EE 178/278A Probabilistic Systems Analysis

Spring 2014 Tse/Hussami Lecture 14

Continuous Probability Continued

We will finish the main part of the material in this lecture, and then move on to applications of probability in inference problems. We ended lecture 13 by introducing some continuous distribution. The normal distribution has two parameters, μ and σ and is defined as:

Definition 14.1 (Gaussian (Normal) distribution): For any μ and $\sigma > 0$, a continuous random variable X with pdf f given by

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

is called a *Gauusian* or *Normal* random variable with parameters μ and σ ($X \sim N(\mu, \sigma)$). In the special case $\mu = 0$ and $\sigma = 1$, X is said to have the *standard Gaussian* distribution and the density becomes

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

A plot of the pdf f reveals a classical "bell-shaped" curve, centered at (and symmetric around) $x = \mu$, and with "width" determined by σ. (The precise meaning of this latter statement will become clear when we discuss the variance below.)

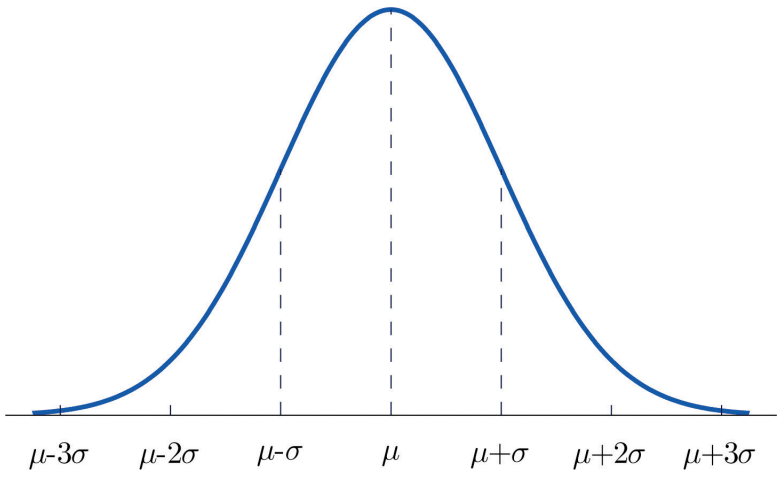

Figure 1: *Pdf of a Gaussian random variable with mean μ and variance σ^2.*

The density integrates to 1:

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1. \tag{1}$$

The fact that this integral evaluates to 1 is a routine exercise in integral calculus, and is left as an exercise (or feel free to look it up in any standard book on probability or on the internet).

In matlab, the function 'randn' gives a random realization drawn from the $N(0,1)$ distribution. Now suppose we would like to generate a realization Y from the distribution $N(\mu,1)$ using the realization we get from $X \sim N(0,1)$. Then the density we want to sample is the same as that of X but shifted to be centered at μ. Therefore to generate such samples Y, it seems natural to generate X samples and add μ to them. Let us confirm this method is accurate by computing the pdf of Y if it is generated as $Y = X + \mu$.

Since $Y = X + \mu$, a natural guess for the pdf of Y in terms of the pdf of X is $f_Y(y) = f_X(y-\mu)$. What if we want to generate $Z \sim N(0,\sigma^2)$ from X? A natural guess is $Z = \sigma X$. Is the pdf of Z $f_Z(z) = f_X(\frac{z}{\sigma})$? These statements are trivially correct if instead of a pdf we had a pmf. However, we should be more careful in our setting, because the probability density function is not a probability. How do we know whether these statements are correct? A trick about continuous probability is that we should always try to go back to 'probability' terms, to avoid any confusion with the density function. So to make sure we are doing things correctly, we start by computing the cdf of Y and Z, and then differentiate the answer to obtain the respective pdfs.

Let us start by computing the cdf of Y:

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \le y) \\ &= \mathbf{P}(X + \mu \le y) \\ &= \mathbf{P}(X \le y - \mu) \\ &= F_X(y-\mu) \end{aligned}$$

We differentiate it to get the pdf:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(y-\mu) \\ &= f_X(y-\mu) \\ &= \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2}} \end{aligned}$$

This verifies the guess we had above for the pdf of Y. As for Z:

$$\begin{aligned} F_Z(z) &= \mathbf{P}(Z \le z) \\ &= \mathbf{P}(\sigma X \le z) \\ &= \mathbf{P}(X \le \frac{z}{\sigma}) \\ &= F_X(\frac{z}{\sigma}) \end{aligned}$$

Finally, we get that

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_X(\frac{z}{\sigma}) \\ &= \frac{1}{\sigma} f_X(\frac{z}{\sigma}) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} \end{aligned}$$

Thus, $Z \sim N(0,\sigma^2)$. On the other hand, our earlier guess that $f_Z(z) = \frac{1}{\sigma} f_X(\frac{z}{\sigma})$ is NOT correct. Going back to computing a quantity's associated probability is important to verify that what you are claiming about the pdf is correct.

