Probability and Markov Matrices

Larry Susanka

February 7, 2024

Introduction

When reasoning about various events in the world, we see that some are precisely predictable or certain while others are not.

For instance if I flip a coin it is safe to assume that it will rise and then fall. Yes, the laws of gravity may be repealed mid-flight or the coin may evaporate in front of my eyes, but there is no need for me to think about those things. I will not waste my time contemplating such imaginary but impossible situations.

On the other hand, if the question is whether the coin will land "heads up" or "tails up" I am not certain at all. Taking into account all I know about coins and coin tossing, my "degree of belief" that it will land heads up is equal to my "degree of belief" that it will show tails. I am uncertain, but the nature of my uncertainty has a pattern.

This degree of belief represents my state of knowledge, not the true nature of the coin. It would be illogical for me to assess my degrees of belief differently without specific knowledge that the coin was rigged in some way to affect how it lands.

In fact, even if I know it has been rigged, but don't know which side is thereby favored, I still would have to assess heads and tails as equally likely to occur.

And it makes no difference if the "flipping" has already occurred. If I reach through a window into a closed room and flip the coin and slowly walk around to the entrance to inspect the coin, I still must assess the two possibilities as equally likely, even though it is certainly in one state or the other, up to the moment I see that coin on the floor.

Similarly, if I have an ordinary six-sided die and roll it, my degree of belief that one dot will show on top after it comes to a stop is equal to my degree of belief that five dots will show on top.

The subject of probability is the study of patterns in the degree of belief a logical person must hold, given all the information he or she possesses, about such events.

This is a huge subject and we will consider only one tiny corner of it here.

For more on the more general subject of Probability I recommend E. T. Jaynes insightful and entertaining *Probability Theory The Logic of Science*.

In the first 50 pages Jaynes makes a very convincing case, using elementary reasoning, that any attempt to measure "plausibility" using real numbers, that has qualitative correspondence with "common sense," and which is consistent **must** correspond to the standard rules of probability.

For more on the subject of Markov Chains in particular I recommend J. R. Norris' *Markov Chains*.

Probability

We specify a collection of **elementary outcomes,** also called **sample points** or **states**, related in some way to a physical situation about whose uncertain aspects we wish to reason. The set of these outcomes is called the **sample space** or **state space** and a sample space is frequently designated by the symbol Ω.

Subsets of the sample space are called **events**.

For instance in the first example our list of elementary outcomes might be H and T, with $\Omega = \{H, T\}$. There are just four events for this sample space, the empty set $\varnothing$ and $\{H\}$ and $\{T\}$ and $\Omega = \{H, T\}$.

In the second example the sample space could be $\Omega = \{1, 2, 3, 4, 5, 6\}$. This sample space yields many more events—64, actually.

We insist (in this talk) that the list of elementary outcomes should be finite. The list must include a description of everything that might occur. And we insist that the elementary outcomes be identified in such a way that no two of them could describe the same physical aspect whose plausibility we wish to discuss.

For instance for the die situation the list $1, 2, 3, 4, 5$ is not good because it neglects the possible 6 which might show up. Also, the list

$$1, 2, 3, 4, 5, \text{even number shows up}$$

is bad because if two dots were to show up two different outcomes would describe it. However $\Omega = \{\text{even}, 1, 3, 5\}$ doesn't have this problem.

The words **exhaustive** and **mutually exclusive** describe these two conditions, required of the members of our sample space.

Many different sample spaces could describe the same physical situation, though one might be more detailed, enabling us to answer questions with more specificity about the situation, or easier to use than another.

In the die-rolling situation, for instance, my degree of belief that any elementary outcome in $\{1,2,3,4,5,6\}$ will occur is the same. Using sample space $\{\text{even}, 1,3,5\}$ this is not true, and that might make some calculations harder or impossible.

Using sample space $\{\text{even}, \text{odd}\}$ restores this symmetry, and if all I care about is whether an even number of dots appears or an odd number this last sample space could be the best.

We decide to assign numerical values between 0 and 1 to describe our degrees of belief that an elementary outcome will occur, with 0 representing our assessment that an elementary outcome is impossible, while 1 corresponds to certainty that the elementary outcome will occur. We will call this number the **probability** of the elementary outcome.

The probability of an event is the sum of the probabilities of the elementary outcomes it contains.

If E is an event we use the notation $P(E)$ to designate the probability of event E. Thus $P(\varnothing) = 0$ (*something* must happen) while $P(\Omega) = 1$ (again, *something* must happen).

P is called a **probability function**, defined on the collective of all subsets of Ω.

For any two events E and F we have

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

To see this, think about the elementary outcomes that are "counted twice" in $P(E) + P(F)$.

So for **disjoint events**—that is if the two events cannot happen at the same time—probabilities are additive.

