Gaussian Distribution

Problem Motivation Summary

Just like in other learning problems, we are given a dataset $x^{(1)}, x^{(2)}, \dots, x^{(m)}$.

We are then given a new example, x_{test}, and we want to know whether this new example is abnormal/anomalous.

We define a "model" $p(x)$ that tells us the probability the example is not anomalous. We also use a threshold ϵ (epsilon) as a dividing line so we can say which examples are anomalous and which are not.

A very common application of anomaly detection is detecting fraud:

- $x^{(i)}$ = features of user i's activities
- Model $p(x)$ from the data.
- Identify unusual users by checking which have $p(x) < \epsilon$.

If our anomaly detector is flagging **too many** anomalous examples, then we need to **decrease** our threshold ϵ

Gaussian (Normal) Distribution

We're going to talk about the Gaussian distribution which is also called **the normal distribution**.

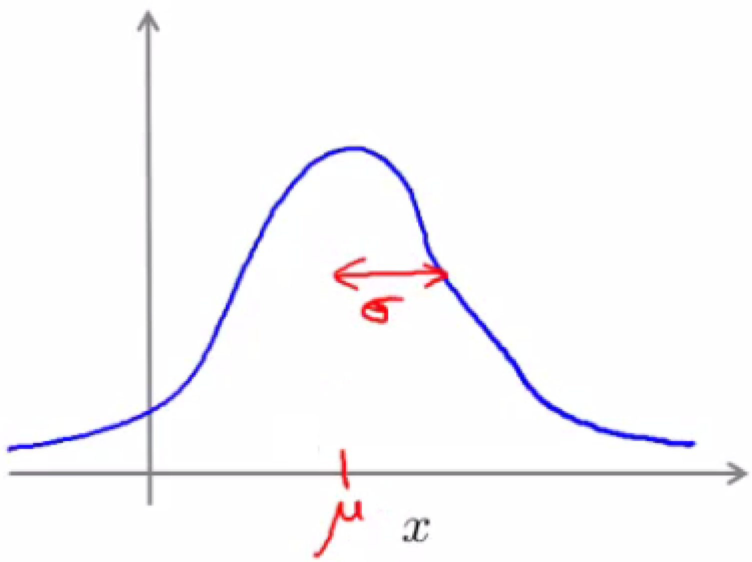
Say x (data set) is made up of real number ($x \in \mathbb{R}$). If x is a distributed Gaussian with:

- mean μ
- variance σ^2
  - σ is also called **the standard deviation** - specifies the width of the Gaussian probability.

The data has a **Gaussian distribution**, we can write this as $x \sim N(\mu, \sigma^2)$

- ~ means = "is distributed as"
- N (should really be "script" N) $\rightarrow$ means normal distribution
- The Gaussian Distribution is parameterized by a mean μ and a variance σ^2.
- μ, σ^2 represent the mean and variance, respectively

If we plot the Gaussian distribution or **Gaussian probability density**. It'll look like the bell shaped curve:



