Module 5

Text mining, Neural Networks, Monte Carlo methods, Markov chains, classification, Market Basket Analysis.

Text mining

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

Text is a one of the most common data types within databases. Depending on the database, this data can be organized as:

- **Structured data:** This data is standardized into a tabular format with numerous rows and columns, making it easier to sto
- re and process for analysis and machine learning algorithms. Structured data can include inputs such as names, addresses, and phone numbers.

- **Unstructured data:** This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.

- **Semi-structured data:** As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn't have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.

Text mining tools and natural language processing (NLP) techniques, like information extraction allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality insights. This, in turn, improves the decision-making of organizations, leading to better business outcomes.

Text mining vs. text analysis

What's the difference between text mining and text analytics or text analysis? Well, the two terms are often used interchangeably, but they do have subtly different meanings.

Both text mining and text analysis describe several methods for extracting information from large quantities of human language. The two concepts are closely related and in practice, text data mining tools and text analysis tools often work together, resulting in a significant overlap in how people use the terms.

- **Text analytics** focuses on turning human language data into a structured format suitable for computers. It's the art of finding numerical data in text documents, such as frequency of word repetition or the presence or absence of themes in different documents. The first text analysis is said to have been carried out in the Middle Ages by French cardinal Hugh of Saint-Cher, who created an early version of a 'concordance' – a cross referencing of terms and concepts in the Bible.
- **Text mining** looks at patterns and trends in textual data, and produces insights that aren't apparent from just looking at the language itself. It works both by analysing what information is already there in the text, and by looking at metadata such as when documents were written and relationships between textual entities such as different web pages or online reviews.
- t can be used to identify semantic themes and even emotions around topics. For example, it might recognise frustration with customer experience or happiness about value for money. Text mining can be valuable in predicting what might happen in the future based on the trends in large volumes of written text over a period of time.

- **How is text mining different from using a search engine?**

- Search engines are powerful tools that make huge quantities of information available to us. However, the level of text analysis a search engine uses when crawling the web is basic compared to the way text analytics tools and text mining techniques work.

- Rather than looking for keywords and other signals of quality and relevance as search engines do, a text mining algorithm can parse and assess every word of a piece of content, often working in multiple languages. Text mining algorithms may also take into account semantic and syntactic features of language to draw conclusions about the topic, the author's emotions, and their intent in writing or speaking.

- **Text mining and text analysis in action**

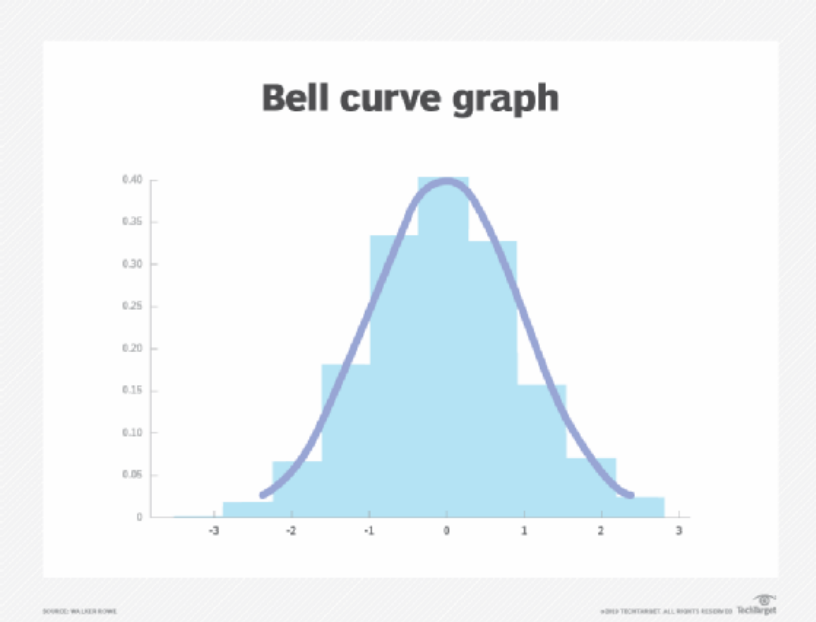
- So what are the applications of these technologies and what are some typical text mining tasks? Here are a few examples:

- Text mining allows a business to monitor how and when its products and brand are being talked about. Using sentiment analysis, the company can detect positive or negative emotion, intent and strength of feeling as expressed in different kinds of voice and text data. Then if certain criteria are met, automatically take action to benefit the customer relationship, e.g. by sending a promotion to help prevent customer churn.

