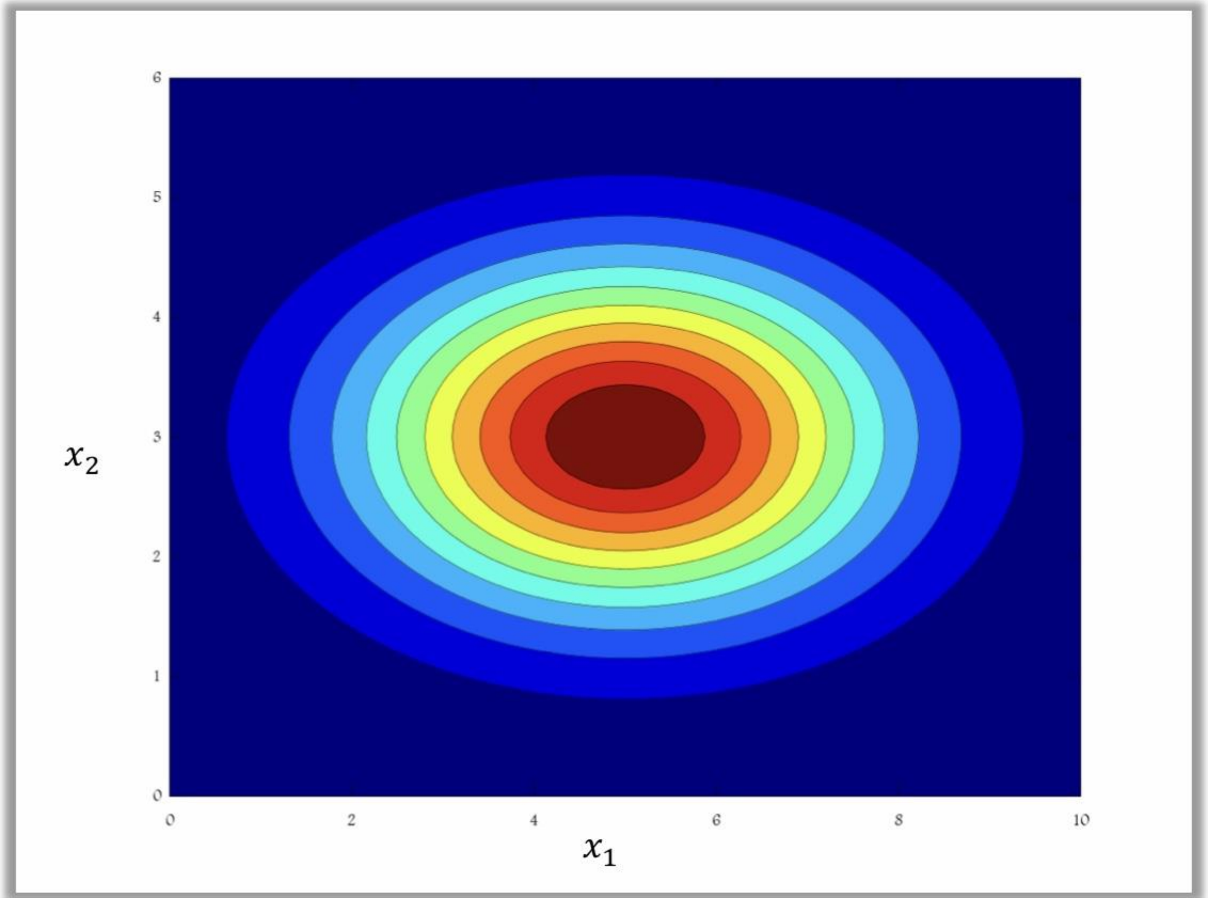

```
x1 = 0:0.1:10;
x2 = 0:0.1:6;
[X1 X2] = meshgrid(x1, x2);
Z1 = normpdf(X1, 5, 2);
Z2 = normpdf(X2, 3, 1);
Z = Z1 .* Z2;
contour(X1, X2, Z);
```

Drawing contours




```
x1 = 0:0.1:10;
x2 = 0:0.1:6;
[X1 X2] = meshgrid(x1, x2);
Z1 = normpdf(X1, 5, 2);
Z2 = normpdf(X2, 3, 1);
Z = Z1 .* Z2;
contour(X1, X2, Z);
```

Drawing contours




```
x1 = 0:0.1:10;
x2 = 0:0.1:6;
[X1 X2] = meshgrid(x1, x2);
Z1 = normpdf(X1, 5, 2);
Z2 = normpdf(X2, 3, 1);
Z = Z1 .* Z2;
contourf(X1, X2, Z);
```

Development and evaluation of anomaly detection systems

Numerical evaluations

- Importance:
  - During the development process of learning systems, if we have a method to evaluate the system, then many decisions (such as feature selection, etc.) will be much easier.
- Suppose we have some labeled data. So that for each data, its normality (y = 0) or abnormality (y = 1) is determined.
- training set (including normal data) $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$
- Validation set $\left\{\left(x_{cv}^{(1)}, y_{cv}^{(1)}\right), \left(x_{cv}^{(2)}, y_{cv}^{(2)}\right), \cdots, \left(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})}\right)\right\}$
- Test set $\left\{\left(x_{test}^{(1)}, y_{test}^{(1)}\right), \left(x_{test}^{(2)}, y_{test}^{(2)}\right), \cdots, \left(x_{test}^{(m_{test})}, y_{test}^{(m_{test})}\right)\right\}$

Example

- Data set: information about the performance of engines
  - 10,000 healthy engine
  - 20 defective engine
- Classification:
  - Educational set: 6000 healthy engines
  - Validation set: 2000 healthy motors and 10 defective motors
  - Test set: 2000 healthy motors and 10 defective motors

Algorithm evaluation

- Training: Development of the p(x) model according to the training set
- Prediction: For samples in the validation or training set

$$y = \begin{cases} 1, & p(x) < \varepsilon \\ 0, & p(x) \geq \varepsilon \end{cases}$$

- Possible evaluation criteria:
  - True positive, false positive, true negative, false negative
  - Accuracy rate and recall rate
  - F1 score
- Note: Validation set can be used to choose a suitable value for epsilon.