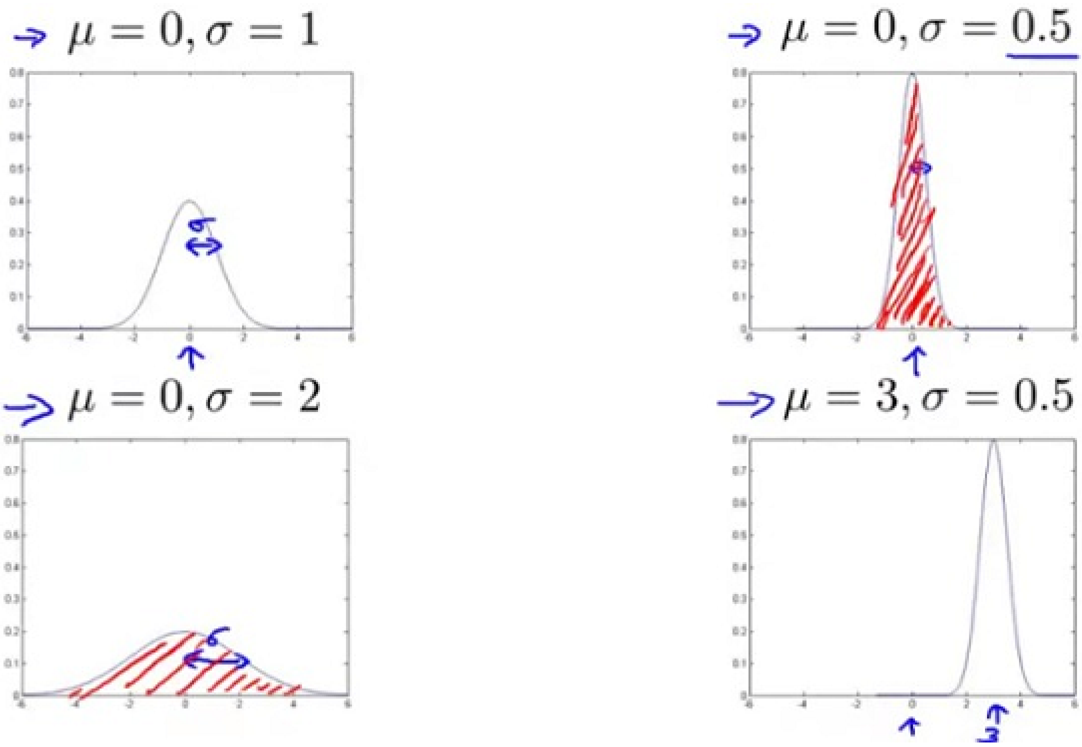
Mu, or μ, describes the center of the curve, called the mean. The width of the curve is described by sigma, or σ, called the standard deviation.

The full function is as follows:

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

Some examples of Gaussians below:

- Area is always the same (must = 1)
- But width changes as standard deviation changes

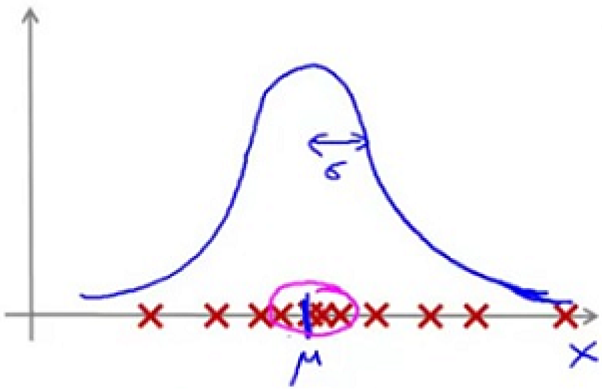


Parameter estimation problem

Given a dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}$, give each example is a real number - we can plot the data on the x axis as shown below:

The problem of parameter estimation is given our data set, we want to try to figure out or we want to estimate what are the values of μ and σ^2.

- Given the dataset can we estimate the distribution?



Seems like a reasonable fit - data seems like a higher probability of being in the central region, lower probability of being further away

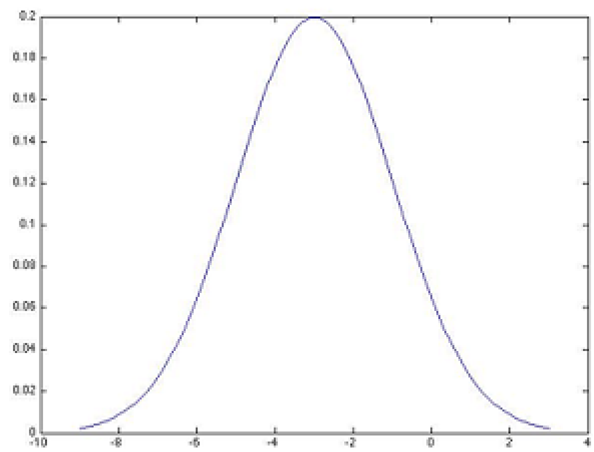
- We can estimate the parameter μ (the mean) from a given dataset by simply taking the average of all the examples:
  - $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$
- We can estimate the other parameter, σ^2 (the variance), with our familiar squared error formula:
  - $\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu)^2$

These parameters are the maximum likelihood estimation values for μ and σ^2

- We can also do 1/(m) or 1/(m - 1) doesn't make too much difference
  - Slightly different mathematical problems, but in practice it makes little difference

Video 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?

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

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

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