Implementations of Probability Theory

Independent Study Report

Andrew Simonson

Compiled on: October 29, 2024

Contents

1 Objective

The educational focus of Implementations of Probability Theory surrounds the application of data models that produce non-deterministic insights through probabilistic methodology. By pursuing this study I hope to gain a deeper understanding of how to apply data in risk calculation for mitigation scenarios as they appear in real life, rather than the experimental lab conditions that enable algorithmic certainty.

In contrast to the path of black-box artificial intelligence and algorithms taught in **CSCI 335: Machine Learning**, this study is tailored to methods designed to produce confidence levels for uncertain events using certain terms, leveraging logical, traceable, and definite, calculations. Current course offerings in the realm of data science focus largely on the storing and management of data, and it is noted that the cluster of data science was until very recently under the branding of data management. Implementations of Probability Theory is intended to extend learnings in previous courses, notably **CSCI 420: Principles of Data Mining**, for more advanced algorithms used at the intersection of data and computing after the preprocessing stage.

After beginning this study the intended deliverable outline was determined to be technically implausible and has been replaced with demonstrations of applied algorithms. Taking inspiration from the retinal mosaic as displayed in **CSCI 431: Intro to Computer Vision** and discussion in **IGME 589: Computational Creativity and Algorithmic Art** on the appearance and nature of randomness in graphics, I hope to create a program that can determine the liklihood that randomly distributed colors on a hexagonal grid appear as they do in an image.

2 Units

2.1 Unit 1: Statistics Review

To ensure a strong statistical foundation for the future learnings in probabilistic models, the first objective was to create a document outlining and defining key topics that are prerequisites for probabilities in statistics or for understanding generic analytical models.

2.1.1 Random Variables

1. **Discrete Random Variables -** values are selected by chance from a countable (including countably infinite) list of distinct values
2. **Continuous Random Variables -** values are selected by chance with an uncountable number of values within its range

2.1.2 Sample Space

A sample space is the set of all possible outcomes of an instance. For a six-sided dice roll event, the die may land with 1 through 6 dots facing upwards, hence:

$$S = [1, 2, 3, 4, 5, 6] \quad \text{where } S \text{ is the sample space}$$

2.1.3 Probability Axioms

There are three probability axioms:

1. **Non-negativity**:

   $$P(A) \geq 0 \quad \text{for any event } A,\ P(A) \in \mathbb{R}$$

   No event can be less likely to occur than an impossible event ($P(A) = 0$). P(A) is a real number. Paired with axiom 2 we can also conclude that $P(A) \leq 1$.

2. **Normalization**:

   $$P(S) = 1 \quad \text{where } S \text{ is the sample space}$$

   Unit Measure - All event probabilities in a sample space add up to 1. In essence, there is a 100% chance that one of the events in the sample space will occur.

3. **Additivity**:

$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \cap B = \emptyset$$

A union between events that are mutually exclusive (events that cannot both happen for an instance) has a probability that is the sum of the associated event probabilities.

2.1.4 Expectations and Deviation

1. **Expectation -** The weighted average of the probabilities in the sample space

$$\sum^{S} P(A) * A = E \quad \text{where } E \text{ is the expected value}$$

2. **Variance -** The spread of possible values for a random variable, calculated as:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Where N is the population size, μ is the population average, and X is each value in the population.
For samples, variance is calculated with **Bessel's Correction**, which increases the variance to avoid overfitting the sample:

$$s^2 = \frac{\sum (X - \bar{x})^2}{n - 1}$$

3. **Standard Deviation -** The square root of the variance, giving a measure of the average distance of each data point from the mean in the same units as the data.

$$\sigma = \sqrt{V} \quad \text{where variance is } V$$

2.1.5 Probability Functions

Probability Functions map the likelihood of random variables to be a specific value.

