Anomaly detection

Problem motivation

Machine Learning

Anomaly detection example

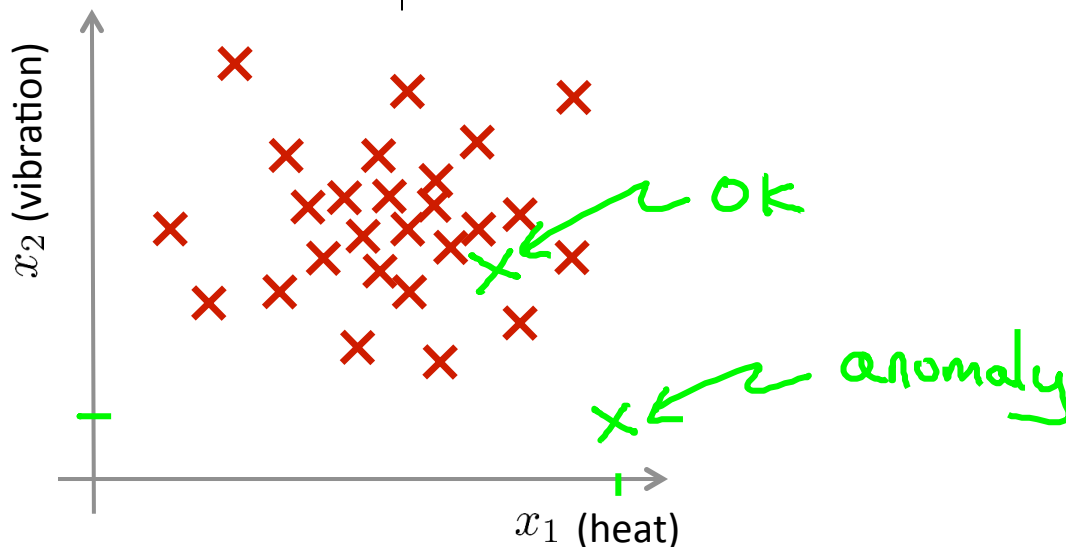Aircraft engine features:

→ x_1 = heat generated

→ x_2 = vibration intensity

...

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
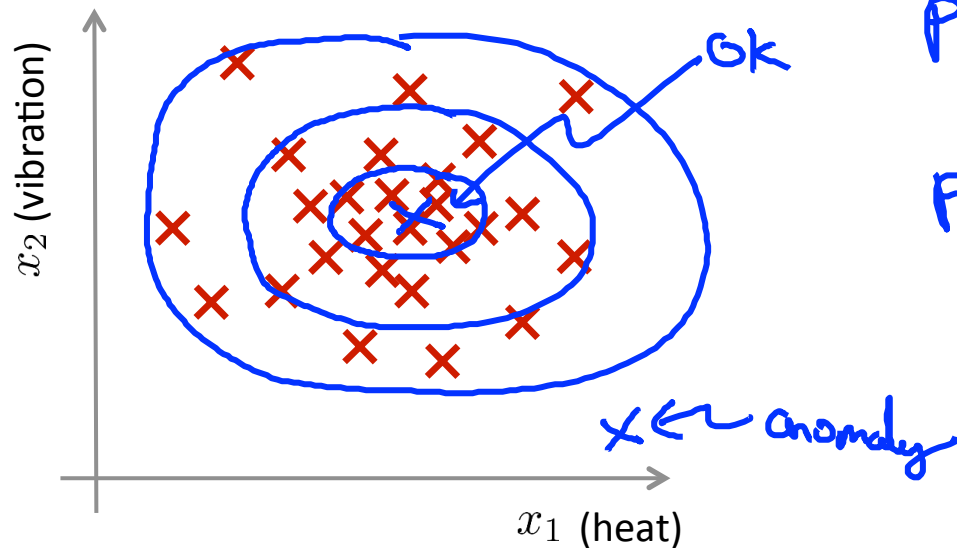
New engine: x_{test}

Density estimation

→ Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

→ Is x_{test} anomalous?

Model $p(x)$.

$p(x_{test}) < \varepsilon \rightarrow$ flag anomaly

$p(x_{test}) \geq \varepsilon \rightarrow$ OK

Anomaly detection example

→ Fraud detection:
  → $x^{(i)}$ = features of user i's activities
  → Model $p(x)$ from data.
  → Identify unusual users by checking which have $p(x) < \varepsilon$

→ Manufacturing

→ Monitoring computers in a data center.
  → $x^{(i)}$ = features of machine i
  x_1 = memory use, x_2 = number of disk accesses/sec,
  x_3 = CPU load, x_4 = CPU load/network traffic.
  ...