Monte Carlo Simulation

A Monte Carlo simulation is a mathematical technique that simulates the range of possible outcomes for an uncertain event. These predictions are based on an estimated range of values instead of a fixed set of values and evolve randomly. Computers use Monte Carlo simulations to analyze data and predict a future outcome based on a course of action.

First, Monte Carlo simulations use a probability distribution for any variable that has inherent uncertainty. Then, it recalculates the results many times, using a different set of random numbers within the estimated range each time. This process generates many probable outcomes, which become more accurate as the number of inputs grows. In other words, the different outcomes form a normal distribution or bell curve, where the most common outcome is in the middle of the curve.



The Monte Carlo method has been described as "faking it a billion times until the reality emerges." It relies on the assumption that many random samples mimic patterns in the total population.

Importance of Monte Carlo simulations

Monte Carlo simulations are simple conceptually but enable users to solve problems in complex systems. They are particularly useful for long-term predictions because of their accuracy. Monte Carlo simulations are also a good alternative to machine learning when there isn't enough data to make a machine learning model accurate. As the number of inputs increases, so does the number of forecasts.

They also enable accurate simulations involving randomness. For a simple example, someone could use a Monte Carlo simulation to calculate the probability of a particular outcome -- say, rolling a seven -- when rolling two dice. There are 36 possible combinations, and six of those combinations add up to seven. The mathematical or expected probability of rolling a seven is 6/36, or 16.67%.

External factors, such as the shape of the dice or the surface they are rolled on, cause the actual or experimental probability to be different from the mathematical probability. Rolling the dice 1,000 times and getting a seven on 170 of those times

would be the actual probability -- 170/1,000, or 17%, which is close to the actual probability but not exact. Each roll would be an iteration of the Monte Carlo simulation, which gets more accurate with each iteration. This property -- that the actual probability gets closer to the exact probability with more iterations -- is known as the *law of large numbers*.

Someone could use Microsoft Excel, IBM SPSS Statistics or a similar program to run this experiment.

The 4 steps in a Monte Carlo simulation

Although they might vary from case to case, the general steps to a Monte Carlo simulation are as follows:

1. Build the model. Determine the mathematical model or transfer algorithm.
2. Choose the variables to simulate. Pick the variables, and determine an appropriate probability distribution for each random variable.
3. Run repeated simulations. Run the random variables through the mathematical model to perform many iterations of the simulation.
4. Aggregate the results, and determine the mean, standard deviation and variant to determine if the result is as expected. Visualize the results on a histogram

Monte Carlo simulation use cases

Monte Carlo simulations can be used for a spectrum of different industries. Finance is one of the most common use case examples, but any industry that involves predicting an inherently uncertain condition has a use for it.

Industry use cases for a Monte Carlo simulation include the following:

- **Finance**, such as risk assessment and long-term forecasting.
- **Project management**, such as estimating the duration or cost of a project.

- **Healthcare and biomedicine**, such as modeling the spread of diseases.

Use cases for Monte Carlo simulations also encompass different technologies. In IT alone, there a many use cases for Monte Carlo simulations. Some of those use cases specific to IT are the following:

- **Network and system design.** Monte Carlo simulations can be used to model different designs, identify potential bottlenecks, and perform capacity planning and resource allocation.
- **Artificial intelligence.** Monte Carlo simulations provide the basis for resampling techniques for estimating the accuracy of a model on a given data set.
- **Cybersecurity.** Monte Carlo simulations can be used to simulate different cyber attacks, evaluate the probability of them occurring, evaluate their hypothetical impact and identify vulnerabilities in IT systems.

Advantages of Monte Carlo simulations

Monte Carlo simulations are used in many different areas for a reason. They are a relatively simple way to make complex predictions. They offer answers to hypothetical questions and assign a certain level of order to randomness. Other advantages of Monte Carlo simulations include the following:

- **Improve decision-making.** Monte Carlo simulations help users make decisions with a degree of confidence.
- **Solve complex problems simply.** Monte Carlo simulations show both what could happen and how likely each outcome is
- **Visualize the range of possible outcomes and their likelihood of occurring.** Monte Carlo simulations make it easy to visualize what the result of a standard decision or outcome might be next to the result of an unusual outcome.

