1 Probability and Bayes Theorem

1.1 Introduction

Probability theory is one of the three central mathematical tools in machine learning, along with multivariable calculus and linear algebra. Tools from probability allow us to manage the uncertainty inherent in data collected from real world experiments, and to measure the reliability of predictions that we might make from that data. In these notes, we will review some of the basic terminology of probability and introduce Bayesian inference as a technique in machine learning problems.

This will only be a superficial introduction to ideas from probability. For a thorough treatment, see this open-source introduction to probability. For a more applied emphasis, I recommend the excellent online course Probabilistic Systems Analysis and Applied Probability and its associated text [1].

1.2 Probability Basics

The theory of probability begins with a set X of possible events or outcomes, together with a "probability" function P on (certain) subsets of X that measures "how likely" that combination of events is to occur.

The set X can be discrete or continuous. For example, when flipping a coin, our set of possible events would be the discrete set $\{H, T\}$ corresponding to the possible events of flipping heads or tails. When measuring the temperature using a thermometer, our set of possible outcomes might be the set of real numbers, or perhaps an interval in $\mathbb{R}$. The thermometer's measurement is random because it is affected by, say, electronic noise, and so its reading is the true temperature perturbed by a random amount.

The values of P are between 0, meaning that the event *will not* happen, and 1, meaning that it is certain to occur. As part of our set up, we assume that the total chance of some event from X occurring is 1, so that $P(X) = 1$; and the chance of "nothing" happening is zero, so $P(\emptyset) = 0$. And if $U \subset X$ is some collection, then $P(U)$ is the chance of an event from U occurring.

The last ingredient of this picture of probability is additivity. Namely, we assume that if U and V are subsets of X that are disjoint, then

$$P(U \cup V) = P(U) + P(V).$$

Even more generally, we assume that this holds for (countably) infinite collections of disjoint subsets $U_1, U_2, \ldots$, where

$$P(U_1 \cup U_2 \cup \cdots) = \sum_{i=1}^{\infty} P(U_i)$$

Definition: The combination of a set X of possible outcomes and a probability function P on subsets of X that satisfies $P(X) = 1$, $0 \leq P(U) \leq 1$ for all U, and

is additive on countable disjoint collections of subsets of X is called a (naive) probability space. X is called the *sample space* and the subsets of X are called *events*.

Warning: The reason for the term "naive" in the above definition is that, if X is an uncountable set such as the real numbers $\mathbb{R}$, then the conditions in the definition are self-contradictory. This is a deep and rather surprising fact. To make a sensible definition of a probability space, one has to restrict the domain of the probability function P to certain subsets of X. These ideas form the basis of the mathematical subject known as measure theory. In these notes we will work with explicit probability functions and simple subsets such as intervals that avoid these technicalities.

1.2.1 Discrete probability examples

The simplest probability space arises in the analysis of coin-flipping. As mentioned earlier, the set X contains two elements $\{H, T\}$. The probability function P is determined by its value $P(\{H\}) = p$, where $0 \leq p \leq 1$, which is the chance of the coin yielding a "head." Since $P(X) = 1$, we have $P(\{T\}) = 1 - p$.

Other examples of discrete probability spaces arise from dice-rolling and playing cards. For example, suppose we roll two six-sided dice. There are 36 possible outcomes from this experiment, each equally likely. If instead we consider the sum of the two values on the dice, our outcomes range from 2 to 12 and the probabilities of these outcomes are given by

2	3	4	5	6	7	8	9	10	11	12
1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

A traditional deck of 52 playing cards contains 4 aces. Assuming that the chance of drawing any card is the same (and is therefore equal to 1/52), the probability of drawing an ace is $4/52 = 1/13$ since

$$P(\{A_{\clubsuit}, A_{\spadesuit}, A_{\heartsuit}, A_{\diamondsuit}\}) = 4P(\{A_{\clubsuit}\}) = 4/52 = 1/13$$

1.2.2 Continuous probability examples

When the set X is continuous, such as in the temperature measurement, we measure $P(U)$, where $U \subset X$, by giving a "probability density function" $f : X \to \mathbb{R}$ and declaring that

$$P(U) = \int_U f(x)dX.$$

Notice that our function $f(x)$ has to satisfy the condition

$$P(X) = \int_X f(x)dX = 1.$$

For example, in our temperature measurement example, suppose the "true" outside temperature is t_0, and our thermometer gives a reading t. Then a good model for the random error is to assume that the error $t - t_0$ is governed by the density function

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

where σ is a parameter. In a continuous situation such as this one, the probability of any particular outcome in X is zero since

$$P(\{t\}) = \int_t^t f_\sigma(x)dx = 0$$

Still, the shape of the density function does tell you where the values are concentrated – values where the density function is larger are more likely than those where it is smaller.