x_1
x_2
x_3
x_4

$p(x)$

$p(x) < \varepsilon$

Anomaly detection

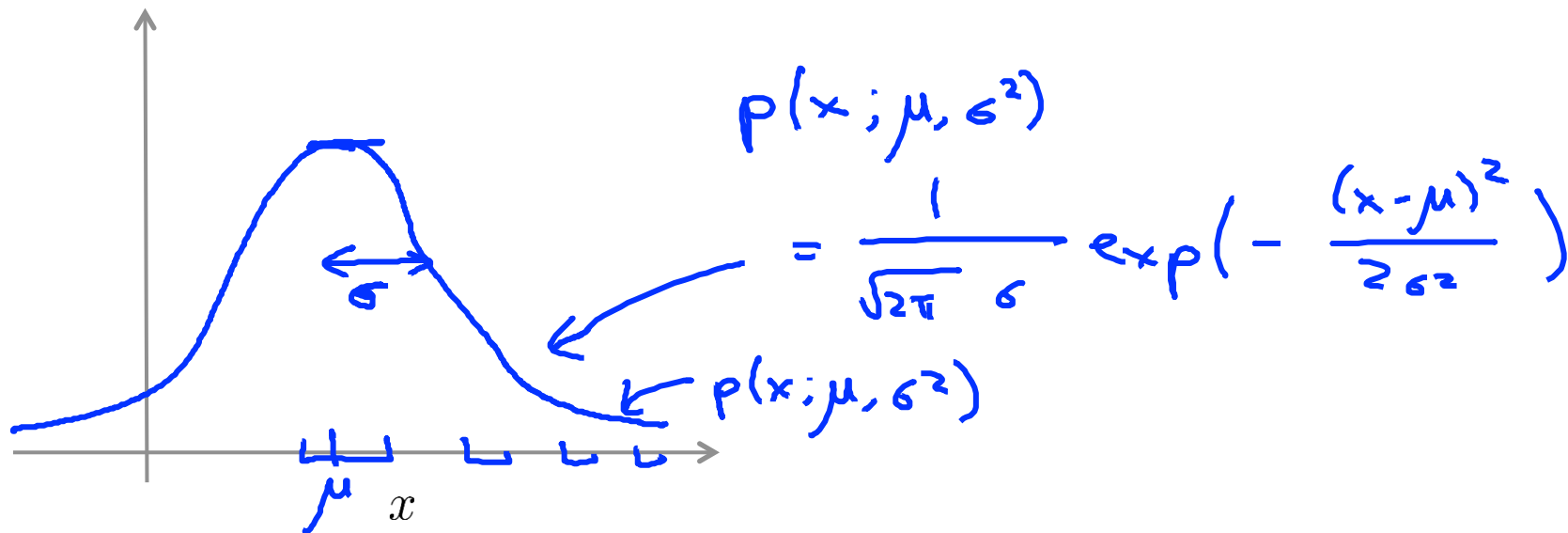## Gaussian distribution

Machine Learning

Gaussian (Normal) distribution

Say $x \in \mathbb{R}$. If x is a distributed Gaussian with mean μ, variance σ^2.
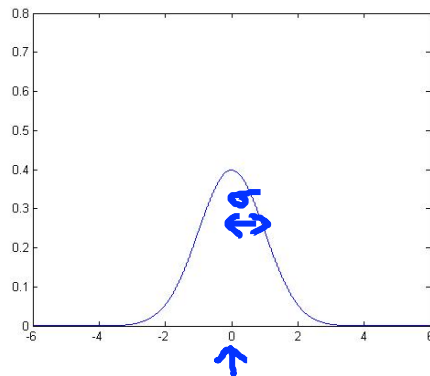
$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$\uparrow$ "distributed as"

σ standard deviation

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

$\leftarrow p(x; \mu, \sigma^2)$

Gaussian distribution example

$\mu = 0, \sigma = 1$

$\mu = 0, \sigma = \underline{0.5}$

$\sigma^2 = 0.25$

$\mu = 0, \sigma = 2$

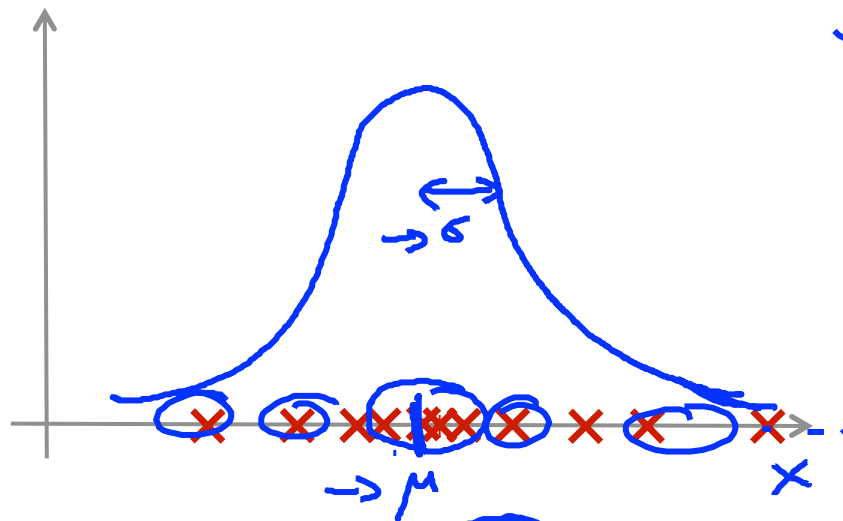$\mu = 3, \sigma = 0.5$

3

Andrew Ng
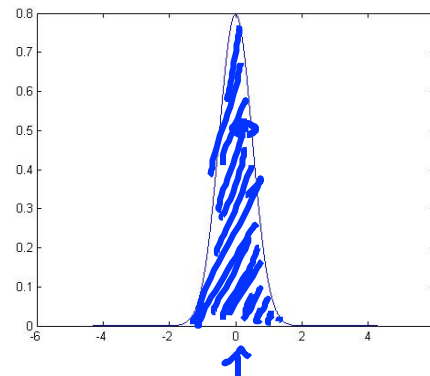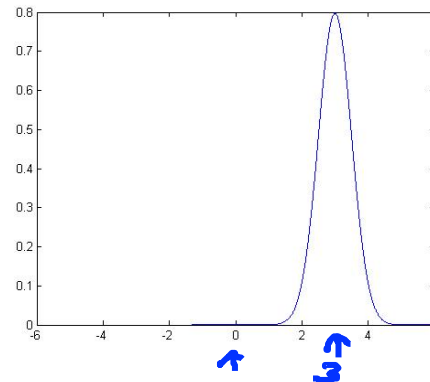
Parameter estimation