Drawbacks of Monte Carlo simulations

Despite the advantages of Monte Carlo simulations, there are disadvantages. Like any simulation, it uses historical data for a future projection, which carries the risk of being inaccurate. Specific drawbacks of Monte Carlo simulations include the following:

- **Processing power.** Monte Carlo simulations require many iterations to be accurate. Running many iterations takes time and energy and can be computationally intensive.
- **Input bias.** This can be damaging if the data used for input is inaccurate or incomplete.
- **Sensitivity to the chosen probability distribution.** It is important to choose an appropriate probability distribution for the problem. Choosing the wrong one can render the results meaningless.

Common probability distributions used in Monte Carlo simulations

Probability distributions represent a range of values between two limits and can consist of discrete or continuous values. Discrete probability distributions are plotted as a sequence of finite numbers in a table, whereas continuous distributions are plotted as a curve between two points on a graph.

Some common probability distributions in Monte Carlo simulations are the following:

- **Normal distributions.** These are continuous distributions where the most data points cluster toward the middle. It is also called a *bell curve* or *Gaussian distribution*.
- **Triangular distributions.** These are continuous distributions with fixed minimum and maximum values. They can be either symmetrical, where the most probable value equals the mean and the median, or asymmetrical.
- **Uniform distributions.** These are continuous distributions by known minimum and maximum values. All outcomes have the same probability of occurring.

- **Lognormal distributions.** These are continuous distributions by mean and standard deviation. The values are positive and create a curve that skews right.

Different probability distributions have different shapes and are suitable for different contexts.

- **Exponential distributions.** These continuous distributions are used to illustrate the time between independent occurrences given the occurrence rate.
- **Weibull distributions.** These continuous distributions can model skewed data and approximate other distributions.
- **Poisson distributions.** These discrete probability distributions describe the probability of an event occurring in X periods of time.
- **Discrete distributions.** These discrete probability distributions help define the finite values of all possible outcome values.

Monte Carlo simulations can be implemented in programming languages. To implement a few common distributions in Python, use the following code:

- Normal distribution: numpy.random.normal.
- Triangular distribution: numpy.random.triangular.
- Uniform distribution: numpy.random.uniform.
- Weibull distribution: numpy.random.weibull.

Markov Chain

WHAT IS A MARKOV CHAIN?

A Markov chain is a stochastic model that uses mathematics to predict the probability of a sequence of events occurring based on the most recent event. A common example of a Markov chain in action is the way Google predicts the next word in your sentence based on your previous entry within Gmail.

A Markov chain is a stochastic model created by Andrey Markov that outlines the probability associated with a sequence of events occurring based on the state in the previous event. It's a very common and easy to understand model that's frequently used in industries that deal with sequential data such as finance. Even Google's page rank algorithm, which determines what links to show first in its search engine, is a type of Markov chain. Through mathematics, this model uses our observations to predict an approximation of future events.

he main goal of the Markov process is to identify the probability of transitioning from one state to another. One of the primary appeals to Markov is that the future state of a stochastic variable is only dependent on its present state. An informal definition of a stochastic variable is described as a variable whose values depend on the outcomes of random occurrences.

What Are the Main Characteristics of a Markov Chain?

As stated above, a Markov process is a stochastic process which has memoryless characteristics. The term "memorylessness" in mathematics is a property of probability distributions. It generally refers to scenarios in which the time associated with a certain event occurring does not depend on how much time has already elapsed. In other words, when a model has a memoryless property, it

implies that the model has “forgotten” which state the system is in. Hence, previous states of the process would not influence the probabilities.

The main characteristic of a Markov process is this property of memorylessness. The predictions associated with a Markov process are conditional on its current state and are independent of past and future states.

This memorylessness attribute is both a blessing and a curse to the Markov model in application. Imagine a scenario in which you wish to predict words or sentences based on previously entered text —similar to how Google does for Gmail. The benefit of using the Markov process to do this is that the newly generated predictions would not be dependent on something you wrote paragraphs ago. However, the downside is that you won’t be able to predict text that’s based on context from a previous state of the model. This is a common problem in natural language processing (NLP) and an issue many models face.