Anomaly detection or supervised learning?

Anomaly detection or supervised learning?

- Anomaly diagnosis:
  - number of samples:
    - The number of positive to negative samples is very low
  - Very different "types" of anomalies:
    - For any algorithm, it is very difficult to learn anomalies from a small number of positive samples.
    - New anomalies may bear no resemblance to previously seen anomalies.
- Supervised learning:
  - number of samples:
    - A large number of positive and negative samples
  - Positive examples:
    - The number of positive samples is enough for the algorithm to get a correct understanding of the positive samples.
    - New positive examples are likely to be similar to positive examples that the algorithm has previously encountered during the training process.

Anomaly detection or supervised learning?

- Anomaly diagnosis:
  - Fraud detection
  - Construction and production (making airplane engines)
  - Monitoring machines in data centers
  - ...
- Supervised learning:
  - Spam detection
  - Weather forecast
  - Diagnosis of malignant cancerous glands
  - ...

Feature Selection

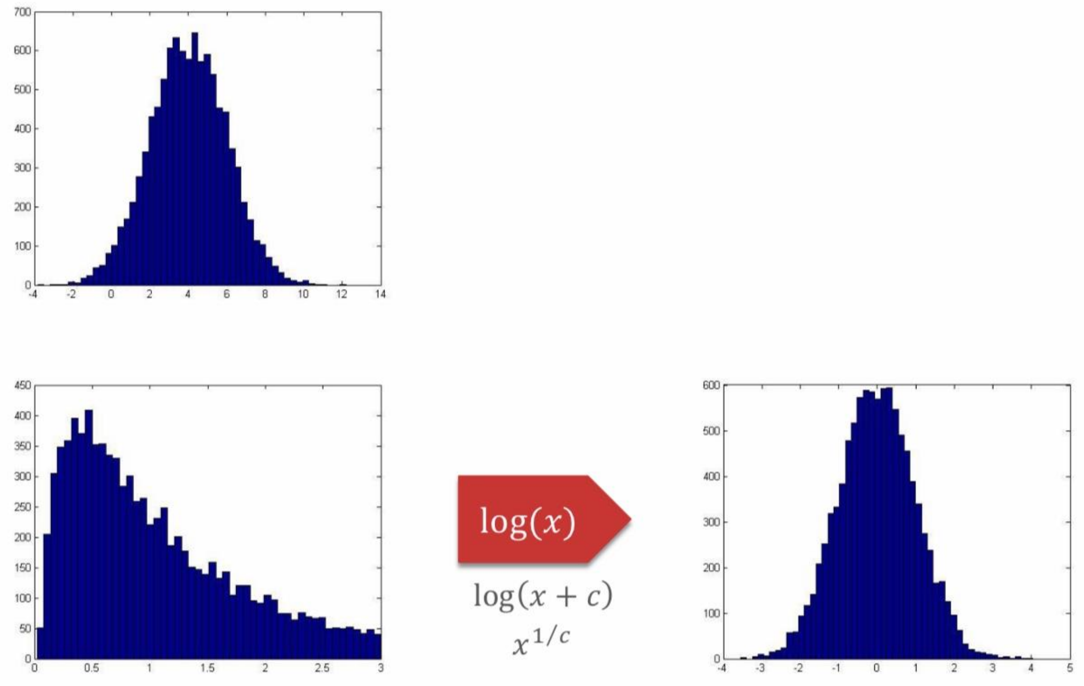
Non-Gaussian Features

- Features with

Gaussian distribution

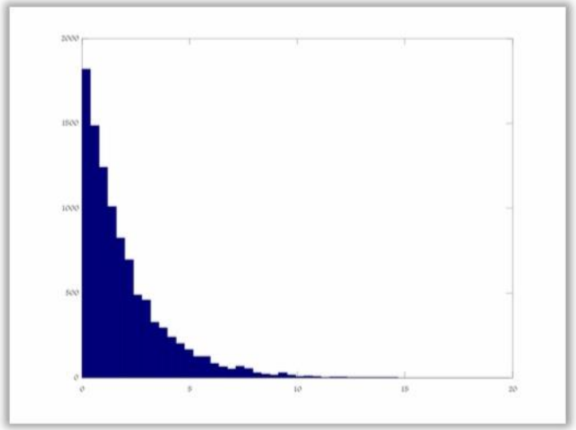
- Features with

non- Gaussian distribution

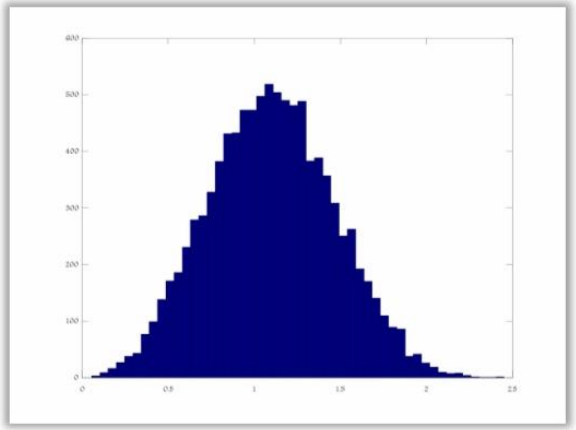


Converting to Gaussian distribution

```
>> x = gamrnd(1, 2, [10000 1]);
>> hist(x, 50);
```