With this density function, the error in our measurement is given by

$$P(|t - t_0| < \delta) = \int_{-\delta}^{\delta} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx \tag{1}$$

The parameter σ (called the *standard deviation*) controls how tightly the thermometer's measurement is clustered around the true value t_0; when σ is large, the measurements are scattered widely, when small, they are clustered tightly. See fig. 1.

1.3 Conditional Probability and Bayes Theorem

The theory of conditional probability gives a way to study how partial information about an event informs us about the event as a whole. For example, suppose you draw a card at random from a deck. As we've seen earlier, the chance that card is an ace is 1/13. Now suppose that you learn that (somehow) that the card is definitely not a jack, king, or queen. Since there are 12 cards in the deck that are jacks, kings, or queens, the card you've drawn is one of the remaining 40 cards, which includes 4 aces. Thus the chance you are holding an ace is now $4/40 = 1/10$.

In terms of notation, if A is the event "my card is an ace" and B is the event "my card is not a jack, queen, or king" then we say that *the probability of A given B* is 1/10. The notation for this is

$$P(A|B) = 1/10.$$

More generally, if A and B are events from a sample space X, and $P(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

so that $P(A|B)$ measures the chance that A occurs among those situations in which B occurs.

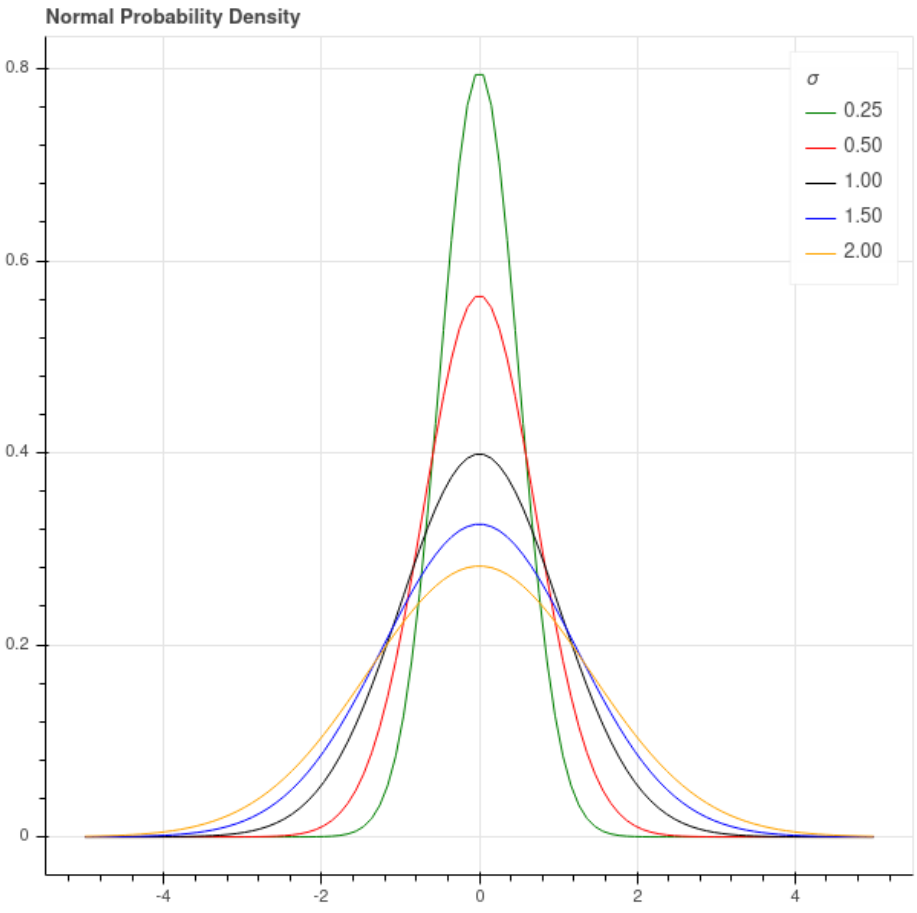

Figure 1: Normal Density

1.3.1 Bayes Theorem

Bayes theorem is a foundational result in probability.