Probability Mass Functions

Probability Mass Functions (PMFs) map discrete random variables. For example, a six-sided die roll creates a uniform random PMF. Each side of the die has a one-sixth chance of landing face-up, so the discrete chances of each x value between 1 and 6 is

represented by a $\frac{1}{6}$th portion of the sample space:

$$P(A) = \begin{cases} 1/6 & \text{if} \quad X = 1 \\ 1/6 & \quad\quad X = 2 \\ 1/6 & \quad\quad X = 3 \\ 1/6 & \quad\quad X = 4 \\ 1/6 & \quad\quad X = 5 \\ 1/6 & \quad\quad X = 6 \end{cases}$$

Probability Density Functions

Probability Density Functions (PDFs) map continuous random variables. For example, this is a PDF representing a vehicle's risk of being stranded as it travels (in a line at a fixed speed). The y value increases as the vehicle puts distance between itself and the starting point but, once the halfway point is reached, the risk decreases as the distance between the vehicle and the destination decreases.

$$P(A) = \begin{cases} X & \text{if} \quad 0 \leq X \leq .5 \\ -X + 1 & \quad\quad .5 < X \leq 1 \\ 0 & \quad\quad otherwise \end{cases}$$

2.1.6 Limit Theorems

Law of Large Numbers

The Law of Large Numbers states that as the number of independent random samples increases, the average of the samples' means will approach the true mean of the population.

$$\text{true average} \approx \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{as } n \to \infty$$

Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of a sample mean is a normal distribution even when the population distribution is not normal.

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{d} N(0,1)$$

Where X_i is the sample mean, $N(0,1)$ is a standard normal distribution, and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
This is a challenging to understand solely as an equation. As an example, take a sample of two six-sided dice rolls and average their numbers. The more sample

averages taken, the more they will resemble a normal distribution where the majority of samples average around 3.5.

2.1.7 Confidence

Confidence is described using a confidence interval, which is a range of values that the true value is expected to be in, and its associated confidence level, which is a probability (expressed as a percentage) that the true value is in the confidence interval.

It is important to note that confidence levels, such as 95%, do not indicate that the real value is within 5% of the point estimate. The confidence level expresses the probability that the real value is in the range provided by the confidence interval.

At the highest level, calculating confidence intervals is simply the observed statistic (generally the mean) plus or minus the standard error. The percentage is identified by applying the z-score coefficient (in the case of nornal distribution, other distributions use non-parametric methods) that corresponds to that level of confidence. For instance, the z-multiplier for a confidence level of 95% is 1.96 so a confidence interval formula around the mean would look like this:

$$\text{interval} = \mu \pm (1.96 * \text{SE})$$

To calculate standard error when the population standard deviation (σ) is known:

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

When σ is unknown:

$$\text{SE} = \frac{s}{\sqrt{n}}$$

where n is the size of the sample and s is the sample standard deviation. Notice how the standard error decreases with a larger sample size because it indicates a resilience in the sample to random events as per the Law of Large Numbers (2.1.6).

2.1.8 Statistical Inference

Statistical Inference is any data analysis to draw conclusions from a sample to make assertions about the population. Methods include estimation via averages and confidence intervals, and hypothesis testing, which attempts to invalidate (never *validate*) a hypothesis.

2.2 Unit 2: Probabilistic Theories and Epistemology

When developing probabilistic models it is vital to use domain expertise to expose the product to the full range of external variables that would be expected of a model applied to the real world. Without an appropriate understanding of both the limitations in research procedures and the true value of the data collected, the integrity of the model becomes inherently compromised.

As data scientists, we are uniquely at risk of falling for this trap because it is hard to fully grasp domain expertise when the nature of data science in a business setting frequently means consulting for many separate projects with a collectively massive scope. Of equal consideration, it is also easy to assume that the sophistication of our tools overrides imperfections in the data, in spite of mantras like 'Garbage In, Garbage Out'.

