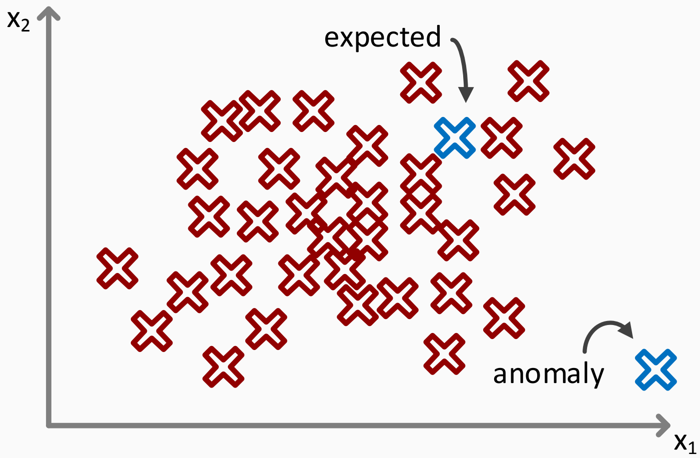
anomaly detection basics

density estimation

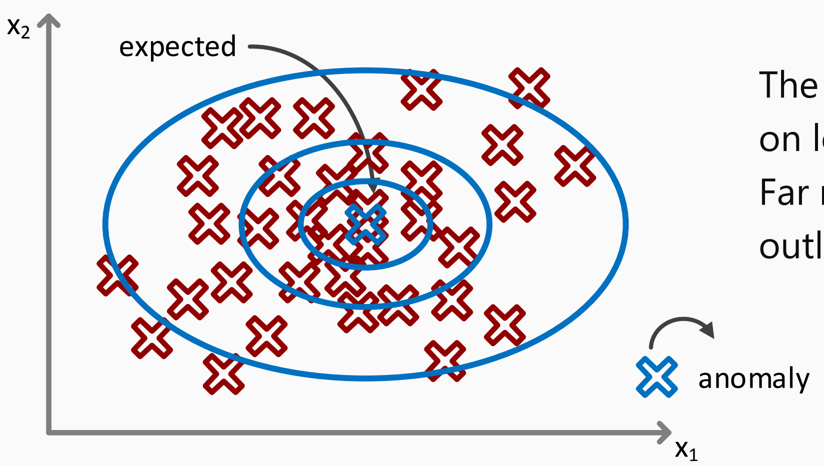
problem motivation



when given a dataset of features $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, the points plotted (red) represent the data correlations. When a new dataset is added to the model (x_{test}), points that fall within the majority or expected, but any far reaching points are considered to be an anomaly

formally, a model $p(x)$will determine anomaly detection: $p(x_{test}) < \varepsilon$ anomaly

$(x_{test}) \geq \varepsilon$ expected

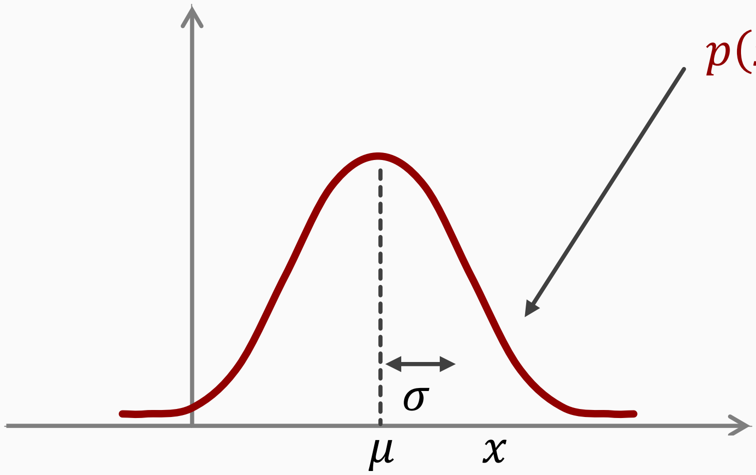


The model p(x) will set thresholds based on logic acceptable to the given data. Far reaching points will be labeled as an outlier or anomaly.