Theorem: Bayes Theorem says

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

If we use the definition of conditional probability given above, this is straightforward:

$$\frac{P(B|A)P(A)}{P(B)} = \frac{P(B \cap A)}{P(B)} = P(A|B).$$

1.3.2 An example

To illustrate conditional probability, let's consider what happens when we administer the most reliable COVID-19 test, the PCR test, to an individual drawn from the population at large. There are two possible test results (positive and negative) and two possible true states of the person being tested (infected and not infected). Suppose I go to the doctor and get a COVID test which comes back positive. What is the probability that I actually have COVID?

Let's let S and W stand for infected (sick) and not infected (well), and let $+/-$ stand for test positive or negative. Note that there are four possible outcomes of our experiment:

- test positive and infected (S+) – this is a *true positive*.
- test positive and not infected (W+) – this is a *false positive*.
- test negative and infected (S-) – this is a *false negative*.
- test negative and not infected (W-) – this is a *true negative*.

The CDC says that the chance of a false positive – that is, the percentage of samples from well people that incorrectly yields a positive result – is about one-half of one percent, or 5 in 1000.

In other words,

$$P(+|W) = P(W+)/P(W) = 5/1000 = 1/200$$

On the other hand, the CDC tells us that chance of a false negative is 1 in 4, so

$$P(-|S) = P(S-)/P(S) = .25.$$

Since $P(S-) + P(S+) = P(S)$. since every test is either positive or negative, we have

$$P(+|S) = .75.$$

Suppose furthermore that the overall incidence of COVID-19 in the population is p. In other words, $P(S) = p$ so $P(W) = 1 - p$. Then

$$P(S+) = P(S)P(+|S) = .75p$$

and

$$P(W+) = P(W)P(+|W) = .005(1-p).$$

Putting these together we get $P(+) = .005 + .745p$

What I'm interested in is $P(S|+)$ – the chance that I'm sick, given that my test result was positive. By Bayes Theorem,

$$P(S|+) = \frac{P(+|S)P(S)}{P(+)} = .75p/(.005 + .745p) = \frac{750p}{5 + 745p}.$$

As fig. 2 shows, if the population incidence is low then a positive test is far from conclusive. Indeed, if the overall incidence of COVID is one percent, then a positive test result only implies a 60 percent chance that I am in fact infected.

Just to fill out the picture, we have

$$P(-) = P(S-) + P(W-) = (P(S) - P(S+)) + (P(W) - P(W+))$$

which yields

$$P(-) = 1 - .005 + .005p - .75p = .995 - .745p.$$

Using Bayes Theorem, we obtain

$$P(S|-) = \frac{P(-|S)P(S)}{P(-)} = .25p/(.995 - .745p) = \frac{250p}{995 - 745p}.$$

In this case, even though the false negative rate is pretty high (25 percent) overall, if the population incidence is one percent, then the probability that you're sick given a negative result is only about .25 percent. So negative results are very likely correct!

1.4 Independence

Independence is one of the fundamental concepts in probability theory. Conceptually, two events are independent if the occurrence of one has does not influence the likelihood of the occurrence of the other. For example, successive flips of a coin are independent events, since the result of the second flip doesn't have anything to do with the result of the first. On the other hand, whether or not it rains today and tomorrow are not independent events, since the weather tomorrow depends (in a complicated way) on the weather today.

We can formalize this idea of independence using the following definition.