A sample space together with an assignment of probabilities to elementary outcomes is called a **probability model.**

Markov Matrices

We are going to expand on our ruminations about probabilities by assigning a probability model to each point in a finite set $\Omega = \{\omega_1, \ldots, \omega_n\}$, where the probability model for each ω_j has Ω as its sample space and probability function P_j.

We will interpret the number $P_j(\omega_i)$ as the probability that a "hopper" at state ω_j will hop to state ω_i. Thus each $P_j(\omega_i)$ is non-negative and

$$P_j(\omega_1) + \cdots + P_j(\omega_n) = 1.$$

Any vector $\boldsymbol{v}$ whose entries are non-negative and sum to 1 is called a **probability distribution vector**.

Form vector $\boldsymbol{p}_j$ whose entries are $p_j^i = P_j(\omega_i)$ for $i, j = 1, \ldots, n$.

Each $\boldsymbol{p}_j$ is the column whose ith entry is the probability of jumping from state ω_j to ω_i.

$\boldsymbol{p}_j$ is said to be the **probability distribution vector for probability function P_j.**

Create matrix $P = (\boldsymbol{p}_1 \, \boldsymbol{p}_2 \, \cdots \, \boldsymbol{p}_n)$.

A matrix of this kind is called a **Markov matrix.**

Markov matrices are, generally, those with non-negative entries whose columns add to 1.

(Other treatments may switch this, defining Markov matrices as having non-negative entries and whose rows add to 1 instead.)

$P\boldsymbol{e}_j = \boldsymbol{p}_j$, so left-multiplying $\boldsymbol{e}_j$ by P is a vector that represents the probabilities of a hopper finding itself at the various states having started, certainly, at state ω_j.

These hoppers need not be indivisible for this interpretation to make sense. $\boldsymbol{e}_j$ could represent 1 kilogram of microscopic hoppers at state ω_j. In that case $P\boldsymbol{e}_j$ would represent the distribution of hoppers after one jump, starting from state ω_j. The ith row of $P\boldsymbol{e}_j$ is the fraction (of the kilogram) of hoppers that end up in state ω_i if they all start at ω_j.

It could also represent, for instance, a gram of protein whose molecules , individually, jump from one geometrical configuration to another, or "reactant" molecules transforming to "products." The key feature here is that the probability of transforming to a different state should depend only on *where you are*, and not on *where you've been.*

Markov matrices have a number of interesting properties, and we assemble a few of these now.

For any Markov matrix P a nonzero vector $\boldsymbol{v}$ for which $P\boldsymbol{v} = \boldsymbol{v}$ is called a **stable** or, synonymously, a **stationary** vector. If stable $\boldsymbol{v}$ is a probability distribution vector it is called a **stationary or stable distribution.** These are eigenvectors for eigenvalue 1.

It is a standard fact from linear algebra that P and P^T have the same eigenvalues.

Let w be the column vector $(1, 1, \ldots, 1)$.

P is an $n \times n$ Markov matrix if and only if $P^T w = w$. So 1 is an eigenvalue for P^T and hence, also, P. Thus, every Markov matrix has a stable vector.

If P and Q are Markov matrices of the same size then PQ is a Markov matrix. And if P is invertible and Markov so is P^{-1}. Therefore the set of invertible Markov matrices is a group with matrix multiplication.

(It follows easily that the Markov matrices of a certain size form a Lie group.)

If $R_1, \ldots, R_n$ are the rows of a Markov matrix M then $R_1 + \cdots + R_n$ is a row whose entries are all 1. And if $v \in \mathbb{R}^n$ then

$$Mv = \begin{pmatrix} R_1 v \\ \vdots \\ R_n v \end{pmatrix}.$$

If the entries of any vector v add to c so too will the entries of Mv.
As a consequence, the entries of any eigenvector for eigenvalues *other* than 1 must add to 0: that is, these eigenvectors are orthogonal to $(1, \ldots, 1)$.

And if v is a probability distribution vector so is Mv.

Suppose v is a stable vector for Markov matrix M. v has at least one non-zero entry and we assume (replace v by $-v$, if necessary) that it has at least one *positive* entry.

Let a be the nonzero vector whose nonzero entries are the positive entries of v and define $b = v - a$. Thus $v = a - b = Ma - Mb$ and for each i at most one of a^i or b^i is nonzero and all entries of the four vectors on the right are non-negative.

Define

$$c = Ma \quad \text{and} \quad d = Mb.$$

The sum of the entries of c is the same as the sum of the entries of a, and the same is true for the pair b and d.

