Key Ideas from Probability Theory

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

January 25, 2021

Outline

Outline for today:

- Describing probability distributions
- Homework-related examples

Building blocks

Probability distributions

A *probability measure* is a function Pr that takes subsets of the sample space[1] as inputs and produces real numbers as outputs, subject to certain constraints called the Kolmogorov axioms:

- $0 \leq \mathrm{Pr}(A) \leq 1$
- $\mathrm{Pr}(S) = 1$
- If $E \cap F = \emptyset$, then $\mathrm{Pr}(E \cup F) = \mathrm{Pr}(E) + \mathrm{Pr}(F)$

[1] Strictly speaking, not every subset can be allowed in some cases. For a few details, see "Finer points" in chapter 1 of the 464 lecture notes. For a lot of details, take MATH 523A.

Describing probability distributions on $\mathbb{R}$

We'll focus on real-valued random variables at first: those whose sample space is a subset of the real line $\mathbb{R}$

- Discrete random variable: sample space is a set of discrete points (e.g., the integers)
  Described by a *probability mass function*

$$p_X(x) = \Pr(X = x)$$

- Continuous random variable: sample space is (typically) an interval (could be half- or fully-infinite)
  Described by a *probability density function*

$$\Pr(a < X < b) = \int_a^b p_X(x) dx$$

Describing probability distributions on $\mathbb{R}$

Cumulative distribution functions

Alternatively, a random variable can be described by its *cumulative distribution function*:

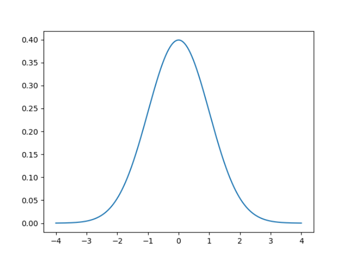
$$F(x) = \Pr(X \leq x)$$

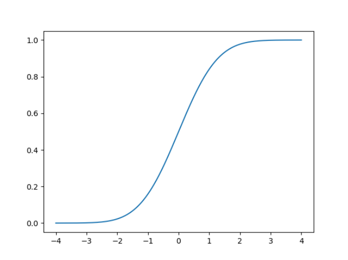
i.e., $F(x)$ gives you the probability that the RV falls below x.
Properties:

- non-decreasing
- goes to 0 as $x \to -\infty$, goes to 1 as $x \to \infty$
- If X is a continuous RV, $F'(x) = p_X(x)$

Cumulative distribution functions

Example: normal distribution pdf vs cdf





CDF example

Select uniformly at random a point in the triangle with vertices at $(0,0),(1,0),(0,2)$; let X be the x-coordinate of this point.

What is the CDF of this random variable? What is the PDF?

CDF example

Distributions with parameters

Almost all of our interest is focused on families of distributions depending on parameters; e.g., the normal distribution has parameters μ, σ and pdf

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

Note we may sometimes write:

- $p(x)$ – probability density function
- $p(x|\mu, \sigma)$ – explicitly noting that $p(x)$ is conditional on the values of the parameters μ, σ

Normalized and unnormalized distribution

A distribution function (mass or density) is *normalized* if $\sum_x p(x) = 1$ or $\int_{\mathbb{R}} p(x)dx = 1$.

For example:

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

is *not* normalized because it integrates to $\sqrt{2\pi\sigma^2}$; so, the "probability distribution" defined by this function isn't really a probability distribution.

As we'll see, when applying Bayes' theorem it is sometimes easier to work with the un-normalized distribution, especially as an intermediate step.

Proper and improper densities

A density function is *improper* if it cannot be normalized; for example, a uniform density

$$p(x) \propto 1$$

can be normalized on any *finite* interval (a, b) to get an honest pdf

$$p(x) = \frac{1}{b - a}$$

However, there is no true uniform probability distribution on $(0, \infty)$ or $(-\infty, \infty)$ – but sometimes we act as if there is.

Working with distributions in SciPy

Distributions in SciPy

In the `scipy.stats` module there are a number of classes for standard probability distributions. These can be used to:

- compute PDF/PMF values
- compute cumulative probabilities (CDF values)
- draw random samples

Let's see examples...

Conditional probability, independence, and Bayes' theorem

Conditional probability and independence

If the probability of an event represents our knowledge about that event, we should be able to "update" this knowledge by incorporating observations:

$$\Pr(E|H) = \text{"probability of E given H"}$$

E and H are said to be *independent* if $\Pr(E|H) = \Pr(E)$.

Multiplication rule for probabilities

The multiplication or *chain rule* for probabilities of intersections of events is:

$$\Pr(E \cap H) = \Pr(E|H)\Pr(H) = \Pr(H|E)\Pr(E)$$

Independence

This leads to an alternative characterization of independence for events; two events are independent if:

$$\Pr(E \cap H) = \Pr(E)\Pr(H)$$

Often this is taken as the starting definition of independence.

Pairwise vs. mutual independence

One of the homework problems deals with the issue of pairwise or mutual independence:

- pairwise independence of $A_1, A_2, A_3, \ldots$: given any two i, j, $\Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j)$.
- mutual independence of $A_1, A_2, A_3, \ldots$: given any subset $i_1, i_2, \ldots, i_n$, $\Pr(A_{i_1} \cap \ldots \cap A_{i_n}) = \Pr(A_{i_1}) \ldots \Pr(A_{i_n})$

Example from the homework

Independence of random variables

Two random variables X, Y are independent if the joint probability mass/density function factors:

$$p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$$

Bayes' theorem

The theorem that gives Bayesian statistics its name is a seemingly trivial rearrangement of the equations above:

$$\mathrm{Pr}(E \cap H) = \mathrm{Pr}(E|H)\mathrm{Pr}(H) = \mathrm{Pr}(H|E)\mathrm{Pr}(E)$$

to

$$\mathrm{Pr}(H|E) = \frac{\mathrm{Pr}(E|H)\mathrm{Pr}(H)}{\mathrm{Pr}(E)}$$

The significance comes when we assign interpretations to H and E of "hypothesis" and "evidence" respectively.

The cookie problem

Suppose we have two bowls of cookies.[2] Bowl 1 has 30 vanilla and 10 chocolate cookies; Bowl 2 has 20 of each.

We select a bowl at random and, without looking at which one we picked, pull a cookie at random from it. The cookie is vanilla.

What is the probability that our randomly selected bowl was Bowl 1?

[2] This example is from *Think Bayes* by Allen Downey.

The cookie problem