Definition: Let X be a sample space and let A and B be two events. Then A and B are *independent* if $P(A \cap B) = P(A)P(B)$. Equivalently, A and B are independent if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

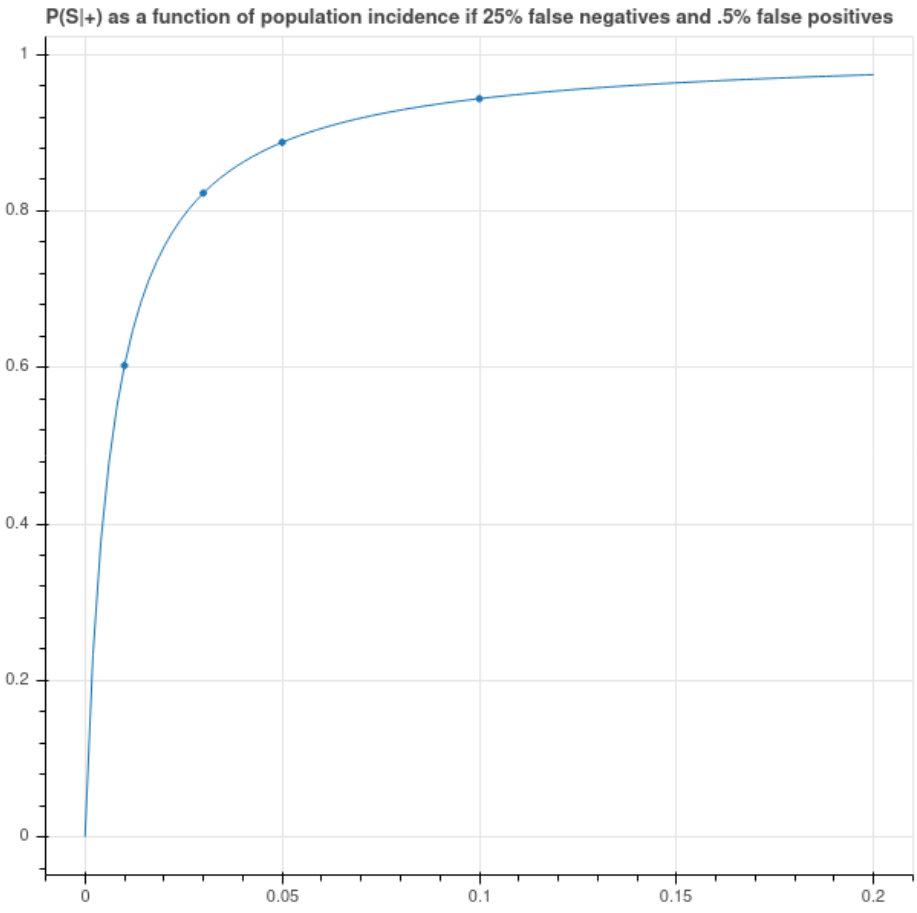

Figure 2: P(S|+) vs P(S)

1.4.1 Examples

1.4.1.1 Coin Flipping Suppose our coin has a probability of heads given by a real number p between 0 and 1, and we flip our coin N times. What is the chance of gettting k heads, where $0 \leq k \leq N$? Any particular sequence of heads and tails containing k heads and $N-k$ tails has probability

$$P(\text{particular sequence of } k \text{ heads among } N \text{ flips}) = p^k(1-p)^{N-k}.$$

In addition, there are $\binom{N}{k}$ sequences of heads and tails containing k heads. Thus the probability $P(k,N)$ of k heads among N flips is

$$P(k,N) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{2}$$

Notice that the binomial theorem gives us $\sum_{k=0}^{N} P(k,N) = 1$ which is a reassuring check on our work.

The probability distribution on the set $X = \{0, 1, \ldots, N\}$ given by $P(k,N)$ is called the *binomial distribution* with parameters N and p.

1.4.1.2 A simple 'mixture' Now let's look at an example of events that are not independent. Suppose that we have two coins, with probabilities of heads p_1 and p_2 respectively; and assume these probabilities are different. We play the a game in which we first choose one of the two coins (with equal chance) and then flip it twice. Is the result of the second flip independent of the first? In other words, is $P(HH) = P(H)^2$?

This type of situation is called a 'mixture distribution' because the probability of a head is a "mixture" of the probability coming from the two different coins.

The chance that the first flip is a head is $(p_1 + p_2)/2$ because it's the chance of picking the first coin, and then getting a head, plus the chance of picking the second, and then getting a head. The chance of getting two heads in a row is $(p_1^2 + p_2^2)/2$ because it's the chance, having picked the first coin, of getting two heads, plus the chance, having picked the second, of getting two heads.

Since

$$\frac{p_1^2 + p_2^2}{2} \neq \left(\frac{p_1 + p_2}{2}\right)^2$$

we see these events are not independent.