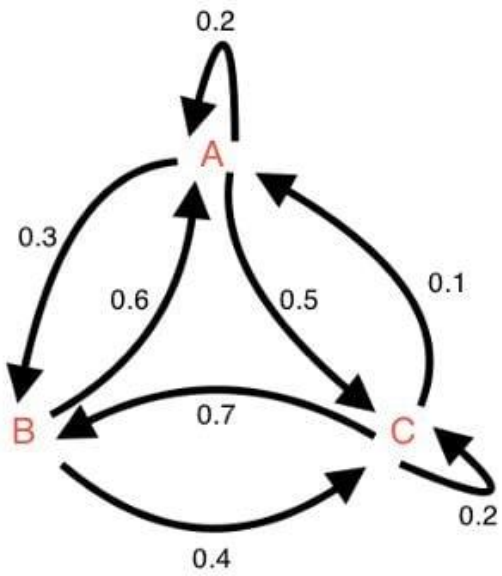
How to Create a Markov Chain Model

A Markov chain model is dependent on two key pieces of information — the transition matrix and initial state vector.

TRANSITION MATRIX

Denoted as “P,” This NxN matrix represents the probability distribution of the state’s transitions. The sum of probabilities in each row of the matrix will be one, implying that this is a stochastic matrix.

Note that a directed, connected graph can be converted into a transition matrix. Each element in the matrix would represent a probability weight associated to an edge connecting two nodes.



This graph outlines the probability associated with moving from one state to another. For example, there is a 60 percent chance to move from state B to state A. | Image: Vatsal Patel

Explain

```
    +-----+-----+-----+
    | A | B | C |   - Represents the network above
+-----+-----+-----+-----+   - NxN transition matrix
| A | .2 | .3 | .5 |   - element hold probabilities
+-----+-----+-----+-----+   - row sum of probabilities = 1
| B | .6 | 0 | .4 |   - .3 is the probability for state A
+-----+-----+-----+-----+     to go to state B
| C | .1 | .7 | .2 |   - .7 is the probability for state C
+-----+-----+-----+-----+     to go to state B
```

INITIAL STATE VECTOR

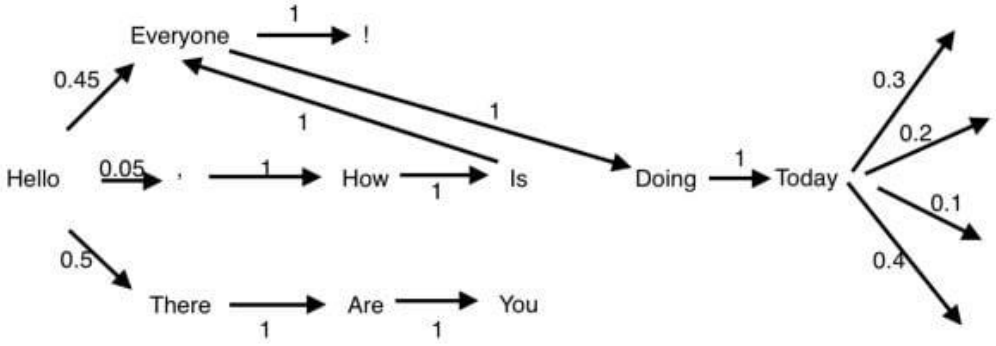
Denoted as “S,” this Nx1 vector represents the probability distribution of starting at each of the N possible states. Every element in the vector represents the probability of beginning at that state.

Given these two dependencies, you can take the product of P x Ito determine the initial state of the Markov chain. In order to predict the probability of future states occurring, you can raise your transition matrix P to the “M’th” power.

What Is an Example of a Markov Chain?

A common application of Markov chains in data science is text prediction. It’s an area of NLP that is commonly used in the tech industry by companies like Google, LinkedIn and Instagram. When you’re writing emails, Google predicts and suggests words or phrases to autocomplete your email. And when you receive messages on Instagram or LinkedIn, those apps suggest potential replies. These are the applications of a Markov chain we will explore. That said, the types of models these large scale companies use in production for these features are more complicated.