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

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

→ μ

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

→ $\sigma^2 = \boxed{\frac{1}{m}} \sum_{i=1}^{m} (x^{(i)} - \underline{\mu})^2$

→ $m-1$ $\frac{1}{m-1} \leftarrow$

Anomaly
detection

Algorithm

Machine Learning

→ **Density estimation**

→ Training set: $\{x^{(1)}, \ldots, x^{(m)}\}$

Each example is $\underline{x \in \mathbb{R}^n}$

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$x_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$$

→ $p(x)$

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

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

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n$$

$$\prod_{i=1}^{n} i = 1 \times 2 \times 3 \times \cdots \times n$$

Anomaly detection algorithm

1. Choose features x_i that you think might be indicative of anomalous examples.

   $\{x^{(1)}, \ldots, x^{(m)}\}$

2. Fit parameters $\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2$

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

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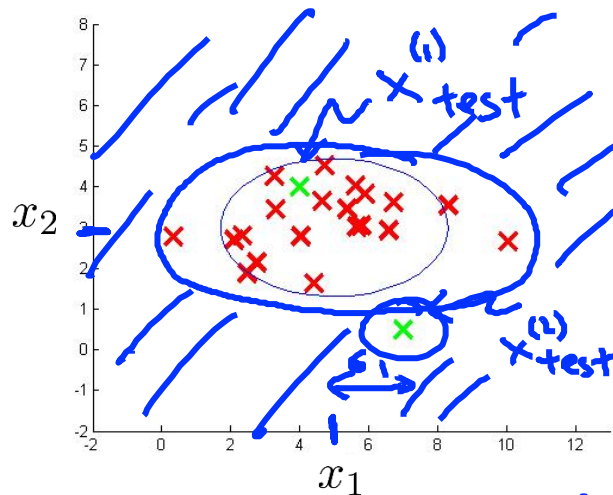
$p(x_j; \mu_j, \sigma_j^2)$

$\mu_1, \mu_2, \ldots, \mu_n$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

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

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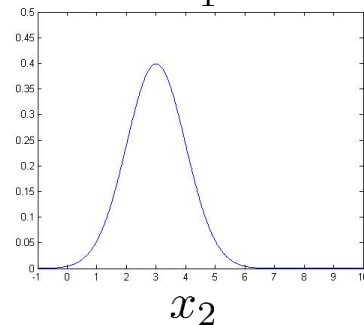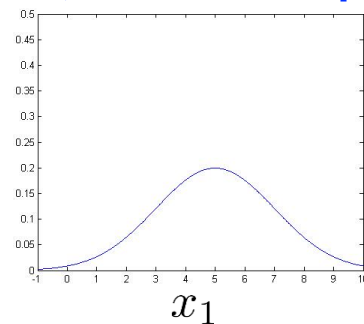
Anomaly if $p(x) < \varepsilon$

Anomaly detection example

$\rightarrow p(x) = p(x_1; \mu_1, \sigma_1^2)$
$\times p(x_2; \mu_2, \sigma_2^2)$

$$\mu_1 = 5, \sigma_1 = 2$$
$$\mu_2 = 3, \sigma_2 = 1$$

$\sigma_1^2, \sigma_2^2 = 4$

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

$p(x_2; \mu_2, \sigma_2^2)$
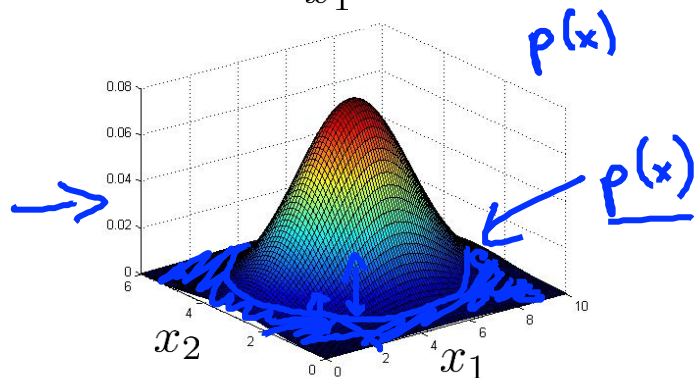
$\varepsilon = 0.02$

$$p(x_{test}^{(1)}) = 0.0426 \geq \varepsilon$$
$$p(x_{test}^{(2)}) = 0.0021 < \varepsilon$$

$p(x)$

Anomaly detection

Developing and evaluating an anomaly detection system

Machine Learning

The importance of real-number evaluation

When developing a learning algorithm (choosing features, etc.), making decisions is much easier if we have a way of evaluating our learning algorithm.

→ Assume we have some labeled data, of anomalous and non-anomalous examples. ($y = 0$ if normal, $y = 1$ if anomalous).