gaussian (normal) distribution

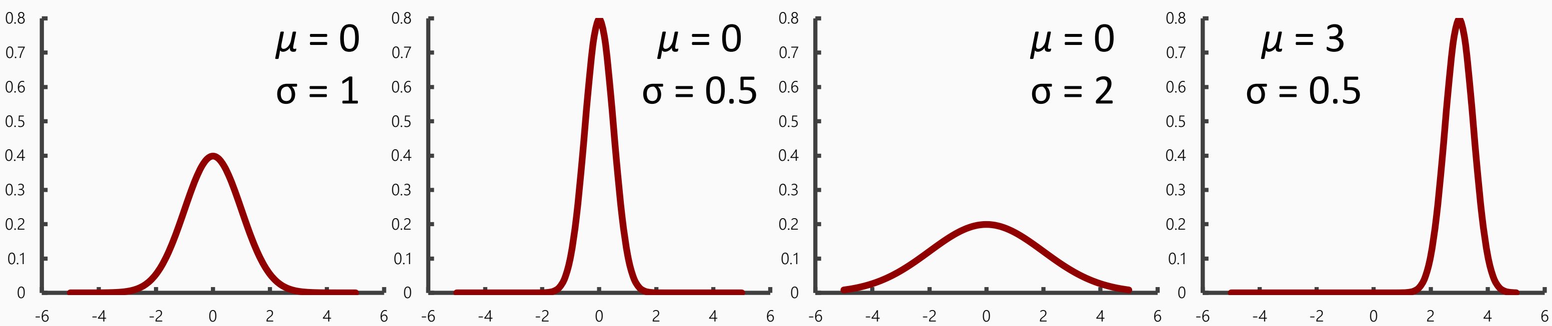
$x \in \mathbb{R}$. if x is a distributed gaussian with mean μ, variance σ^2, denoted as $\boldsymbol{x \sim \mathcal{N}(\mu, \sigma^2)}$

$\boldsymbol{x}$ represents a row values random variable (x is a row number). $\boldsymbol{\sim}$ tilde represents the variable "distributed as". The script $\boldsymbol{\mathcal{N}}$ represents the "normal" gaussian distribution and is parameterized by two parameters $\boldsymbol{\mu}$ mu (mean) and $\boldsymbol{\sigma^2}$ sigma squared (variance)