In terms of conditional probabilities, the chance that the second flip is a head, given that the first flip is, is computed as:

$$P(HH|H) = \frac{p_1^2 + p_2^2}{p_1 + p_2}.$$

From the Cauchy-Schwartz inequality one can show that

$$\frac{p_1^2 + p_2^2}{p_1 + p_2} > \frac{p_1 + p_2}{2}.$$

Why should this be? Why should the chance of getting a head on the second flip go up given that the first flip was a head? One way to think of this is that the first coin flip contains a little bit of information about which coin we chose. If, for example $p_1 > p_2$, and our first flip is heads, then it's just a bit more likely that we chose the first coin. As a result, the chance of getting another head is just a bit more likely than if we didn't have that information. We can make this precise by considering the conditional probability $P(p = p_1|H)$ that we've chosen the first coin given that we flipped a head. From Bayes' theorem:

$$P(p = p_1|H) = \frac{P(H|p = p_1)P(p = p_1)}{P(H)} = \frac{p_1}{p_1 + p_2} = \frac{1}{1 + (p_2/p_1)} > \frac{1}{2}$$

since $(1 + (p_2/p_1)) < 2$.

Exercise: Push this argument a bit further. Let $p_1 = \max(p_1, p_2)$ Let P_N be the conditional probability of getting heads assuming that the first N flips were heads. Show that $P_N \to p_1$ as $N \to \infty$. All those heads piling up make it more and more likely that you're flipping the first coin and so the chance of getting heads approaches p_1.

1.4.1.3 An example with a continuous distribution Suppose that we return to our example of a thermometer which measures the ambient temperature with an error that is distributed according to the normal distribution, as in eq. 1. Suppose that we make 10 independent measurements $t_1, \ldots, t_{10}$ of the true temperature t_0. What can we say about the distribution of these measurements?

In this case, independence means that

$$P = P(|t_1 - t_0| < \delta, |t_2 - t_0| < \delta, \ldots) = P(|t_1 - t_0| < \delta)P(|t_2 - t_0| < \delta) \cdots P(|t_{10} - t_0| < \delta)$$

and therefore

$$P = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{10} \int_{-\delta}^{\delta} \cdots \int_{-\delta}^{\delta} e^{-(\sum_{i=1}^{10} x_i^2)/2\sigma^2} dx_1 \cdots dx_{10}$$

One way to look at this is that the vector $\mathbf{e}$ of errors $(|t_1 - t_0|, \ldots, |t_{10} - t_0|)$ is distributed according to a *multivariate gaussian distribution*:

$$P(\mathbf{e} \in U) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{10} \int_U e^{-\|x\|^2/2\sigma^2} d\mathbf{x}$$

where U is a region in $\mathbf{R}^{10}$.

The multivariate gaussian can also describe situations where independence does not hold. For simplicity, let's work in two dimensions and consider the probability density on $\mathbf{R}^2$ given by

$$P(\mathbf{e} \in U) = A \int_U e^{-(x_1^2 - x_1 x_2 + x_2^2)/2\sigma^2} d\mathbf{x}.$$

where the constant A is chosen so that

$$A \int_{\mathbf{R}^2} e^{-(x_1^2 - x_1 x_2 + x_2^2)/2\sigma^2} d\mathbf{x} = 1.$$

This density function as a "bump" concentrated near the origin in $\mathbf{R}^2$, and its level curves are a family of ellipses centered at the origin. See fig. 3 for a plot of this function with $\sigma = 1$.