For convenience, permute the rows of these four vectors so that the k nonzero entries of a are first. Thus v is

$$\begin{pmatrix} a^1 \\ \vdots \\ a^k \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b^{k+1} \\ \vdots \\ b^n \end{pmatrix} = \begin{pmatrix} c^1 \\ \vdots \\ c^k \\ c^{k+1} \\ \vdots \\ c^n \end{pmatrix} - \begin{pmatrix} d^1 \\ \vdots \\ d^k \\ d^{k+1} \\ \vdots \\ d^n \end{pmatrix} = \begin{pmatrix} c^1 \\ \vdots \\ c^k \\ c^{k+1} \\ \vdots \\ c^n \end{pmatrix} - \begin{pmatrix} c^1 - a^1 \\ \vdots \\ c^k - a^k \\ b^{k+1} + c^{k+1} \\ \vdots \\ b^n + c^n \end{pmatrix}.$$

The entries of the vector on the far right are all non-negative and sum to $b^{k+1} + \cdots + b^n$. It follows that $c^j = 0$ for $j = k+1, \ldots, n$ and $c^j = a^j$ for $j = 1, \ldots, k$.

So $c = Pa = a$. Dividing a by the sum of its entries we conclude that every Markov matrix has a stable probability distribution.

If M is Markov so too is M^k for any positive integer k, so none of the entries of M^k, which are non-negative, can exceed 1. If $\boldsymbol{v}$ is an eigenvector for eigenvalue λ then $M^k\boldsymbol{v} = \lambda^k\boldsymbol{v}$.

From this, we conclude that $|\lambda| \leq 1$.

We are now finished accumulating general facts about Markov matrices.

Markov Chains

For some probability models, such as the act of coin-tossing or die-rolling or hopping from one state to another, it makes sense to imagine performing the procedure more than once and creating thereby more interesting compound events. In the case of coin-tossing or die-rolling the outcome of an earlier toss or roll has no effect on our thoughts about what will happen in later results.

Well, actually, maybe it does.

We interpret probabilities here as "degrees of belief" of a rational believer.

In fact, if I flip a coin assumed to be "fair" 10 times and it comes up heads each time my degree of belief that this is a "fair" coin would begin to erode and a fair-coin model would no longer match my degree of belief.

It would violate "common sense" to completely ignore strong evidence of this kind. Mathematicians should not be *obliged* to be idiots or fools, though in some cases it could help.

Further, if you flip a coin a million times it is no longer the coin you started with. The edges will become worn and other physical changes will likely occur.

In a class on statistics or probability you will learn how to use **Bayes' Theorem** to incorporate new information into a probability model "on the fly." This is an important topic.

We will, here, assume that the physical scenario is such that the probability model may be *assumed* to be fixed for the number of repetitions we care about.

So we have some accumulated ideas about plausible reasoning, a finite state space and facts about a Markov matrix P governing movement probabilities from any one of these states to the other states.

There is nothing to stop us from iterating this process.

$P^2\boldsymbol{e}_j$ is the jth column of P^2 and should be the distribution of hopper-mass among the various states after two hops, starting with 1 kilogram of microscopic hoppers in state ω_j. It also should represent the probability distribution vector for the location of a *single* hopper, having started at state ω_j, after two hops.

To justify and clarify these statements, consider the (k, j) entry of the matrix P^2. It consists of the sum

$$p_j^1 p_1^k + p_j^2 p_2^k + \cdots p_j^n p_n^k.$$

The first term in the sum is the fraction of 1 kilogram of hoppers that start in state ω_j and travel to state ω_1 multiplied by the fraction of those that subsequently go to state ω_k. This product is the mass of the original kilogram of hoppers that make it to ω_k via ω_1.

Any hopper that makes it to ω_k does so by passing through one of the n possible intermediary states, so the sum provides the total mass of hoppers that arrive at ω_k from their original location at ω_j by *any* intermediary.

We can use the previous calculation and an induction argument on the exponent to conclude that the (k, j) entry of the matrix P^L is the probability of starting at state ω_j and ending at state ω_k after L steps.

The behavior of the sequences of probability distributions $P^L \boldsymbol{v}$ for initial probability distribution $\boldsymbol{v}$ is of interest.

The sequence of matrices P^L and their implied effect on movement among the members of the state space is called a **Markov chain.**

Examples

Let's consider an example.

$$P = \begin{pmatrix} 0.5 & 0.25 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0.25 & 0.5 \end{pmatrix}.$$

Starting from any state there is a 50% chance of staying in that state. In the middle states there is a 25% chance of moving "left or right" but in the first and last states there is a 50% chance of moving to the single neighbor state.

$$P^2 = \begin{pmatrix} 0.3750 & 0.2500 & 0.0625 & 0 & 0 & 0 \\ 0.5000 & 0.4375 & 0.2500 & 0.0625 & 0 & 0 \\ 0.1250 & 0.2500 & 0.3750 & 0.2500 & 0.0625 & 0 \\ 0 & 0.0625 & 0.2500 & 0.3750 & 0.2500 & 0.1250 \\ 0 & 0 & 0.0625 & 0.2500 & 0.4375 & 0.5000 \\ 0 & 0 & 0 & 0.0625 & 0.2500 & 0.3750 \end{pmatrix}$$

$$P^{90} = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

So after two steps our kilogram of hoppers is spreading away, as expected, from their starting place but after 90 they have settled down with equal numbers of hoppers entering and leaving each state, with interior states having equal numbers and favored over end states.