In this unit I explored some common fallacies and assumptions held by analysts who may not fully grasp the content that they work with, nor the problems they intend to solve. This required extensive research that I found was best digested in the form of books whose chapters chronicle multiple examples of a given principle. As such, the reading was not confined to just the timeslot designated for this unit. Research started during the months leading up to the start of the semester[1] and have continued through the independent study. This structure was particularly helpful to pull me back and gain perspective of what my goal was when I was knee-deep in feature construction and model formulation.

2.2.1 Moral Hazards and The Bob Rubin Trade

Picking pennies in front of a steamroller. When studying the effectiveness of a model the scope of review must capture the entire range of the sample space. Discarding black swans that don't impact the client does not mean the results will not reflect on the client for an oversight. There is therefore a question of obligation for data scientists to include flags for significant events in reality that do not effect the proposed course of action to the client.

The 2009 recession, attributed to the collapse of the housing market bubble, is the most common example of a moral hazard because the displacement of risk from banks who were federally required to give subprime loans to the taxpayer meant that banks could profit from subprime loans but would not be harmed when the inevitable occurred. In popular media, the housing bubble bursting is attributed to the banks where those in the industry passed off the event as something that nobody could have foreseen[2]. In reality, banks only ignored a probabilistic eventuality because

[1] Only research during the semester was logged in the timesheet

[2] For instance, in the 2015 movie *The Big Short*, only a few savvy traders who bothered to look into the details find that banks had, in their ignorance, built the bundled mortgages on an unstable foundation.

their models did not need to account for such an event.

Most emphasize the problems with risk transferrence when creating models. For this study's purposes, the important learning is that probabilistic models should not drop evaluations as soon as an event leaves the scope of the immediate client.

2.2.2 Ignoring Improbable Outliers with Outsized Impact

In machine learning it is common for algorithms to drop the most extreme (or a random selection of) datapoints to avoid overfitting and errors in data collection. One issue with the current implementation of this procedure is that it is often done blindly, ignorant of information that these outliers may relay. For instance, in a selection of 300 water samples from a stream, all but a few show a normal amount of oxygen in the stream. A citizen scientist may discount the remaining pockets as a statistical implausibility that is most likely indicative of a failure in sample testing and drop the most extreme 5% of datapoints. However, if these few pockets show a complete disruption of the dissolution process, the vast majority of aquatic life in the stream will eventually pass through these pockets without oxygen and die, resulting in an outsized impact from just a few sources.

Nassim Taleb in *Fooled By Randomness* describes this event with an analogy to Russian Roulette: If there was a 5/6 chance of winning a million dollars and a 1/6 chance of killing yourself, many people would at least hesitate before pulling the trigger. But what if the barrel is 10,000 rounds and it was only a 1/10,000 chance of harm? In this case, many less-than-rational actors use the game repeatedly to acquire wealth indefinitely, forgetting or even outright ignorant that eventually the unlikely, or, as the actor would see it, the unthinkable, happens and all of the gains are completely negated.

2.2.3 Fooled By Randomness

While most statisticians are familiar with techniques to remove noise to get a clearer picture of long-term trends, many forget that noise over longer terms can materialize as highly improbable events. For instance, it is improbable to flip a fair coin and have heads land face up 5 times in a row, but if the coin is flipped millions of times, it's exceedingly unlikely that a 5-head sequence does not occur.

In Nassim Taleb's namesake book, *Fooled By Randomness*, this concept is applied to ongoing timeseries analysis in stock markets. By accounting for the scope of the prior evidence, Taleb models the probability that daily events are the effect of noise, a number that remains high even in the face of multiple point swings in the market. Understanding this chance is critical because often observers attempt to justify random market events to events with high publicity that in reality had a negligible on the market, fooling investors out of acting on prices deviating from their

target.

2.2.4 Lindy Effect

The Lindy Effect describes the importance of historical evidence of continuity when estimating its continuity in the future. For items with a set lifespan, such as perishable goods, each passing day is indicative of a shorter remaining life expectancy, but the same is not true for nonperishables like tools and concepts. For example, consider the lifespan of a news story or hot book. Many such stories may take the world by storm, but then be nearly forgotten months later. However, older writings are incredibly unlikely to be forgotten in the next few months. It would be truly bizzare if everyone decided Shakespeare was not worth learning in the next few years because its value has been determined for so long to be high enough to maintain its popularity.