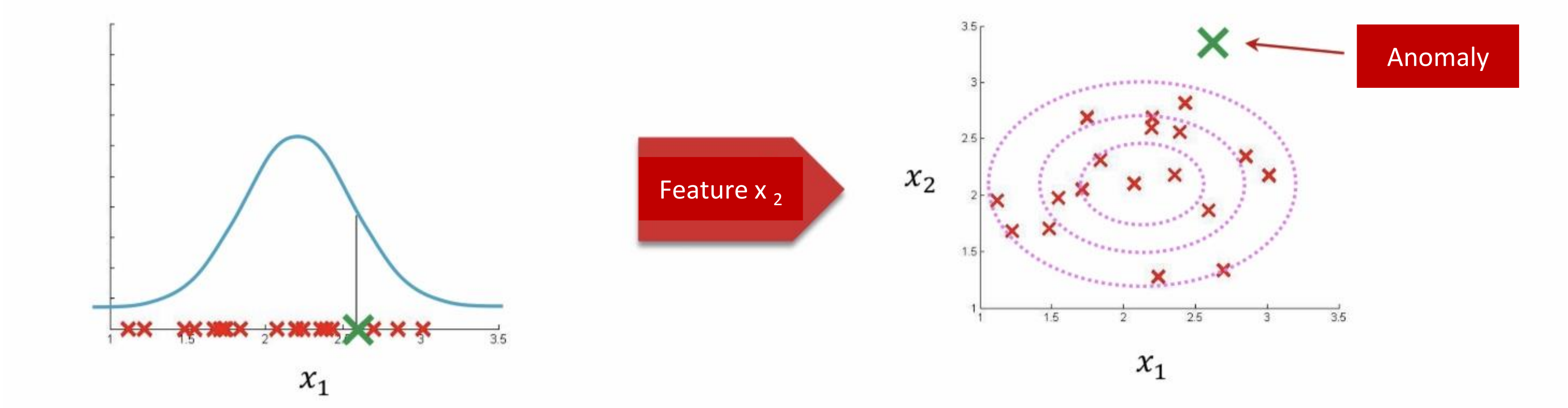


```
>> hist(x .^ 0.3, 50);
>> xnew = x .^ 3;
```



Error analysis to aid in anomaly detection

- Objective: We want the value of p(x):
  - be large for normal data.
  - be small for unusually small data.
- A common problem:
  - p(x) is not significantly different for normal and non-normal data.



Monitor computers in data centers

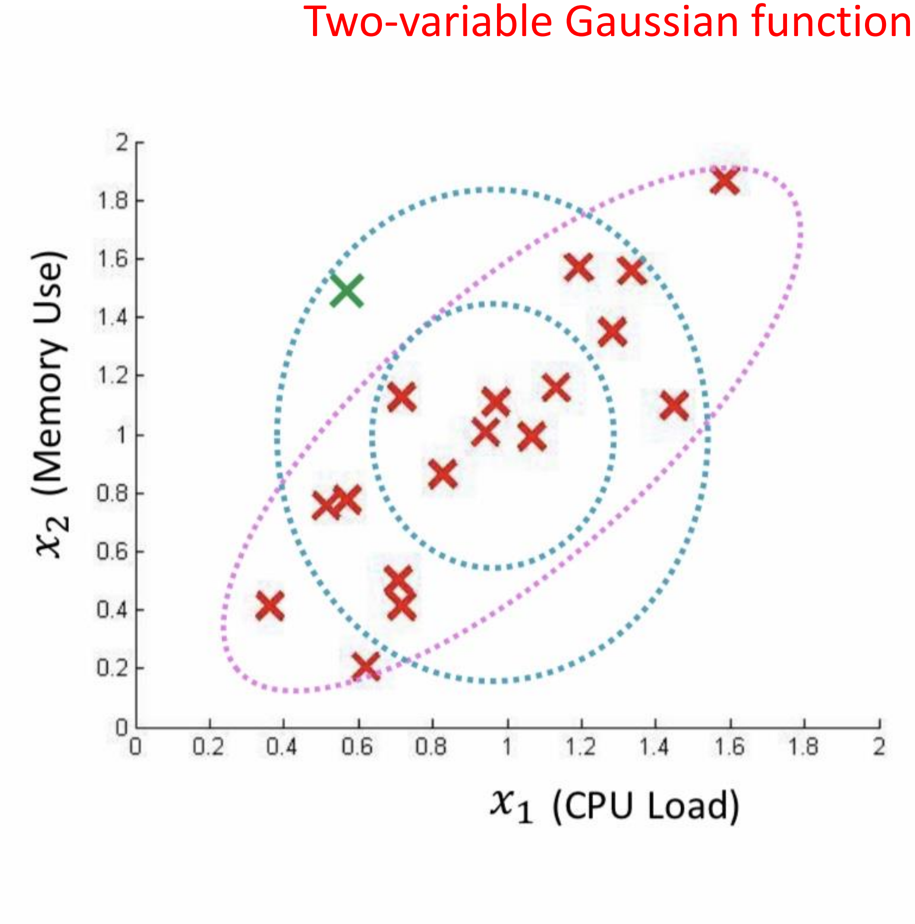
- Selection of features: Select features that are very small or very large if there is an anomaly.
  - Memory consumption
  - The number of disk accesses per second
  - CPU load
  - Network traffic
- Adding new features to detect unusual conditions:
  - The ratio of processor load to network traffic