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

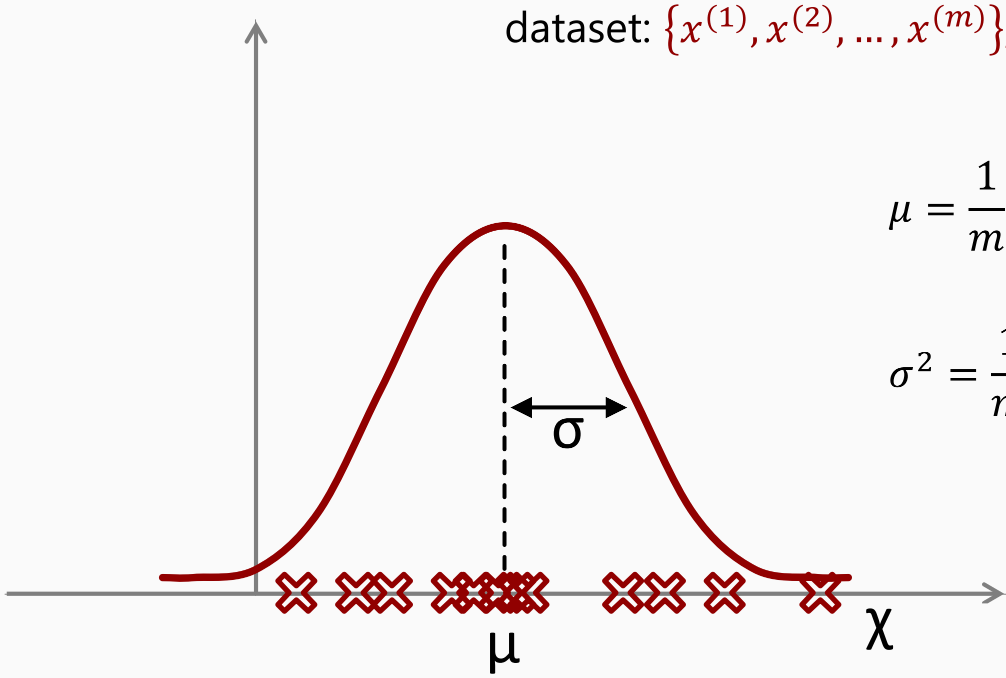
gaussian distribution examples



parameter estimation

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

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



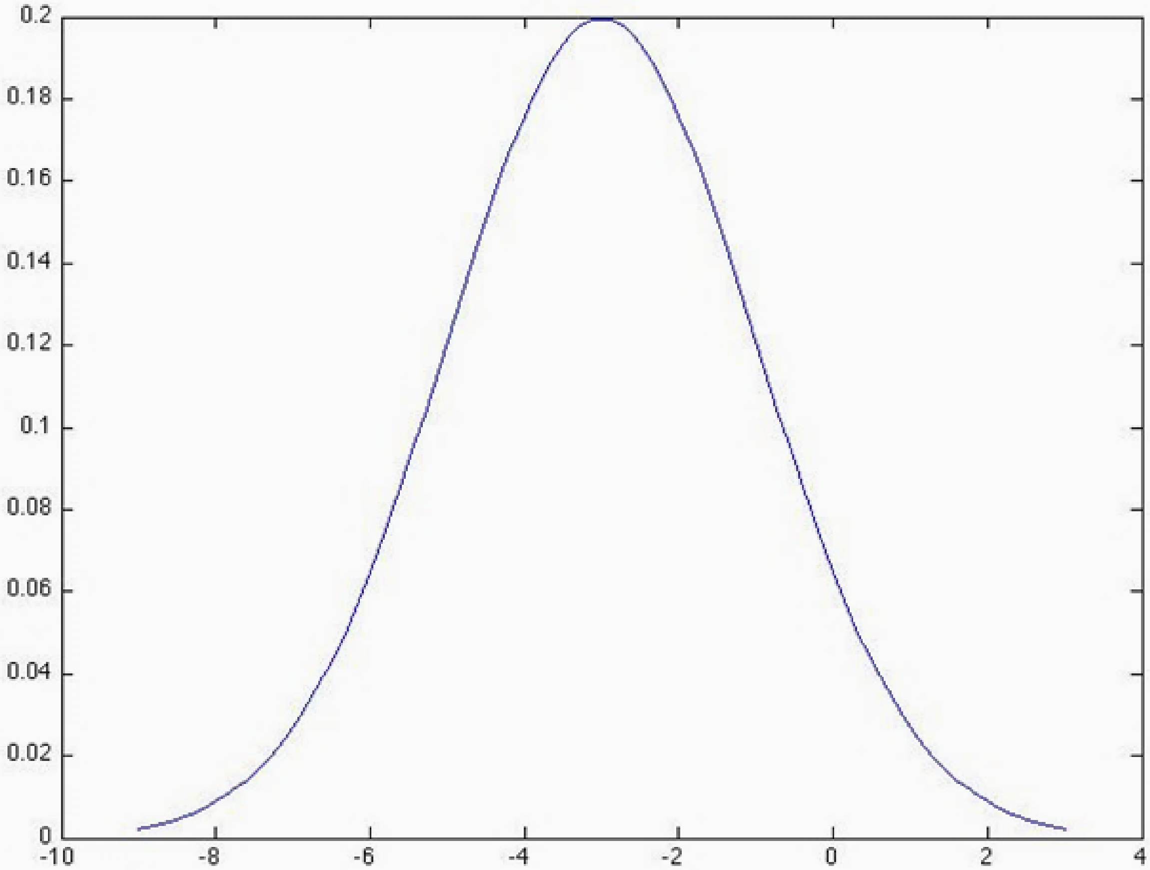
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

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

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?



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

Correct Response

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

anomaly detection algorithm

density estimation

dataset: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$,

each example is $x^{(i)} \in \mathbb{R}^n$

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

$$
\begin{aligned}
p(x) \\
&= p(x_1; \mu_1, \sigma_1^2)p(x_2; \mu_2, \sigma_2^2)p(x_3; \mu_3, \sigma_3^2) \dots p(x_n; \mu_n, \sigma_n^2) \\
&= \prod_{j=1}^{n} p\left(x_j; \mu_j, \sigma_j^2\right)
\end{aligned}
$$

to describe the product operator in compassion to the summation operator:

$$\sum_{i=1}^{n} 1 + 2 + 3 + \cdots + n \qquad \text{versus} \qquad \prod_{i=1}^{n} 1 \times 2 \times 3 \times \dots \times n$$

estimating the above distribution $p(x)$ is referred to as the density estimation problem

given training set $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, each would be estimated as (note $\mu_j \in \mathbb{R}, \sigma_j^2 \in \mathbb{R}$):

