Chris Piech
CS109

Lecture #10
Oct 15th, 2018

Gaussian

The single most important random variable type is the Normal (aka Gaussian) random variable, parametrized by a mean (μ) and variance (σ^2). If X is a normal variable we write $X \sim N(\mu, \sigma^2)$. The normal is important for many reasons: it is generated from the summation of independent random variables and as a result it occurs often in nature. Many things in the world are not distributed normally but data scientists and computer scientists still model them as Normal distributions anyways. Why? Because it is the most entropic (conservative) distribution that we can apply to data with a measured mean and variance.

0.1 Properties

The Probability Density Function (PDF) for a Normal is:

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

By definition a Normal has $E[X] = \mu$ and $Var(X) = \sigma^2$.

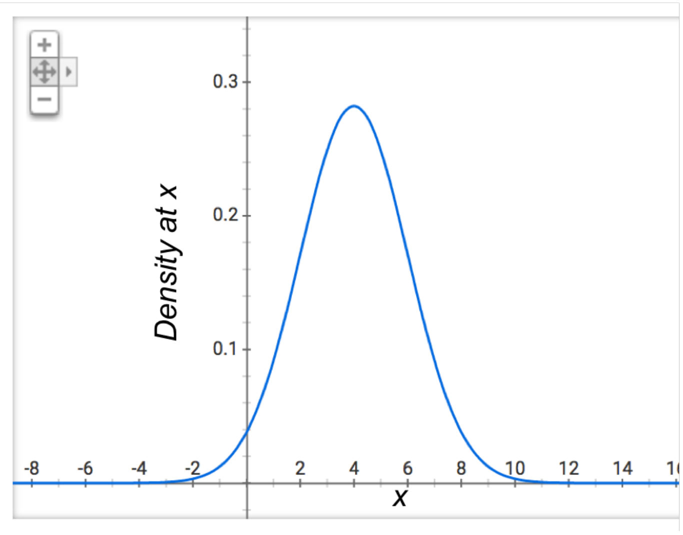
If X is a Normal such that $X \sim N(\mu, \sigma^2)$ and Y is a linear transform of X such that $Y = aX + b$ then Y is also a Normal where $Y \sim N(a\mu + b, a^2\sigma^2)$.

There is no closed form for the integral of the Normal PDF, however since a linear transform of a Normal produces another Normal we can always map our distribution to the "Standard Normal" (mean 0 and variance 1) which has a precomputed Cumulative Distribution Function (CDF). The CDF of an arbitrary normal is:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Where Φ is a precomputed function that represents that CDF of the Standard Normal.

A concrete example of random variable that is treated as Gaussian is the number of roses on a rosebush. For a species of roses grown in Lake Naivasha, Kenya it has been observed that the number of roses on a mature bush is X and it is distributed as $X \sim N(\mu = 4, \sigma^2 = 2)$. Here is a graphical representation of the probability density function for number of roses:



What is the probability that a bush is a “super-bush” meaning it has more than 6 roses?

$$\begin{aligned} P(X > 6) &= 1 - F_X(6) \\ &= 1 - \Phi\Big(\frac{6-\mu}{\sigma}\Big) && \text{Recall that: } F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \\ &= 1 - \Phi\Big(\frac{6-4}{\sqrt{2}}\Big) \\ &\approx 1 - \Phi(1.414) \\ &\approx 0.079 \end{aligned}$$

Projection to Standard Normal

For any Normal X we can define a random variable $Z \sim N(0,1)$ to be a linear transform

$$\begin{aligned} Z &= \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} \\ &\sim N(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}) \\ &\sim N(0,1) \end{aligned}$$

Using this transform we can express $F_X(x)$, the CDF of X, in terms of the known CDF of Z, $F_Z(x)$. Since the CDF of Z is so common it gets its own Greek symbol: $\Phi(x)$

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

The values of $\Phi(x)$ can be looked up in a table. We also have an online calculator.

Example 1

Let $X \sim N(3,16)$, what is $P(X > 0)$?

$$\begin{aligned} P(X > 0) &= P\left(\frac{X-3}{4} > \frac{0-3}{4}\right) = P\left(Z > -\frac{3}{4}\right) = 1 - P\left(Z \leq -\frac{3}{4}\right) \\ &= 1 - \Phi(-\frac{3}{4}) = 1 - (1 - \Phi(\frac{3}{4})) = \Phi(\frac{3}{4}) = 0.7734 \end{aligned}$$

What is $P(2 < X < 5)$?

$$\begin{aligned} P(2 < X < 5) &= P\left(\frac{2-3}{4} < \frac{X-3}{4} < \frac{5-3}{4}\right) = P\left(-\frac{1}{4} < Z < \frac{2}{4}\right) \\ &= \Phi(\frac{2}{4}) - \Phi(-\frac{1}{4}) = \Phi(\frac{1}{2}) - (1 - \Phi(\frac{1}{4})) = 0.2902 \end{aligned}$$

Example 2

You send voltage of 2 or -2 on a wire to denote 1 or 0. Let X = voltage sent and let R = voltage received. $R = X + Y$, where $Y \sim N(0,1)$ is noise. When decoding, if $R \geq 0.5$ we interpret the voltage as 1, else 0. What is $P(\text{error after decoding}|\text{original bit} = 1)$?

$$P(X + Y < 0.5) == P(2 + Y < 0.5) = P(Y < -1.5) = \Phi(-1.5) = 1 - \Phi(1.5) \approx 0.0668$$