Applying this concept to probability theory, information and evidence that has been important for a long time is likely to stick around long after hot new examples or tactics that contradict it fade into obscurity. When measuring risk of startups, the concept and foundations may indeed be strong, but they have to be contrasted with the robustness of past ideas as proven over time. This concept also has applications for how people think about new things in their day to day life. In the news and papers outlining new developments, "Inaccurate science. . . is constantly being published. The Lindy-conscious consumer of scientific data will take seriously only information that has held up over a period of time"[3] because time has removed uncertainty associated with volatility of untested (or tested less than the alternative) information.

2.2.5 Decision Theory

Decision theory is the study of how people make decisions with uncertain information. There are two main branches of decision theory:

Normative/Rational Decision Theory

This branch studies how people *should* make decisions. In problems with other actors, as in game theory, it is assumed that all other actors will also act with perfect rationality, allowing for precise calculation of the actions of all of the others and their expected utility to the agent.

Descriptive Decision Theory

This branch studies how people actually make decisions which includes factors such as psychological and emotional biases. It applies subjective value measurements, frequently working in parallel with Dempster-Shafer Theory (2.2.7).

[3] `https://www.nytimes.com/2021/06/17/style/lindy.html`

2.2.6 Info Gap Decisions

In info gap decision theory there is not enough information to assign probabilities to events. The goal, then, is to select a course of action that is robust in the face of uncertainty. Where decision theory can predict expectations in irrationality to determine expected values, info gap decisions approximate the range of probabilities and weight them to estimate expected value. In essence, it applies probabilities to probabilities, adding an additional layer to insulate calculations from a lack of data or lack of understanding of a topic. Tying this into the Lindy Effect (2.2.4), we can compare the large range of probabilities of new, untested information with the narrower range from old, tested information which has experienced more challenges, just as confidence increases with a larger sample size.

2.2.7 Dempster-Shafer Theory

This section is an extra theory chosen to coincide with the unit 3 focus on Bayesian statistics. The Dempster-Shafer theory is a derivative application of Bayes Theorem (2.3.1) where subjective beliefs are applied to independent variables not tracked by the belief network. Shafer so eloquently describes this process by supposing that two friends, both of whom he subjectively believes are 90% reliable, tell him that a limb has fallen on his car [4]. Without observing Shafer's car we can calculate that there is only a 1% chance that both friends are unreliable, so there's a high liklihood that the statement is true.

However, if both friends are unreliable, they are not necessarily lying. Thus, there is actually less than 1% chance that a limb fell on the car. The exact probability can only be calculated by determining how likely it is that the friends would find it funny to tell Shafer that a limb fell on his car, contrasted with the odds that such a friend may also be willing to throw limbs at his car so as to maintain their ever-reliable facade. If one also considers the possibility that Shafer's friends mistakenly believed a limb fell on his car, this uncertainty must also be combined with the evidence for the most accurate picture.

2.2.8 Methodology Considerations

I have taken 10134023 instances of the last 40 years, during all of which Obama has been alive. Therefore I can say with a high degree of certainty that Obama is immortal.

An event never occurring in history does not discount its possiblity of occurring in the future. Similarly, events that may have been impossible in the past are not necessarily impossible in the future. Also, psychology. Someone who knows they are

[4] `http://glennshafer.com/assets/downloads/articles/article48.pdf`

being studied will act differently than someone who isn't being studied so models will be inaccurate.

2.3 Unit 3: Bayesian Statistics

This unit was deliberately separated from statistical review due to the percieved complexity of the topic and the magnitude of usage in recent data science breakthroughs. Bayes Theorem is a part of the cirriculum for both **MATH 351 - Probability and Statistics** and **CSCI 420 - Principles of Data Mining**. However, as both approached the topic from different perspectives and while neither solidified my personal confidence in its use, I chose to take extra time to learn this important topic in my own way.