Why does the $\frac{1}{\sigma}$ factor appear for the pdf of $Z \sim N(0,\sigma^2)$? Suppose we have a standard Gaussian random variable for X, and let Y$= 3X$ and $\sigma = 3$. The variable Y will be spread out 3 times as much as the standard Gaussian. So the probability density curve will become 'fatter', having a bigger variation. However, since it still needs to integrate to 1, it will need to be rescaled so that the area under the curve is equal to 1. This gives a geometric interpretation to the additional scalar factor.

Alternatively, we can explain the extra factor because we are dealing with densities and not probabilities. The difference between densities and probabilities is in the units: probability does not have a unit, whereas the probability density function has the inverse of the unit of the random variable. Suppose for example X is in meters, and we would like to change it in feet. We multiply X by some constant c to achieve that. Then the probability density function should change units as well. That is why we scale it by $\frac{1}{c}$ which represents the change in the units of the pdf.

We conclude that if $X \sim N(0,1)$,

(1) $X + \mu \sim N(\mu, 1)$

(2) $\sigma X \sim N(0,\sigma^2)$

More generally,

$$W = \sigma X + \mu \sim N(\mu, \sigma^2).$$

Finally, we compute the mean and variance of Gaussian random variables. First if $X \sim N(0,1)$,

(1) $\mathbb{E}[X] = 0$ (by symmetry of the pdf around 0)

(2) $\text{Var}(X) = 1$ (integration by parts)

For $W = \sigma X + \mu$,

(1) $\mathbb{E}[W] = \mu$

(2) $\text{Var}(W) = \sigma^2 \times 1 = \sigma^2$

As a consequence, we see that μ and σ^2 are the mean and the variance of a $N(\mu,\sigma^2)$ random variable.

Central Limit Theorem

If we look at all the distributions we have covered so far, they all had a motivation for their underlying probability model. For example the binomial models the number of successes, the Poisson was a limit of the binomial, etc. However, we have not given a motivation for the Gaussian distribution so far. So why is it important?

Consider the random variables $X_1 \cdots, X_n$, that are independent and identically distributed (i.i.d) each with mean μ and variance σ^2. The Central Limit Theorem says that the sum of these variables $S_n := X_1 + \cdots + X_n \approx$ Gaussian with mean $\mathbb{E}[S_n] = n\mu$ and variance $\text{Var}(()S_n) = n\sigma^2$ for large n . Therefore the Gaussian distribution can be used to approximate any situation where we add up many independent effects.

The Central Limit Theorem applies to continuous and discrete random variables. We will use the cdf of the sum to state that it becomes close to the cdf of the Gaussian. Let $S_n = X_1 + \cdots + X_n$. We want to say that the cdf of S_n is close to the cdf of a Gaussian random variable as n goes to infinity.

Suppose $W \sim N(\mu, \sigma^2)$, then

$$\frac{W-\mu}{\sigma} \sim N(0,1).$$

We can standardize any Gaussian random variable using that trick. How do we standardize S_n? We can do that by subtracting $n\mu$ and dividing by $n\sigma^2$. We know

$$X_1 + \cdots + X_n \approx N(n\mu, n\sigma^2)$$

which is equivalent to

$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \approx N(0,1)$$

So $\forall z\,,\ \lim_{n\to\infty} \mathbf{P}(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le z) = \mathbf{P}(Z \le z),$ where $Z \sim N(0,1)$.

Theorem 14.1: **[Central Limit Theorem]** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \text{Var}(X_i)$ (both assumed to be $< \infty$). Define $A_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$. Then as $n \to \infty$, the distribution of A_n approaches the standard Gaussian distribution in the sense that, for any real z,*

$$\mathbf{P}(A_n \le z) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx \quad \text{as } n \to \infty.$$

The Central Limit Theorem is a very striking fact. What it says is the following. If we take an average of n observations of absolutely any r.v. X, then the distribution of that average will be approximately a bell-shaped curve centered at $\mu = \mathbb{E}[X]$. Thus all trace of the distribution of X disappears as n gets large: all distributions, no matter how complex,[1] look like the Gaussian distribution when they are averaged. The only effect of the original distribution is through the variance σ^2, which determines the width of the curve for a given value of n, and hence the rate at which the curve shrinks to a spike.

Let's reconsider the polling example. $X_i = 1$ if ith person is a democrat, and 0 otherwise. Suppose the X_i variables are independent and identically distributed. We would like to estimate p, the fraction of the population that is democrat which is equal to $\mathbf{P}(X_i = 1)$. We estimate $\hat{p} = \frac{1}{n}\sum X_i$, and would like our estimate to satisfy $\mathbf{P}(|\hat{p} - p| > 0.1) \le 0.05$. How many people do we have to survey to achieve that? We used the Chebyshev's inequality in a previous lecture to solve this problem. However, Chebyshev's inequality gives

[1] We do need to assume that the mean and variance of X are finite.

a bound and we would like to use the CLT to get an approximation of the actual value. Please continue this question as a homework exercise.