→ Training set: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ (assume normal examples/not anomalous)

→ Cross validation set: $(x^{(1)}_{cv}, y^{(1)}_{cv}), \ldots, (x^{(m_{cv})}_{cv}, y^{(m_{cv})}_{cv})$

→ Test set: $(x^{(1)}_{test}, y^{(1)}_{test}), \ldots, (x^{(m_{test})}_{test}, y^{(m_{test})}_{test})$

$y = 1$

Aircraft engines motivating example

→ 10000 good (normal) engines

→ 20 flawed engines (anomalous) $\quad$ 2-50 $\qquad\qquad$ $y = 1$

$\mu_1, \sigma_1^2, \ldots, \mu_n, \sigma_n^2.$

→ Training set: 6000 good engines $(y = 0)$ $\quad p(x) = p(x_1; \mu_1, \sigma_1^2) \cdots p(x_n; \mu_n, \sigma_n^2)$

CV: 2000 good engines $(y = 0)$, 10 anomalous $(y = 1)$

Test: 2000 good engines $(y = 0)$, 10 anomalous $(y = 1)$

Alternative:

Training set: 6000 good engines

→ CV: 4000 good engines $(y = 0)$, 10 anomalous $(y = 1)$

→ Test: 4000 good engines $(y = 0)$, 10 anomalous $(y = 1)$

Algorithm evaluation

→ Fit model $p(x)$ on training set $\{x^{(1)}, \ldots, x^{(m)}\}$

→ On a cross validation/test example x , predict

$$ y = \begin{cases} 1 & \text{if } p(x) < \varepsilon \text{ (anomaly)} \\ 0 & \text{if } p(x) \geq \varepsilon \text{ (normal)} \end{cases} $$

$\left(x^{(i)}_{\text{test}} , y^{(i)}_{\text{test}} \right)$

$y = 0$

Possible evaluation metrics:

→   - True positive, false positive, false negative, true negative

→   - Precision/Recall

→   - F$_1$-score ←

CV

Test set

Can also use cross validation set to choose parameter ε ←

Anomaly detection

Anomaly detection vs. supervised learning

Machine Learning

Anomaly detection   vs.   Supervised learning

→ Very small number of positive examples ($y = 1$). (0-20 is common).

skewed data

→ Large number of negative ($y = 0$) examples. $\boxed{p^{(x)}}$ ← ←

→ Many different "types" of anomalies. Hard for any algorithm to learn from positive examples what the anomalies look like;

→ future anomalies may look nothing like any of the anomalous examples we've seen so far.

Large number of positive and ← negative examples.

Enough positive examples for ← algorithm to get a sense of what positive examples are like, future ← positive examples likely to be similar to ones in training set.

Spam ←

Anomaly detection   **vs.**   **Supervised learning**

Anomaly detection	Supervised learning
• Fraud detection $\quad y = 1$	• Email spam classification
• Manufacturing (e.g. aircraft engines)	• Weather prediction (sunny/rainy/etc).
• Monitoring machines in a data center	• Cancer classification
⋮	⋮

Anomaly detection

Choosing what features to use

Machine Learning

Non-gaussian features
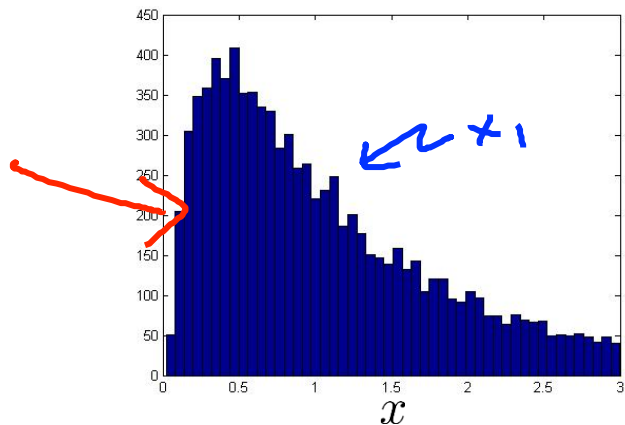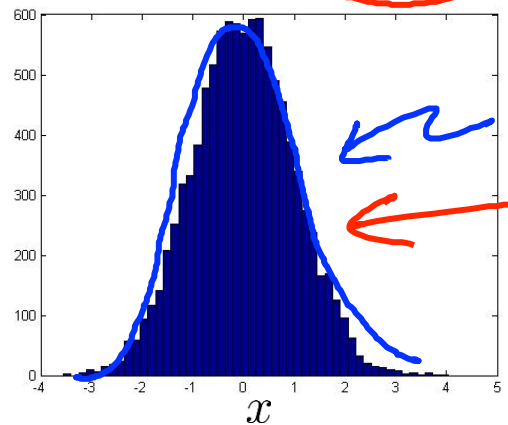
→ Error analysis for anomaly detection

Want $p(x)$ large for normal examples x.

$p(x)$ small for anomalous examples x.

Most common problem:

$p(x)$ is comparable (say, both large) for normal and anomalous examples
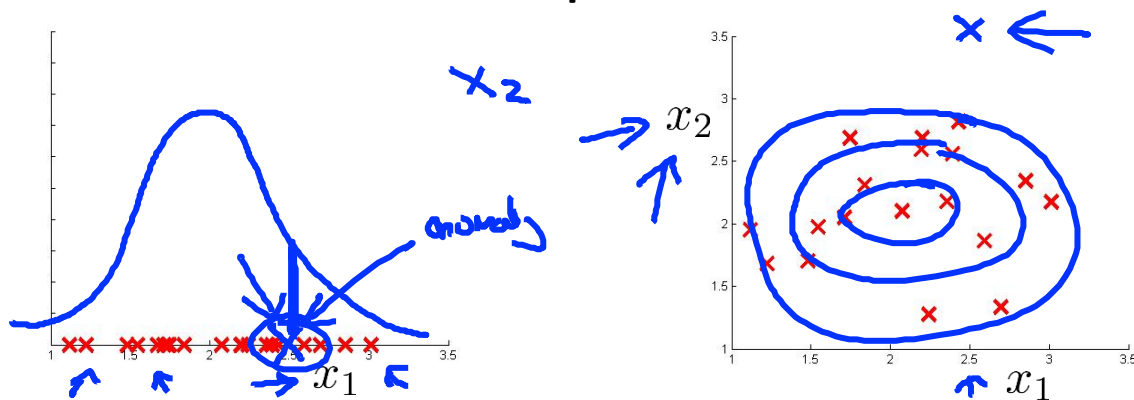
Monitoring computers in a data center