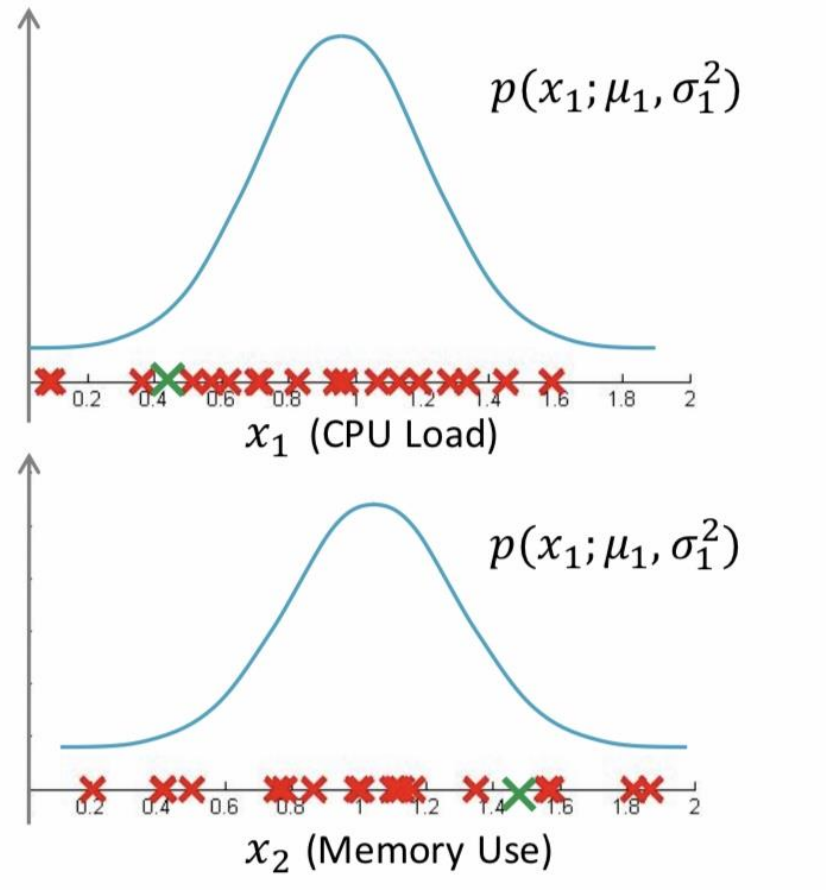
  (For example, if the processor is stuck in an infinite loop, the value of this property will be very large.)

Multivariate Gaussian distribution

Introductory example

Two-variable Gaussian function





Multivariate Gaussian distribution

Multivariate Gaussian function:

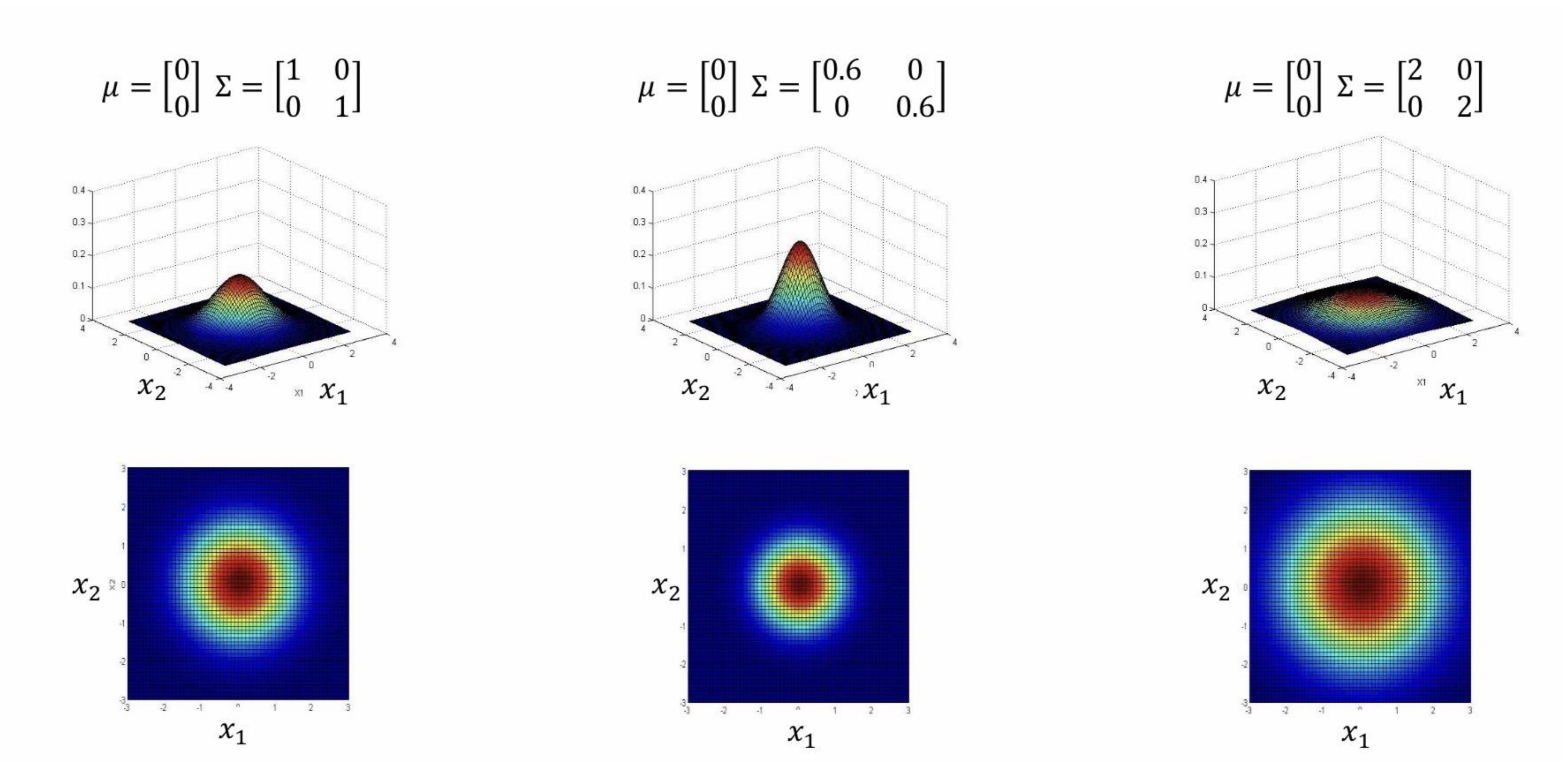
$$p(x;\mu,\Sigma) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Parameters:

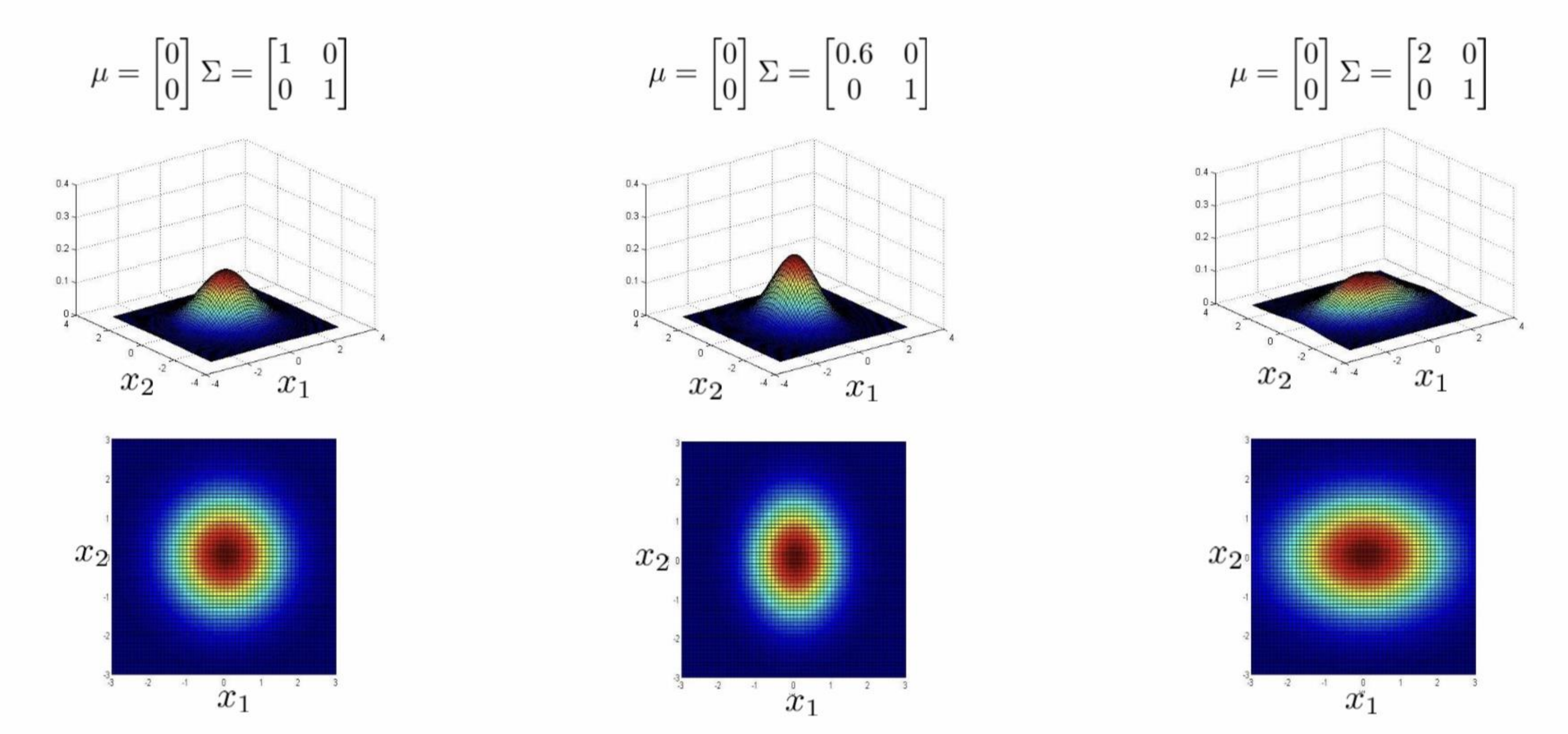
Covariance matrix

$\mu \in \mathbb{R}^n$ $\Sigma \in \mathbb{R}^{n\times n}$

Some examples



Some examples