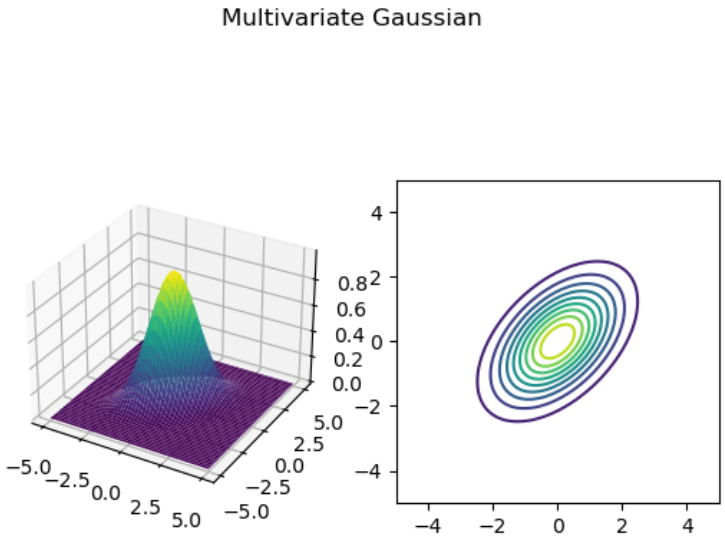

Figure 3: Multivariate Gaussian

In this situation we can look at the conditional probability of the first variable given the second, and see that the two variables are not independent. Indeed, if we fix x_2, then the distribution of x_1 depends on our choice of x_2. We could see this by a calculation, or we can just look at the graph: if $x_2 = 0$, then the most likely values of x_1 cluster near zero, while if $x_2 = 1$, then the most likely values of x_1 cluster somewhere above zero.

1.5 Random Variables, Mean, and Variance

Typically, when we are studying a random process, we aren't necessarily accessing the underlying events, but rather we are making measurements that provide us with some information about the underlying events. For example, suppose our sample space X is the set of throws of a pair of dice, so X contains the 36 possible combinations that can arise from the throws. What we are actually interested is the sum of the values of the two dice – that's our "measurement" of this system. This rather vague notion of a measurement of a random system is captured by the very general idea of a *random variable*.

Definition: Let X be a sample space with probability function P. A *random variable* on X is a function $f : X \to \mathbb{R}$.

Given a random variable f, we can use the probability measure to decide how likely f is to take a particular value, or values in a particular set by the formula

$$P(f(x) \in U) = P(f^{-1}(U))$$

In the dice rolling example, the random variable S that assigns their sum to the pair of values obtained on two dice is a random variable. Those values lie between 2 and 12 and we have

$$P(S = k) = P(S^{-1}(\{k\})) = P(\{(x, y) : x + y = k\})$$

where (x, y) runs through $\{1, 2, \ldots, 6\}^2$ representing the two values and $P((x, y)) = 1/36$ since all throws are equally likely.

Let's look at a few more examples, starting with what is probably the most fundamental of all.

Definition: Let X be a sample space with two elements, say H and T, and suppose that $P(H) = p$ for some $0 \leq p \leq 1$. Then the random variable that satisfies $f(H) = 1$ and $f(T) = 0$ is called a Bernoulli random variable with parameter p.

In other words, a Bernoulli random variable gives the value 1 when a coin flip is heads, and 0 for tails.

Now let's look at what we earlier called the binomial distribution.

Definition: Let X be a sample space consisting of strings of H and T of length N, with the probability of a *particular string* S with k heads and $N - k$ tails given by

$$P(S) = p^k(1 - p)^{N-k}$$

for some $0 \leq p \leq 1$. In other words, X is the sample space consisting of N independent flips of a coin with probability of heads given by p.
Let $f : X \to \mathbb{R}$ be the function which counts the number of H in the string. Then f is a random variable called a *binomial random variable* with parameters

p and N. Clearly, a binomial random variable with $N = 1$ is just a Bernoulli variable with parameter p.

If f is a binomial random variable with parameters p and N, then

$$P(f = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

since $f^{-1}(\{k\})$ is the number of elements in the subset of strings of H and T of length N containing exactly k H's.

For an example with a continuous random variable, suppose our sample space is $\mathbf{R}^2$ and the probability density is the simple multivariate normal

$$P(\mathbf{e} \in U) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_U e^{-\|x\|^2/2} d\mathbf{x}.$$

Let f be the random variable $f(\mathbf{x}) = \|x\|$. The function f measures the Euclidean distance of a randomly drawn point from the origin. The set

$$U = f^{-1}([0, r)) \subseteq \mathbf{R}^2$$

is the circle of radius r in $\mathbf{R}^2$. The probability that a randomly drawn point lies in this circle is

$$P(f < r) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_U e^{-\|x\|^2/2} d\mathbf{x}.$$

We can actually evaluate this integral in closed form by using polar coordinates. We obtain

$$P(f < r) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_{\theta=0}^{2\pi} \int_{\rho=0}^{r} e^{-\rho^2/2} \rho d\rho d\theta.$$