Choose features that might take on unusually large or small values in the event of an anomaly.

x_1 = memory use of computer

x_2 = number of disk accesses/sec

x_3 = CPU load

x_4 = network traffic

$$x_5 = \frac{CPU\ load}{network\ traffic}$$

$$x_6 = \frac{(CPU\ load)^2}{network\ traffic}$$

Anomaly detection

Multivariate Gaussian distribution

Machine Learning

Motivating example: Monitoring machines in a data center

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

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

x_1 (CPU Load)

x_2 (Memory Use)

x_2 (Memory Use)

x_1 (CPU Load)
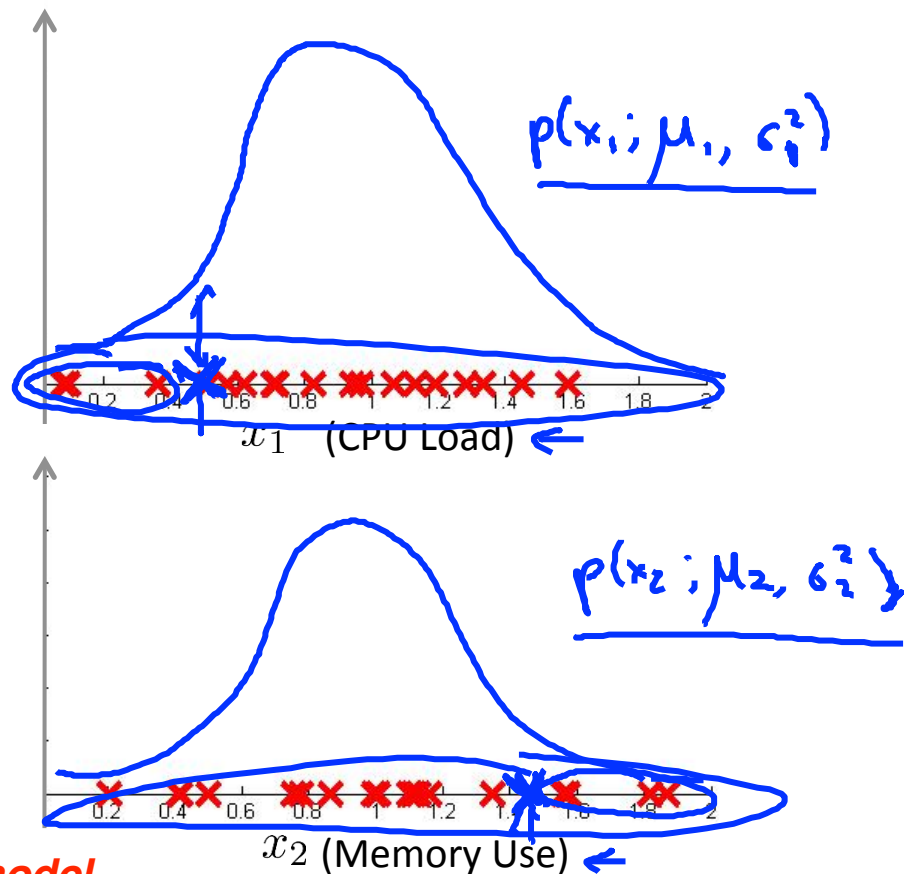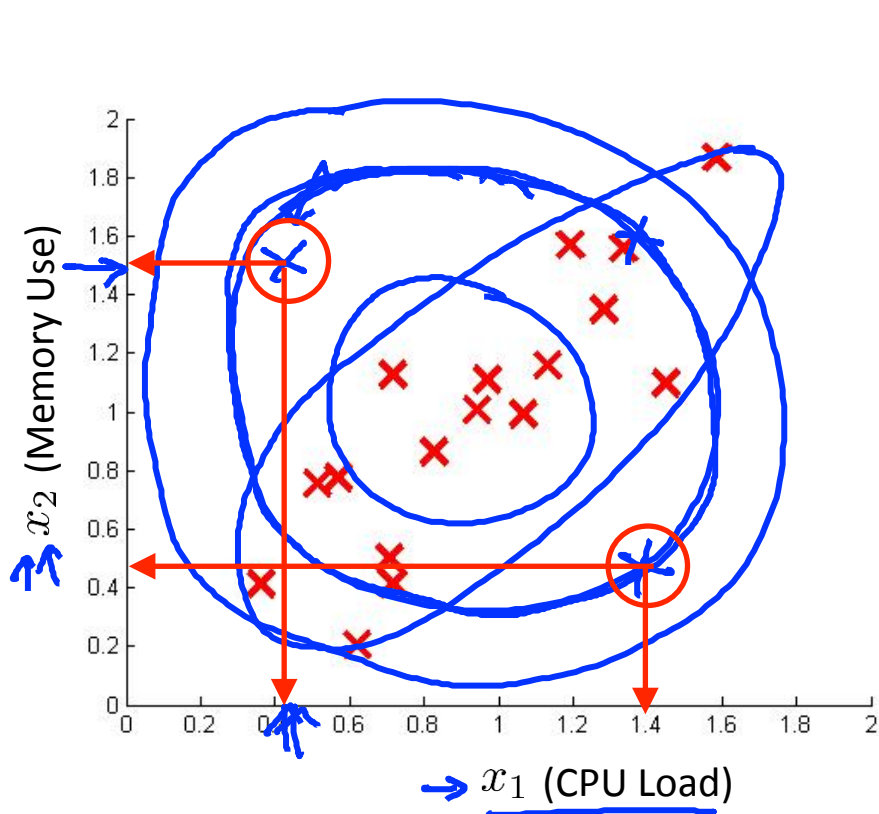
Normal, and can not be detected by original model