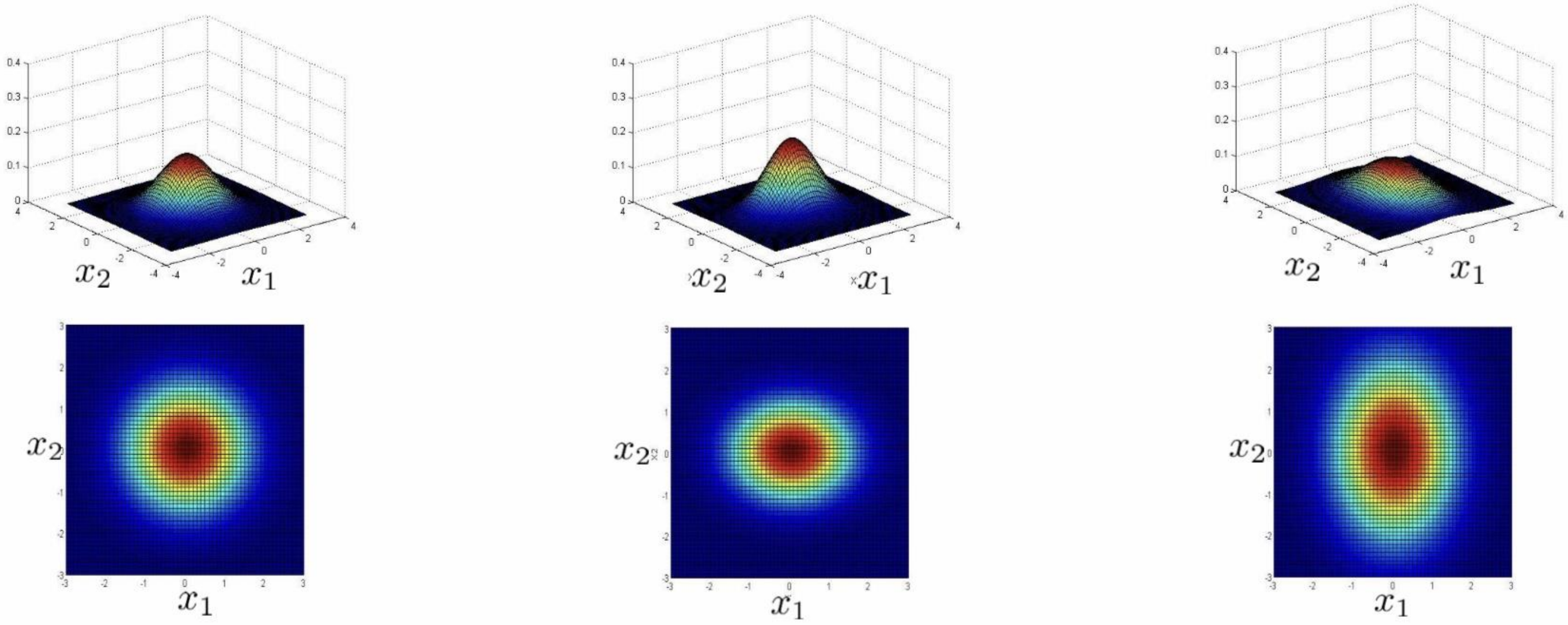


Some examples

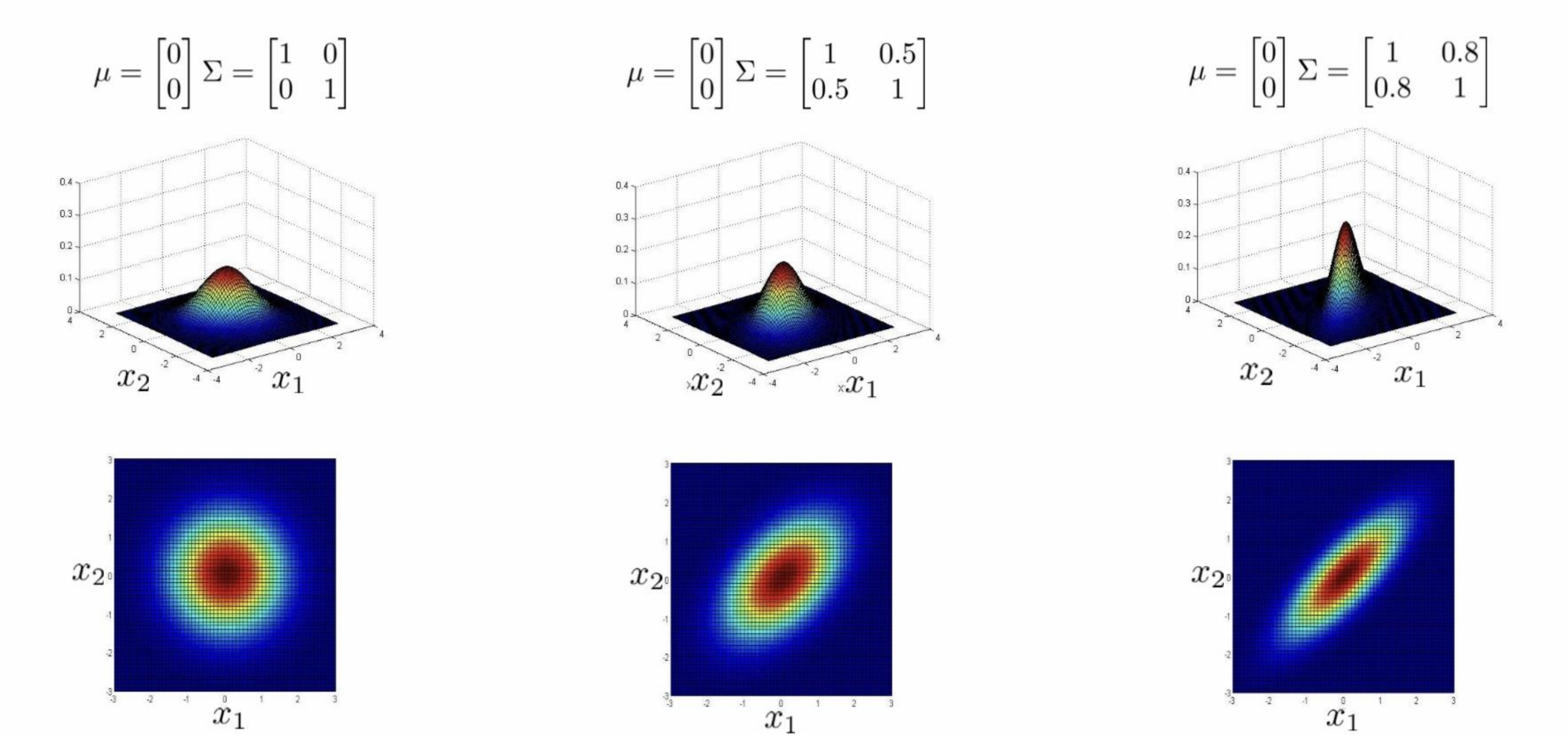
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$$