It has been said that statistics does not come naturally to the human brain, hence statistics is, by mathematical standards, a young discipline. Resulting research on Bayesian statistics has led me to the conclusion that the opposite may be true - Bayes Theorem is quite intuitive, but its discipline has not had the time to crystallize best practices for instructing it. For instance, updating one's beliefs to compare probabilities with the number of documented occurrences is frequently used in philosophical discussion in the form of explanations that subsets with high liklihood of fufilling terms are valid classifications even when the subset size results in overall fufilled terms to be infrequently categorized as the proposed subset. Most people understand these expressions but, when shown a table and how to calculate those ratios, the content enters the realm of collegiate instruction.

2.3.1 Bayes Theorem

Bayes Theorem is a rule for conditional probability that calculates the probability of a cause given an event has occurred. The equation for Bayes Theorem is as follows:

$$P(A|E) = \frac{P(A) * P(E|A)}{P(A) * P(E|A) + (1 - P(A)) * P(E|\neg A)}$$

This formula appears more complex as it is. The denominator, while directly translating to "The probability of A times the probability of event E occuring given A divided by the probability of A times the probability of event E occuring in A plus the probability of not A times the probability of E occuring in not A" can be more easily expressed as $P(E)$ or the probability of event E occuring:

$$P(A|E) = \frac{P(A) * P(E|A)}{P(E)}$$

Finally, this equation is updated to replace descriptions with technical terms:

$$\text{Posterior Probability} = \frac{\text{prior} * \text{likelihood}}{\text{Evidence}}$$

By utilizing venacular more familiar to everyday life, Bayes Theorem can be translated as:

$$\text{P(occurence stems from A)} = \frac{\text{\# of occurences from A}}{\text{total \# of occurences}}$$

To appeal to mental visualization, the sample space can be imagined geometrically as a 1 unit by 1 unit square[5]. The area of this square, 1 unit squared, represents a probability of 1 (or 100%) and the probability of any possible outcome fits inside this square. Intuitively, this visualization can also be thought of as a confusion matrix where the squares are drawn proportional to their representative probabilities.

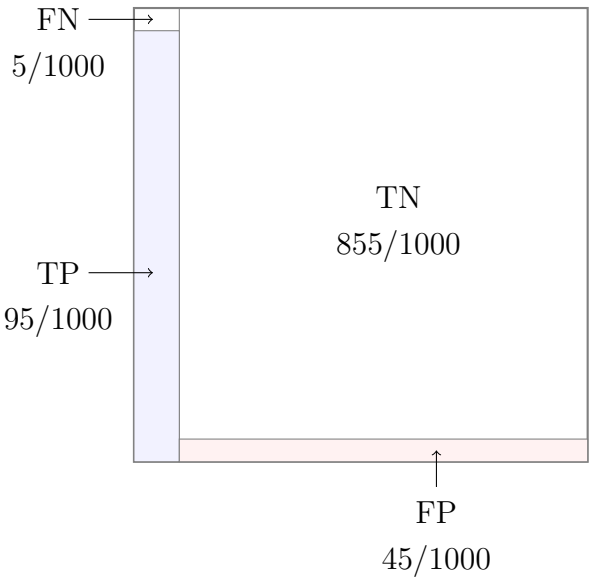
Consider an example where a patient wants to know if their positive cancer test is actually a false negative. Reviewing the test history, it's found to be accurate 95% across 1,000 uses. Given that we want to find the chances that a positive test is truly from a patient with cancer, let's highlight only the cases where a test is positive. A confusion matrix for this example would look like this:

	Cancer (100 patients)	No Cancer (900 patients)
Negative	False Negatives 5 patients	True Negatives 855 patients
Positive	True Positives 95 patients	False Positives 45 patients

Notice that the test does make the correct identification 95% of the time (and in this example, 95% regardless of actual value) but that there are almost half as many false positives as there are true positives, meaning having a positive test is not representative of a 95% chance of having cancer.