Andrew Ng

Multivariate Gaussian (Normal) distribution

$\rightarrow$ $x \in \mathbb{R}^n$. Don't model $p(x_1), p(x_2), \ldots$, etc. separately.
Model $p(x)$ all in one go.
Parameters: $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ (covariance matrix)

$$p(x; \mu, \Sigma) =$$

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

$|\Sigma|$ = determinant of Σ det (Sigma)
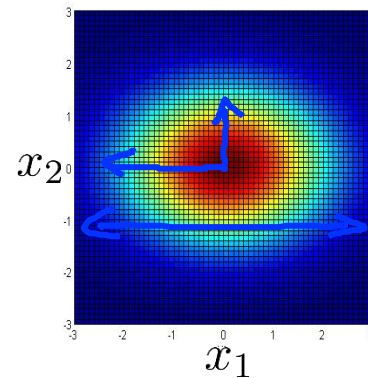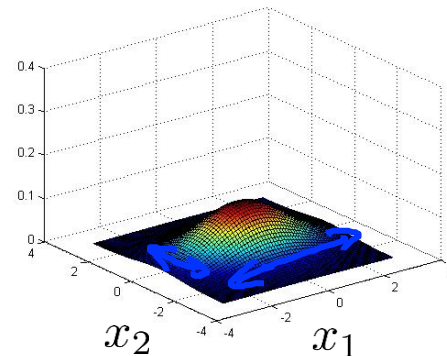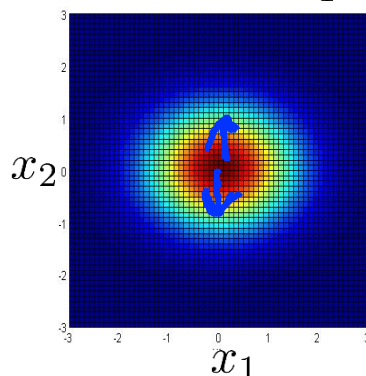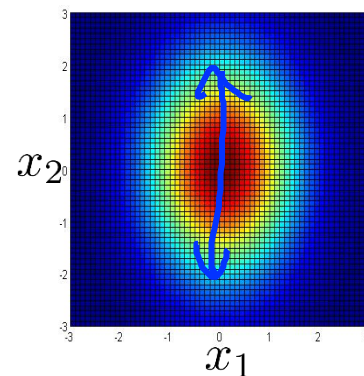
Multivariate Gaussian (Normal) examples

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

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

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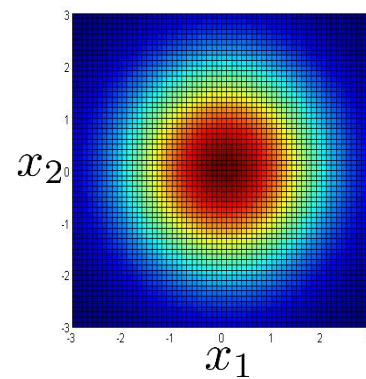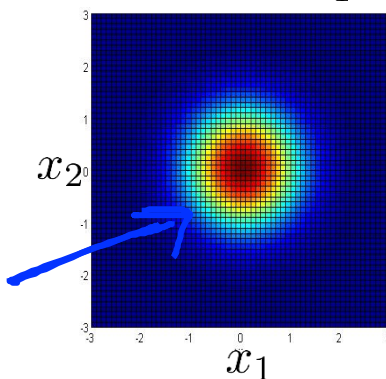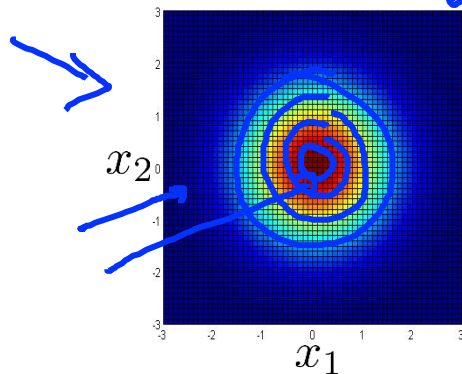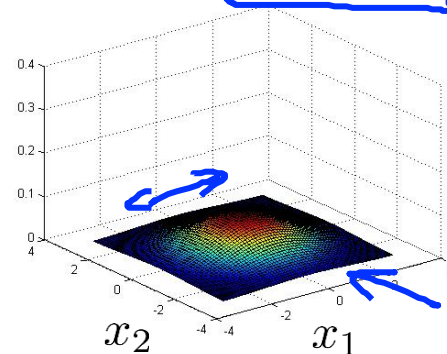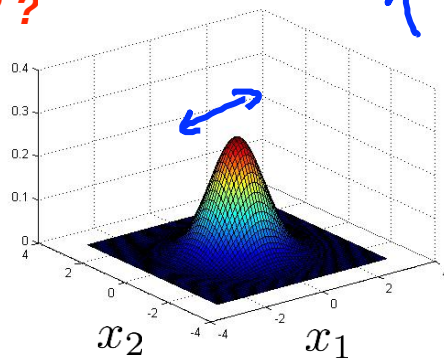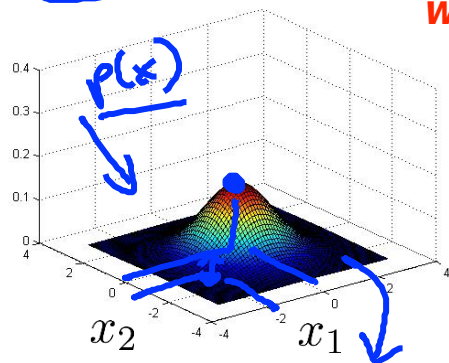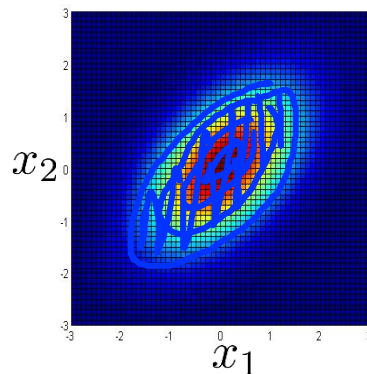
why ? ? ?