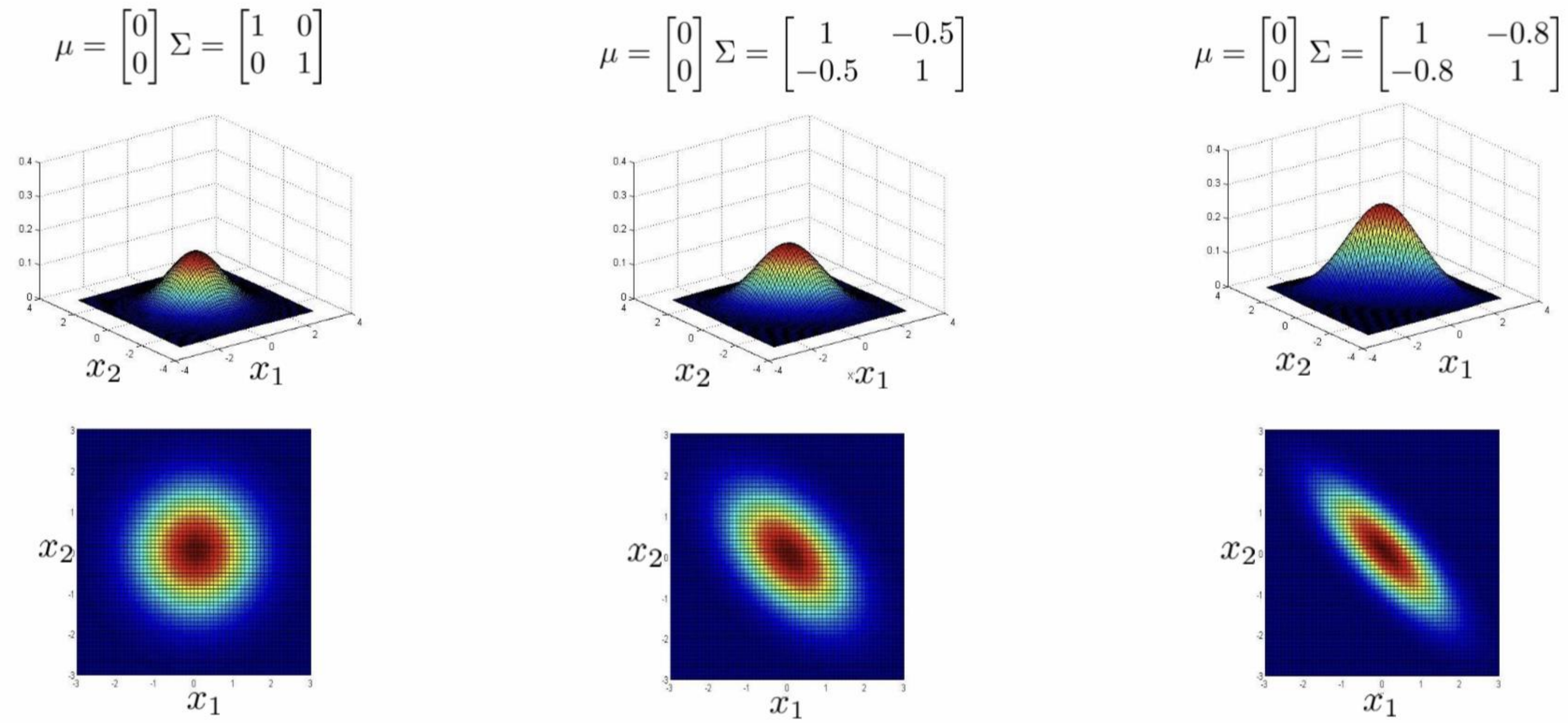
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