Since

$$\frac{d}{d\rho} e^{-\rho^2/2} = -\rho e^{-\rho^2/2}$$

we have

$$\begin{aligned} P(f < r) &= -\frac{1}{2\pi} \theta e^{-\rho^2/2} |_{\theta=0}^{2\pi} |_{\rho=0}^{r} \\ &= 1 - e^{-r^2/2} \end{aligned} \tag{3}$$

The probability density associated with this random variable is the derivative of $1 - e^{-r^2/2}$

$$P(f \in [a, b]) = \int_{r=a}^{b} r e^{-r^2/2} dr$$

as you can see by the fundamental theorem of calculus. This density is drawn in fig. 4 where you can see that the points are clustered at a distance of 1 from the origin.

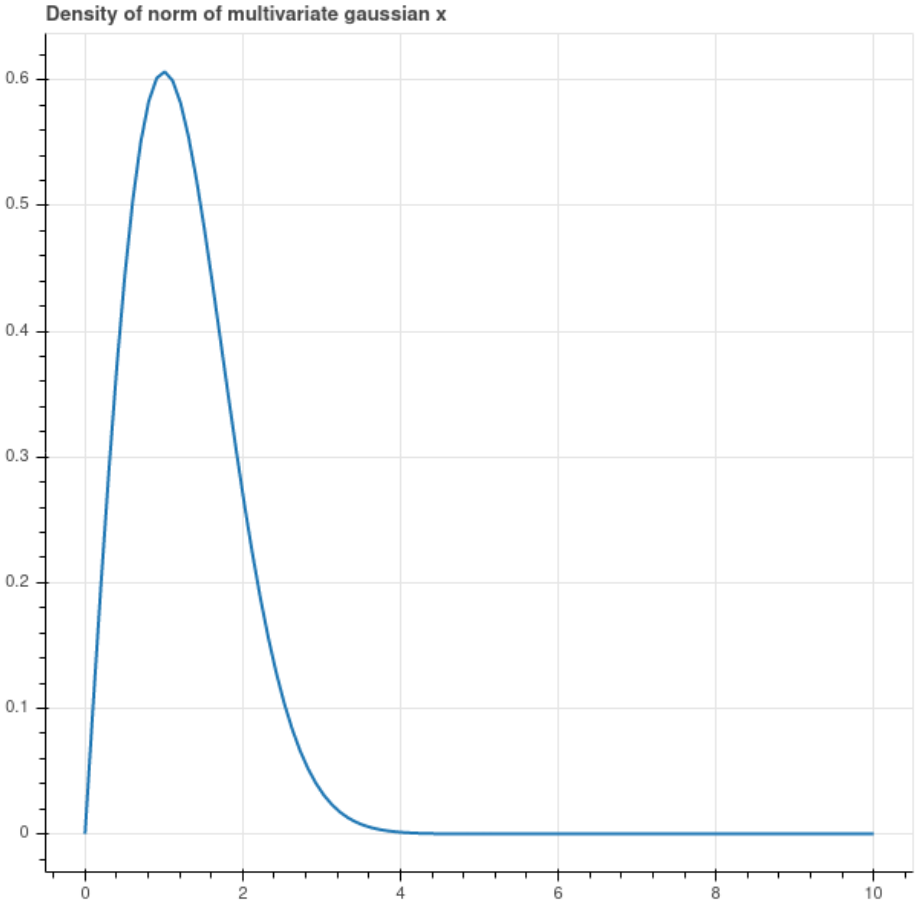

Figure 4: Density of the Norm

1.5.1 Models and Likelihood

A *statistical model* is a mathematical model that accounts for data via a process that incorporates random behavior in a structured way. We have seen several examples of such models in our discussion so far. For example, the Bernoulli process that describes the outcome of a series of coin flips as independent choices of heads or tails with probability p is a simple statistical model; our more complicated mixture model in which we choose one of two coins at random and then flip that is a more complicated model.
Our description of the variation in temperature measurements as arising from perturbations from the true temperature by a normally distributed amount is another example of a statistical model, this one involving a continuous random variable.

