COS 324, Precept #2:
A Crash Course on Continuous Probability

October 14, 2017

1 Review (written on side board before class) (5 min)

Recall from discrete probability:

- A *probability space* (Ω, Pr) is a countable set Ω with a function $\text{Pr} : \Omega \to \mathbb{R}_+$.

- An *event* is a set of outcomes $A \subseteq \Omega$. It has probability $\text{Pr}[A] = \sum_{\omega \in A} \text{Pr}[\omega]$.

- A (real-valued) *random variable* $X : \Omega \to \mathbb{R}$ assigns a real number to each outcome. $X = a$ is an event for each $a \in \mathbb{R}$.

- Random variables X and Y are *independent* if

$$\text{Pr}[X = x \wedge Y = y] = \text{Pr}[X = x] \cdot \text{Pr}[Y = y].$$

- The *distribution* of a random variable X is the function mapping a to $\text{Pr}[X = a]$. Its nonzero domain is called the *support* and denoted $\text{supp}(X)$.

- The *expected value* of a r.v. X is $\mathbb{E}[X] = \sum_{x \in \text{supp}(X)} x \cdot \text{Pr}[X = x]$. It is a *linear* operator: $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$.

- The *variance* of a r.v. X is $\text{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

And the *Gaussian integral* (you're encouraged to find a proof online):

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}.$$

1

2 Continuous Probability Spaces (10 min)

We understand continuous probability intuitively: e.g. if we throw a dart randomly on a dartboard, the probability that it lands in the upper right quadrant is 1/4. But the probability that it lands anywhere in particular (like the exact center), or even a "set of negligible area" (like the boundary) is zero.

It's clear that we should reconcile this using a calculus of infinitesimal outcomes and events. But to pin down a completely rigorous formulation is a subtle matter.[1]

Throughout this precept, let's keep in mind as a reference the uniform probability space on the unit interval: we want to construct a probability space on $\Omega = [0,1]$.

We can't start by assigning a probability to each outcome. Instead let's try to build our probability space from events (subsets of Ω):

1. As before, $\Pr[\Omega]$ should be 1.

2. If A and B are disjoint intervals, $\Pr[A \cup B]$ should be $\Pr[A] + \Pr[B]$.

3. The probability of an interval $[a,b] \subseteq \Omega$ should be $b - a$.

This is almost the entire story! As long as $\Pr[\cdot]$ satisfies (1) and (2), it is a valid probability function on Ω. In (3) we defined a *concrete* probability space $(\Omega, \Pr)$. [2]

3 Continuous Random Variables (20 min)

The definition of a random variable carries over straightforwardly: it is a function that assigns a real number to each outcome.

But in applied settings, the distribution of a random variable is often the primary workhorse. Most continuous random variables we care about have a *probability density function*, which assigns an infinitesimal weight to each outcome. Rather than defining it immediately, let us first informally state that it is the "histogram" associated with a random variable and the area under the histogram should be 1.

Some examples:

- The *uniform* distribution $\text{Unif}([0,1])$: $\rho(x) = 1$ if $x \in [0,1]$ and 0 otherwise.

- In general, for $\text{Unif}([a,b])$, we have $\rho(x) = \frac{1}{b-a}$ on its support.

- The "dartboard distribution" $\text{Unif}(D^2)$: $\rho(x,y) = \frac{1}{\pi}$ on its support. Notice that we sneakily introduced a vector-valued random variable.

---

[1] In real analysis you can learn in depth about the Lebesgue measure, σ-algebras, and Borel sets.

[2] The last part is notoriously hard to define in full generality. It's not quite as easy as assigning a weighted area to each set. For instance, the Banach-Tarski paradox gives a way to dissect a sphere into two spheres of the same volume. To avert this, we need to give up trying to assign probabilities to some *non-measurable* sets, which (very informally) tend to look like fractals, or countable number of holes. Thankfully, we almost never run into these in machine learning.

- The triangle distribution is a non-uniform distribution on $[0, 1]$: $\rho(x) = 2x$. It is more likely to be close to 1 than 0.

This brings us to the definition we need: the pdf of X is the function ρ so that $\int_{x \in A} \rho(x) dx = \Pr[X \in A]$. It's the continuous analogue of the probability mass function (though it is less common to call the pdf the distribution). If we have the pdf of a random variable, we can compute the probability of any event using an integral. In the last example,

$$\Pr[X > 1/2] = \int_{1/2}^{1} \rho(x) \, dx = 3/4.$$

Or, for the dartboard,

$$\Pr[X > 0, Y > 0] = \int_{0}^{1} \int_{0}^{\pi/2} \rho(x) \, dr \, d\theta = 1/4.$$

Let's add one more example, an important distribution whose support is all of $\mathbb{R}$:

- The *normal* (or *Gaussian*) distribution $\mathcal{N}(\mu, \sigma^2)$: $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$. (Sketch a plot!) For example, $\mathcal{N}(100, 15^2)$ is the pdf of X, where X is a randomly selected person's IQ.

If $X \sim \mathcal{N}(100, 15^2)$, then
$$\Pr[X \le 100] = 1/2,$$

$$\Pr[X \le 115] = \int_{-\infty}^{115} \rho(x) \, dx \approx 0.84,$$

and

$$\Pr[X \le 160] = \int_{-\infty}^{160} \rho(x) \, dx \approx 0.99997$$

Sometimes it is convenient to work with $\Pr[X \le a]$, which called the *cumulative distribution function* of X. Notice that it is increasing, and its derivative is the pdf.

4 Expectations and Variance (10 min)

The expectation of a continuous random variable X with pdf ρ is just

$$\mathbb{E}[X] = \int_{\text{supp}(X)} x \cdot \rho(x) \, dx.$$

For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$,
$$\mathbb{E}[X] = \mu,$$

which you can prove by brute force or symmetry.

3

Now, let's compute the variance

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mu)^2\right].$$

This is

$$\int_{\mathbb{R}} (x - \mu)^2 \rho(x) = \int_{\mathbb{R}} x^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-u^2/2\sigma^2} \, du = \sigma^2.[3]$$

In summary: $\mathcal{N}(\mu, \sigma^2)$ is the "bell curve-shaped" distribution that has mean μ and variance σ^2. They're all "equivalent" up to translation and scaling, which is true of many families of distributions.

Finally, we note the following fundamental property of Gaussians.

Theorem 4.1. *Let* $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ *be independent. Then,*

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

This is deceptively tricky, and you should try to prove it yourself. Letting $Z = X + Y$, write

$$\mathbb{E}[Z] = \int_{\mathbb{R}} z \cdot \rho_Z(z) \, dz = \iint_{\mathbb{R}^2} (x + y) \cdot \rho_X(x)\rho_Y(y) \, dx \, dy.$$

The last inequality uses independence, and it may be helpful to sit down later and convince yourself that it is true. From there, it is an integration exercise.

A couple of final remarks, as a preview of where we'll see Gaussians in this course:

- In real life, Gaussians are a good model for many noisy quantities, like "the IQ of a human is around 100, with variation on the scale of 15".

- The fact that a sum of Gaussians is a Gaussian is very convenient and fundamental. It tells us how sources of error accumulate in experiments. It's also related to the *central limit theorem*, which (informally!) says that if you flip 1000 coins, the number of heads is close to $\mathcal{N}(500, 250)$.

- We can choose to view X and Y as a 2-dimensional random vector. Then, we say that (X, Y) has the *multivariate normal distribution*. You'll become well-acquainted with multivariate Gaussians in machine learning.

---

[3]You can prove this by using Feynman's favorite technique "differentiating the integral sign."