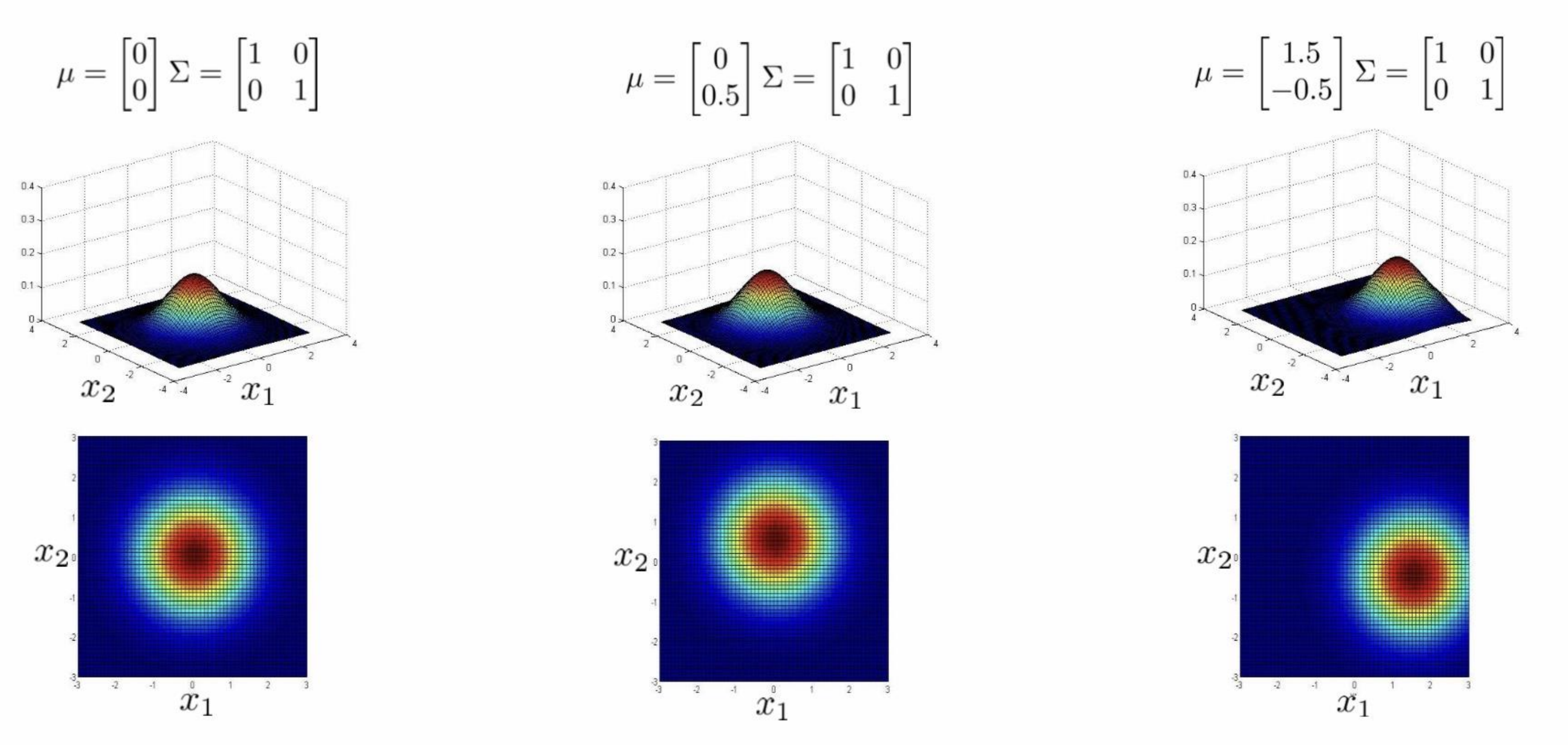
Some examples



Some examples



Some examples

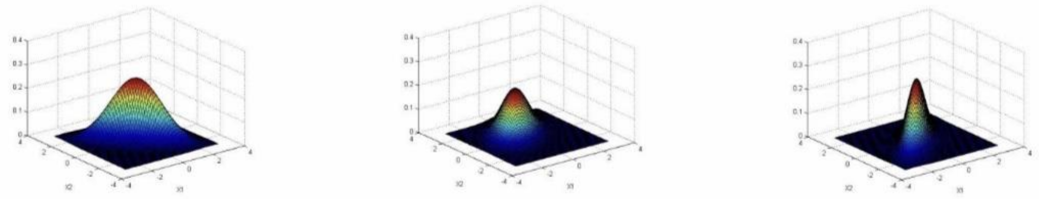


Anomaly detection with multivariate Gaussian function

Multivariate Gaussian distribution

Multivariate Gaussian distribution function:

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

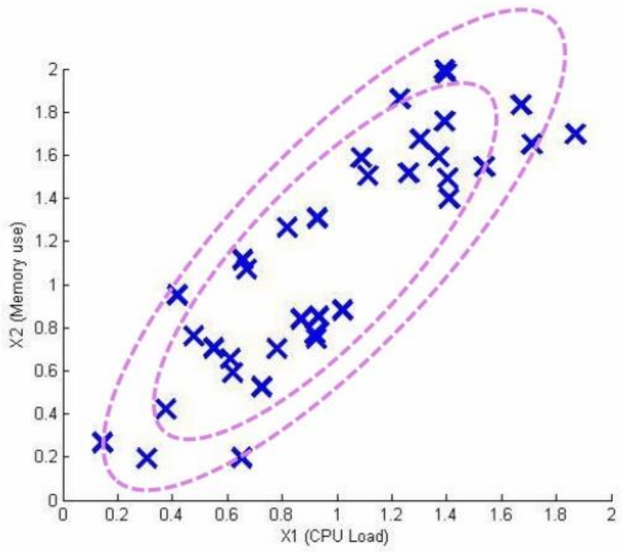


Estimation of parameters:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)\left(x^{(i)}-\mu\right)^T$$

Algorithm

Estimation of model parameters p(x):



$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

Calculate the value of p(x) for the new data x:

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

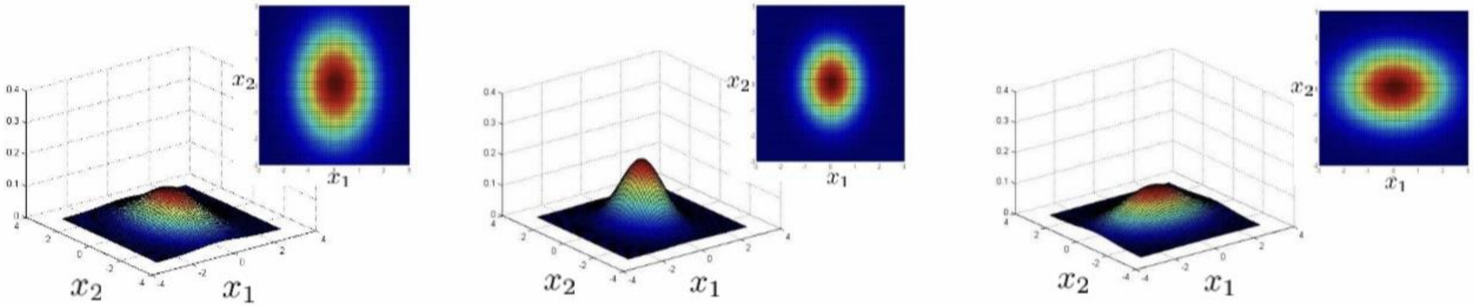
produce the output "yes" if p(x)< e

Relationship with the primary model

Basic model:

$$p(\boldsymbol{x}) = p(x_1; \mu_1, \sigma_1^2)p(x_2; \mu_2, \sigma_2^2)p(x_3; \mu_3, \sigma_3^2) \cdots p(x_n; \mu_n, \sigma_n^2)$$

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



Relationship with multivariate Gaussian distribution:

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

Basic or multivariate model

- Basic model:
  - Creating features is done manually.
  - Computational costs are relatively low.
  - If the number of training samples is small, it still works correctly.
- Multivariate Gaussian distribution:
  - It automatically learns the correlation between features.
  - High computational costs (inverse calculation of the covariance matrix)
  - The number of training samples should be more than the number of features. (invertibility of the sigma matrix)