Suppose you had a large amount of text associated with a topic. You can imagine each sentence as a sequence of words in that corpus of text. Each word would then be its own state, and you would associate the probability of moving from one state to another based on the available words to which you are connected. This would allow you to transition from one state to another based on the probabilities associated with the transition matrix. This can be visualized below.



This sequence can be envisioned as a connected, directed network. | Image: Vatsal Patel

For visualization purposes, the network above has a small amount of words in its corpus. When dealing with large amounts of text like the Harry Potter series, this network would become very large and complicated. If you look at the beginning word “Hello,” there are three other potential words and symbols following it: Everyone , There with their associated probabilities. The simplest way to calculate these probabilities is by frequency of the word in the corpus.

Hello --> ['Everyone', ',', 'Everyone', 'Everyone', 'There', 'There', 'There', There', There', ...]

If there were 20 words in the list above, where each word is stated after the word “Hello” then the probability of each word occurring would follow the following formula:

Explain
P(word) = Frequency of Word / Total Number of Words in List
P(Everyone) = 9 / 20
P(,) = 1 / 20
P(There) = 10 / 20

Your initial state vector would be associated with the probability of all the words you could’ve started your sentence with. In the example above, since “Hello” is the only word to start with, the initial state vector would be a Nx1 vector with

100 percent of the probability associated with the word “Hello.” You can now predict the future states through this Markov model. I’ll show you how to implement this in Python next.

.**Market Basket Analysis**

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

Association Rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.

An example of Association Rules

- Assume there are 100 customers
- 10 of them bought milk, 8 bought butter and 6 bought both of them.
- bought milk => bought butter
- support = P(Milk & Butter)/100 = 6/100 = 0.06
- confidence = support/P(Butter) = 0.06/0.08 = 0.75
- lift = confidence/P(Milk) = 0.75/0.10 = 7.5

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

- **Types of Market Basket Analysis**
- **Market basket analysis is classified into two types:**
- **Predictive Market Basket Analysis**
- Predictive market basket analysis is a type of data mining technique that uses historical data on customer purchases to make predictions about future customer behavior. The goal of predictive market basket analysis is to identify items that are likely to be purchased together and use this information to inform business decisions such as product placement, marketing strategies, and inventory management.
- This type of analysis often involves using statistical and machine learning models to analyze the relationships between items, such as association rules and sequence analysis. The model is trained on historical data and can be used to make predictions about future purchases, such as suggesting items that a customer is likely to buy in the future or identifying products that are likely to be out of stock.
- Predictive market basket analysis is a valuable tool for retailers and other businesses that want to gain a deeper understanding of their customers and improve their operations.
- **Differential Market Basket Analysis**

- Differential Market Basket Analysis (DMBA) is a statistical technique used to identify the difference between two or more market baskets, or sets of items, typically purchased together by customers. It is commonly used in retail and marketing to understand the purchasing behavior of customers, as well as to identify trends and patterns in sales data. The goal of DMBA is to find items that are unique to each market basket and determine the associations between them, which can then be used to inform promotional strategies, product placement, and other marketing decisions.

- **Applications of Market Basket Analysis**
- Market basket analysis has several uses in various sectors, the most popular of which are:
- **The Retail Industry**
- Market basket research can assist retailers to find goods that are commonly purchased together, which can help them make product placement, marketing, and price decisions. This can result in greater revenue and better client satisfaction.
- **E-Commerce**
- Market basket analysis can be used by online merchants to evaluate client purchase data and discover which goods are often purchased together. This data may be utilized to develop targeted product bundles and upsell chances.
- **Healthcare**
- Market basket analysis can be used by healthcare organizations to evaluate patient data and find co-occurring illnesses or treatments. This data may be utilized to enhance patient outcomes while also lowering healthcare expenses.
- **Financial Services and Banking**
- Market basket analysis can be used by banks and financial organizations to evaluate client data and uncover trends in their purchasing habits. This data may be utilized to create customized marketing initiatives and boost consumer loyalty.
- **Telecommunications**

- Telecommunications firms can use market basket analysis to study consumer data and detect trends in their service consumption. This data may be utilized to enhance the customer experience and boost revenue.