Multivariate Gaussian (Normal) examples

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

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$
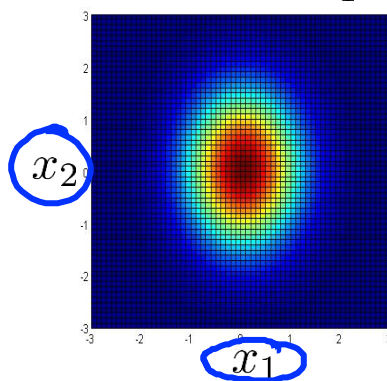
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$
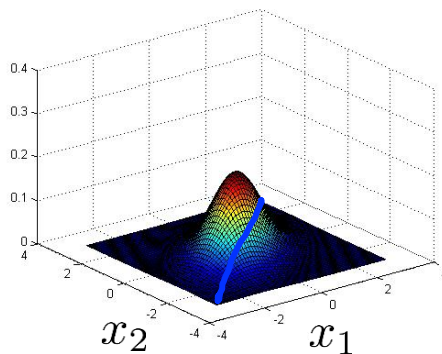

Andrew Ng

Multivariate Gaussian (Normal) examples
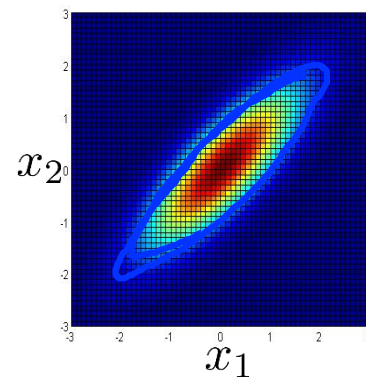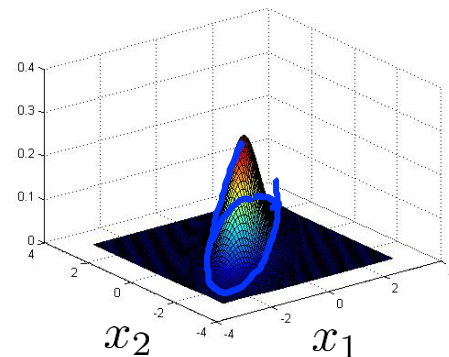
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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

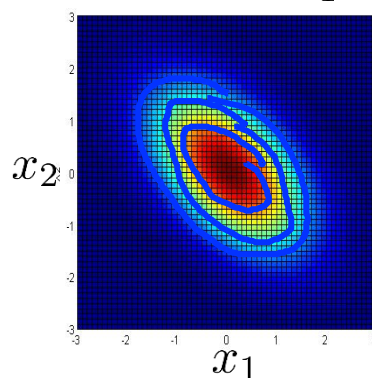$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

Andrew Ng

Multivariate Gaussian (Normal) examples

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Multivariate Gaussian (Normal) examples

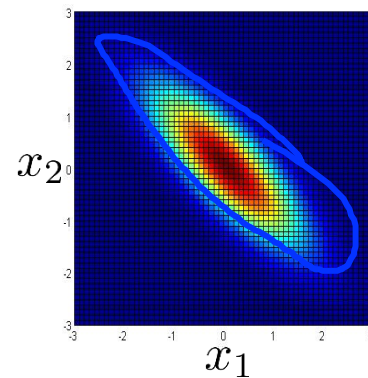$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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

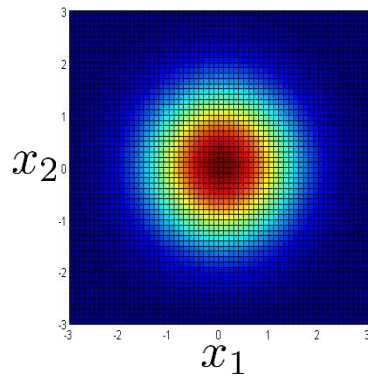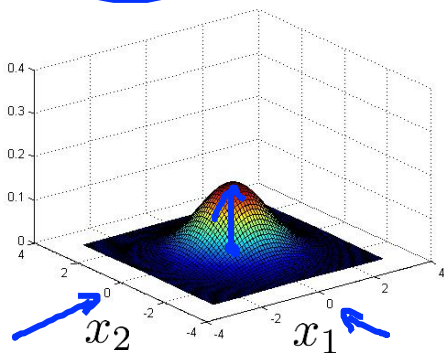$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$
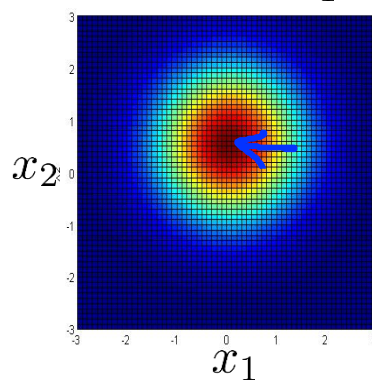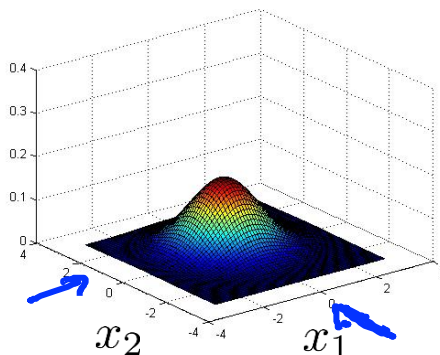