Proportinally scaling the probability matrix squares to create the sample space square defined earlier, we can see that the TP box appears to be approximately twice the size of the FP box. Logically, then, if we chose a random positive test, there's a two-thirds chance of the patient selected being from the true positive category:

[5]Concept credit to 3Blue1Brown on Youtube, this video is what finally clarified in my mind what the frankly simple equation behind Bayes Theorem meant.
`https://www.youtube.com/watch?v=HZGCoVF3YvM`



Bayes Theorem as applied to this problem can be simply expressed as:

$$P(\text{has cancer given positive test}) = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\frac{95}{1000}}{\frac{95}{1000} + \frac{45}{1000}} = 67.9\%$$

Meaning that, given a random positive test, there is a 67.9% chance of the patient actually having cancer, not far off from the two-thirds visual trick.

2.3.2 Bayesian Updating

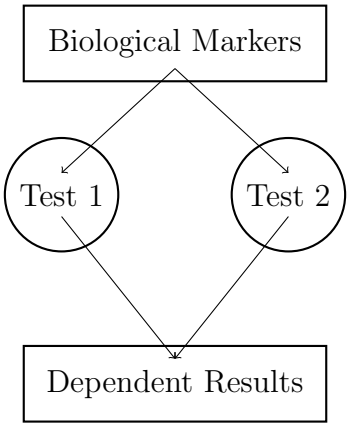
Bayesian Updating is another term that has been added to buzzword vocabulary to describe a process that isn't directly related to Bayesian Statistics but appears to have been rediscovered by academia through study of applied Bayes Theorem. In essence, Bayesian Updating simply states that observed occurrences should not override previous evidence and that it should instead be added to it in equal weight (equal value being a naive assumption). This evidence updating makes applications of Bayes Theory calculate posterior probabilities continuously as new information enters the system rather than a frequentist approach where the calculation only performed once.

2.3.3 Bayesian Belief Networks

Bayesian Belief Networks are probabilistic graphical models that preserve conditional dependence between random variables. In spite of its name, Bayesian Belief Networks do not necessarily apply Bayesian models, though they are a way to utilize Bayes Theorem for domains with greater complexity beyond a single posterior probability. In this type of network, edges are directed and the structure is utilized in a single direction. This is in contrast to undirected Hidden Markov Models (to be covered in

the next unit) that do not assume the order of aquisition of random variables. While it may not be practical to calculate the full conditional probability of a variable, Bayesian Belief Networks allow us to identify conditionally dependent variables that are weighted on the basis of an earlier random variable.

Following the example in the Bayes Theorem section of this report (2.3.1), let's suppose that a patient with a positive test takes a hypothetical second test. However, the second test's results are partially dependent on the first since they measure overlapping biological markers.



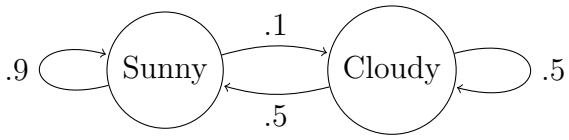
Test 1 Result	Test 2 Result	P(A)
Prior beliefs of test 1		
Unknown	Unknown	10%
Positive	Unknown	67.857%
Negative	Unknown	0.581%
Prior beliefs of test 2		
Unknown	Positive	55%
Unknown	Negative	1%
Dependent results from both tests		
Positive	Positive	75%
Positive	Negative	1.5%
Negative	Positive	0.6%
Negative	Negative	0.087%

Note that this probability of positive results in both tests (which both have greater than 50% of positives being true positives) is only equally certain as two positives from two independent tests each with 50% of positives being true. If the dependence was not included in the calculation and we ignored the fact that the tests partially measure the same thing, as would have occured in a Naive Bayes model, the tests' combined accuracy would be unjustly inflated.