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

steps in the anomaly detection algorithm

choose features x_i that might be indicative of anomalous examples: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

fit parameters $\mu_1, \dots, \mu_n, \sigma_1^2, \dots, \sigma_n^2$

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

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

vectorized version:

$p\left(x_j; \mu_j, \sigma_j^2\right) \qquad u_1, \mu_2, \dots \mu_n$

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

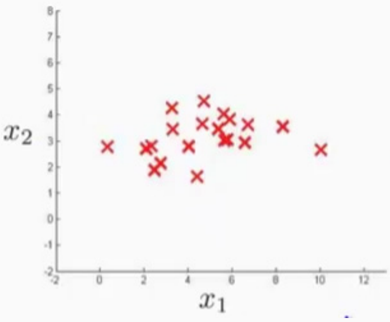
given new example x, compute $p(x)$:

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

anomaly if $p(x) < \varepsilon$

anomaly detection example

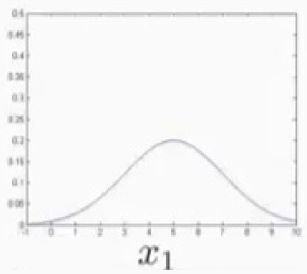
The dataset below has an average measured feature x_1 of ~5 with a standard deviation of ~2 and the feature x_2 has an average value of ~3 with a standard deviation ~1



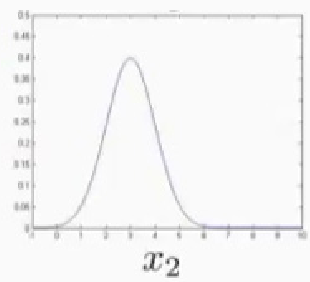
$$\mu_1 = 5, \sigma_1 = 2$$
$$\mu_2 = 3, \sigma_2 = 1$$

Note the use of σ (standard deviation) as opposed to σ^2 (variance) used previously

The features above plotted as in terms of $p(x_i; \mu_i, \sigma_i^2)$ distribution would appear as:

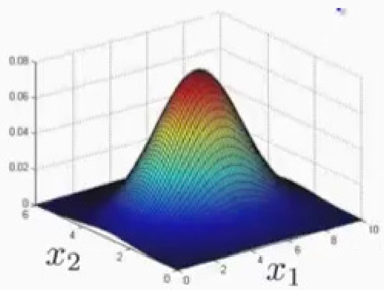


$p(x_1; \mu_1, \sigma_1^2)$



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

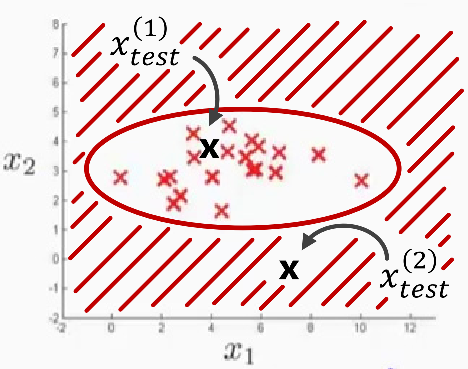
p(x) for both can be taken and plotted in a surface plot as seen below. The height of the surface given a particular values is p(x):



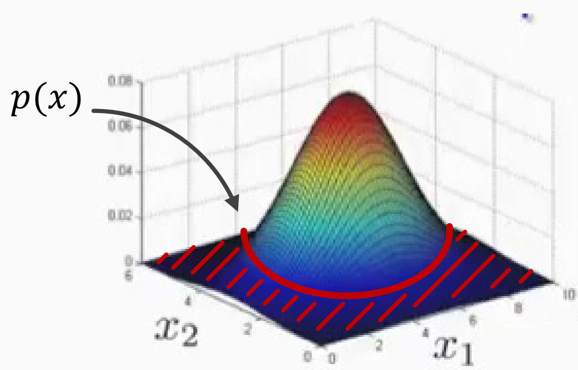
the height of the plot is literally equal to:

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

if additional data points are added to the dataset:



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



to determine if either is an anomaly: a value is set for epsilon ε, (in this case, ε = 0.02):

$p\left(x_{test}^{(1)}\right) = 0.0426 \geq \varepsilon$ (non-anomaly) and $p\left(x_{test}^{(2)}\right) = 0.0021 < \varepsilon$ (anomaly)