Andrew Ng

Multivariate Gaussian (Normal) examples

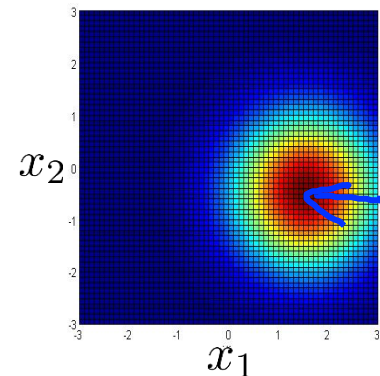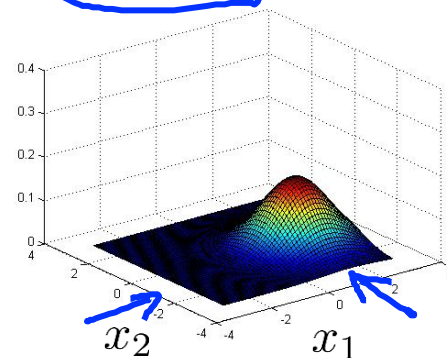$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Andrew Ng

Anomaly detection

Anomaly detection using the multivariate Gaussian distribution

Machine Learning
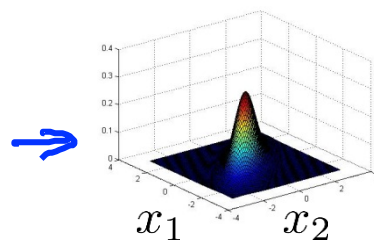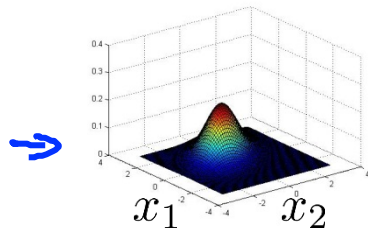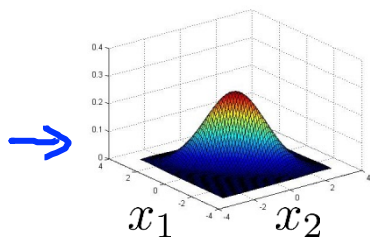
Multivariate Gaussian (Normal) distribution

Parameters μ, Σ

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

co-variance matrix

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

Parameter fitting:

Given training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$x \in \mathbb{R}^n$

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

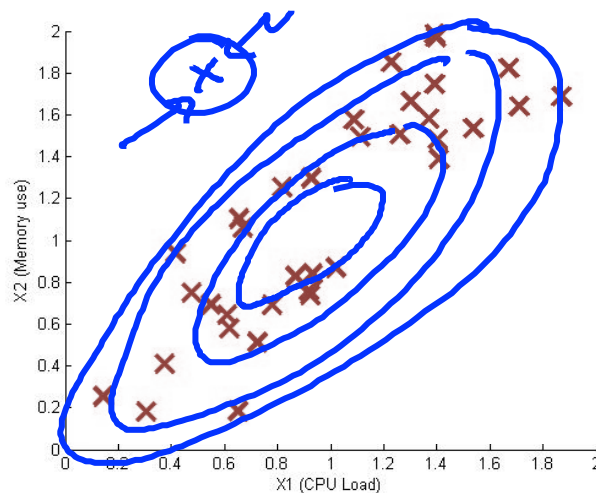# Anomaly detection with the multivariate Gaussian

1. Fit model $p(x)$ by setting

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

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

2. Given a new example x, compute

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

Flag an anomaly if $p(x) < \varepsilon$

Relationship to original model

Original model: $p(x) = p(x_1; \mu_1, \sigma_1^2) \times p(x_2; \mu_2, \sigma_2^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$

Corresponds to multivariate Gaussian

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

where

→ Original model   **vs.**   **→ Multivariate Gaussian**

$$p(x_1; \mu_1, \sigma_1^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$

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

Manually create features to capture anomalies where x_1, x_2 take unusual combinations of values.

$$\rightarrow X_3 = \frac{X_1}{X_2} = \frac{CPU\ load}{memory}$$

→ Automatically captures correlations between features

$$\Sigma \in \mathbb{R}^{n \times n} \qquad \Sigma^{-1}$$

→ Computationally cheaper (alternatively, scales better to large
$$n = 10,000, \qquad n = 100,000$$
)

OK even if m (training set size) is small

Computationally more expensive

$$\rightarrow \Sigma \qquad \sim \frac{n^2}{2} \swarrow$$

$$\rightarrow X_1 = \cancel{X_2}$$
$$\cancel{X_3} = X_4 + X_5$$

Must have $\boxed{m > n}$ or else Σ is non-invertible. $\rightarrow \boxed{m \geq 10n}$