2.4 Unit 4: Markov Methods

2.4.1 Markov Chains

Markov Chains are a form of probabilistic automaton where, the likelihood of transitioning to a new state depends solely on the current state, with no memory of prior states. For example[6], suppose a weather prediction program wants to know whether tomorrow will be a sunny or cloudy day, based on the current weather. Using the current weather as a state, the program identifies that there is a 10% chance of a sunny day transitioning into a cloudy day and a 50% chance that a cloudy day transitions into a sunny day:



Note that there is no information preserved between steps. Markov Chains are memoryless, so any information that must be available to them must be expressed as the state, such as the sunny and cloudy states in the example above. One benefit of such a straightforward structure is that it enables easy calculation of the probabilities of reaching a state k-steps from the current position. By expressing the chain as a transition matrix where rows represent the current state, the column represents the next state, and each cell contains the probability of the state moving from the column state to the row state, we get a 1-step transition matrix:

$$\begin{pmatrix} .9 & .1 \\ .5 & .5 \end{pmatrix}$$

or, expressed as a table:

Current State	Next: Sunny	Next: Cloudy
Sunny	90%	10%
Cloudy	50%	50%

To turn this into a k-steps transition matrix, this 1-step matrix only needs to be raised to the k-th power:

$$\begin{pmatrix} .9 & .1 \\ .5 & .5 \end{pmatrix}^k$$

To find the probability of the weather two days from the current state, plug 2 into k:

$$\begin{pmatrix} .9 & .1 \\ .5 & .5 \end{pmatrix}^2 = \begin{pmatrix} .86 & .14 \\ .7 & .3 \end{pmatrix}$$

[6]example sourced from:
`https://towardsdatascience.com/introduction-to-markov-chains-50da3645a50d`

From this matrix we can determine that if it is currently sunny, there is a 86% chance that it will be sunny in two days and, if it is currently cloudy, there is a 70% chance that it will be sunny in two days. As k approaches infinity, the model approaches its equilibrium where the starting state becomes irrelevant. In this example, any random day would be 83.333% likely to be sunny, representative of the long-term behavior of the system (climate), so the matrix of the equilibrium looks like this:

$$\begin{pmatrix} .9 & .1 \\ .5 & .5 \end{pmatrix}^{\infty} \approx \begin{pmatrix} .83333 & .16666 \\ .83333 & .16666 \end{pmatrix} \text{ OR: } \begin{pmatrix} .83333 \\ .16666 \end{pmatrix}$$

2.4.2 Hidden Markov Models

maybe add notes on mixed

2.5 Unit 5: Monte Carlo Simulations

what is this shit

2.5.1 How To Make a Monte Carlo Simulation

2.5.2 Monte Carlo Integration

2.5.3 Markov Chain Monte Carlo (MCMC) methods

3 Applied Projects

3.1 Randomness of Retinal Mosaic layout

hexagonal grid of marbles. are colors randomly distributed? Hexagonal basis vectors, retinal mosaic, entropy

3.2 Bayes Server Ripoff

I planned to create a trickle-down density belief network using probability density functions as nodes that choose the direction of rows in a relational database. Found this later, it's sort of similar. `https://www.bayesserver.com/`

Even better than their jank bayesian belief network I may be able to make mixed bayesian/markov chain models. This is a big project.

3.3 Cost-Benefit Analysis of Asychronous Education

This section covers a calculation I devised to make me feel better about my life decisions. The data is based on implicit guesswork and, while I will be taking it seriously for my decision to do either the online or on-campus RIT Data Science Masters Program, it should not be taken seriously as a probabilistic model. Since there is no framework for making a subjective decision weighting the potential benefits of on-campus life with the value of entering the workforce 18 months sooner, I decided to make one. Inshallah I shall reach my true potential and fulfill destiny.

with archaic knowledge imbued by Dr. Pepper flowing through my veins, I have selected $y = 3x^2 - 2y$ as the equation for covariance.