When we apply a mathematical model to understand data, we often have a variety of parameters in the model that we must adjust to get the model to best "fit" the observed data. For example, suppose that we observe the vibrations of a block attached to a spring. We know that the motion is governed by a second order linear differential equation, but the dynamics depend on the mass of the block, the spring constant, and the damping coefficient. By measuring the dynamics of the block over time, we can try to work backwards to figure out these parameters, after which we will be able to predict the block's motion into the future.

To see this process in a statistical setting, let's return to the simple example of a coin flip. The only parameter in our model is the probability p of getting heads on a particular flip. Suppose that we flip the coin 100 times and get 55 heads and 45 tails. What can we say about p?

We will approach this question via the "likelihood" function for our data. We ask: for a particular value of the parameter p, how likely is this outcome? From eq. 2 we have

$$P(55H, 45T) = \binom{100}{55} p^{55}(1-p)^{45}.$$

This function is plotted in fig. 5. As you can see from that plot, it is extremely unlikely that we would have gotten 55 heads if p was smaller than .4 or greater than .7, while the *most likely* value of p occurs at the maximum value of this function, and a little calculus tells us that this point is where $p = .55$. This *most likely* value of p is called the *maximum likelihood estimate* for the parameter p.

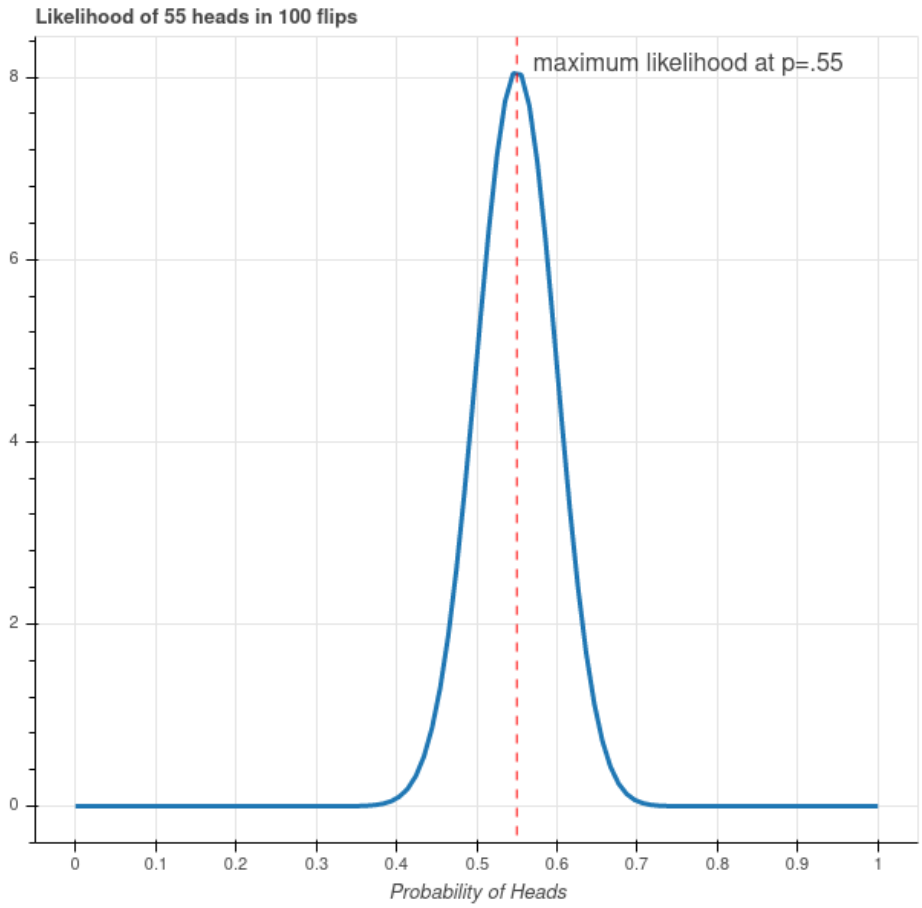

Figure 5: Likelihood Plot

1.6 Bayes Theorem

1.6.1 Conditional probability

1.6.2 Bayes Theorem

1.6.3 Bayesian inference

1.6.3.1 Priors

1.6.3.2 Data

1.6.3.3 Posterior

References

[1] Bertsekas, D. P. and Tsitsiklis, J. N. (2008). *Introduction to probability*. Athena Scientific.