If we modify our Markov matrix to

$$Q = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}$$

we have different behavior. A narrative for this Markov matrix might be that hoppers at interior states have an equal chance of moving to either neighbor, but if they land on an end state bug spray kills them and their carcasses pile up there.

$$Q^{90} = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.4 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.4 & 0.6 & 0.8 & 1 \end{pmatrix}$$

So if a kilogram of hoppers starts out at state ω_3 we would expect to find about 600 grams of dead bugs at state ω_1 and 400 grams at ω_6 after 90 hops.

And if the vector $\boldsymbol{v} = (1, 2, 3, 4, 5, 6)$ represents a distribution of 21 kilograms of bugs among the 6 states,

$$Q^{90}\boldsymbol{v} = (7, 0, 0, 0, 0, 14).$$

We will end up with 7 kilograms at state ω_1 and 14 kilograms at state ω_6.

Let's think of five states organized as beads on a wire loop, with possible movement to neighboring beads only. Consider Markov matrix

$$R = \begin{pmatrix} 0.2 & 0.2 & 0 & 0 & 0.6 \\ 0.6 & 0.2 & 0.2 & 0 & 0 \\ 0 & 0.6 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 & 0.2 \\ 0.2 & 0 & 0 & 0.6 & 0.2 \end{pmatrix}$$

The "hopper" narrative for R would be the following. Except for hoppers in state ω_3, hoppers have a 60% chance of moving one direction (say right) on the ring, 20% chance of "staying put" and 20% chance of moving left. But hoppers who find themselves in state ω_3 have an 80% chance of not moving and a 20% chance of moving left.

Exercise

(i) Starting with hoppers in state ω_3, movement governed by Markov matrix R, what fraction will be in state ω_5 after many hops?

(ii) Does the answer to (i) depend on where the hoppers started?

(iii) If a single person were to start in state ω_2 and make random movements governed by these probabilities, what is the probability that you will find that person in state ω_1 after 70 movements? After 71 hops? After $1,000$?

In the example P above we see that $\boldsymbol{v} = (1, 2, 2, 2, 2, 1)$ is a stable distribution, and each of the columns to which Q^L converges is stable. This is generally true. If P is any Markov matrix and *if* a column of P^L converges then the limit column is a stable distribution.

(Proof: Suppose for some i that $\boldsymbol{v}_k = P^k \boldsymbol{e}_i \to \boldsymbol{v}$. Then $P\boldsymbol{v}_k \to P\boldsymbol{v}$. But we also have $P\boldsymbol{v}_k = P^{k+1}\boldsymbol{e}_i \to \boldsymbol{v}$.)

However sometimes the columns of Q^L for a Markov matrix Q fail to converge, so stable distributions cannot always be found this way. For instance if

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{then} \quad Q^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{so} \quad Q^3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and the powers oscillate between the two forms—so a hopper known to be at one state will (certainly) hop to the other.

This kind of cyclic behavior can be interesting.

Exercise

Suppose $A = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}$ for $0 \le p \le 1$ and $0 \le q \le 1$.

(i) Consider the cases $p = q = 1$ and $p = q = 0$. Invent narratives to describe the behavior of particles changing their state governed by this matrix in each case.

(ii) Suppose $p \neq q$. Find eigenvalues and eigenvectors for matrix A. (hint: 1 is an eigenvalue and $(-1, 1)$ is an eigenvector for the *other* eigenvalue.)

(iii) Suppose we start out with one kilogram of "hoppers" distributed in the two states who hop according to Markov matrix A as in (ii). After many hops, how will the hoppers be distributed on the two states?

Exercise

Consider Markov matrix $S = \begin{pmatrix} 0 & 0 & 0 & 0 & 0.5 & 1 \\ 0 & 0 & 0 & 0.1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0.1 & 1 & 0.5 & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0 & 0 \\ 1 & 0.5 & 0 & 0 & 0 & 0 \end{pmatrix}$.

(i) Concoct a narrative describing what is happening on a ring with six beads.

(ii) Does S^L converge?

(iii) We know 1 is an eigenvalue of S. Find a stationary distribution.

(iv) Given various starting places, are there useful descriptions of the distributions of bugs among these six states after L hops for large L? What about the probability that a single hopper, starting at a specific state, will end up at one of the six?