Learning Module 11: Introduction to Big Data Techniques

LOS 11b: describe Big Data, artificial intelligence, and machine learning

Big data is a term that describes large, complex datasets. These datasets are analyzed with computers to uncover patterns and trends, particularly those related to human behavior. Big data includes **traditional sources** like company reports and government data and **non-traditional** sources like social media, sensors, electronic devices, and data generated as a byproduct of a company's operations.

Characteristics of Big Data

Volume: The amount of data collected in various forms, including files, records, tables, etc. Quantities of data reach almost incomprehensible proportions.

Velocity: The speed of data processing can be extremely high. In most cases, we deal with real-time data.

Variety: The number of types/formats of data. The data could be structured (e.g., SQL tables or CSV files), semi-structured (e.g., HT ML code), or unstructured (e.g., video messages).

Veracity: This is the trustworthiness and reliability of data sources. Veracity is crucial when using big data for making predictions or drawing conclusions. Big data makes it challenging to distinguish between data quality and quantity.

Types of Big Data

Big Data can be structured, unstructured, or semi-structured:

Structured data refers to information with a high degree of organization. Items can be organized in tables and stored in a database where each field represents the same type of information.

Unstructured data refers to information with a low degree of organization. Items are unorganized and cannot be presented in tabular form, such as text messages, tweets, emails, voice recordings, pictures, blogs, scanners, and sensors.

Semi-structured data may have the qualities of both structured and unstructured data.

Sources of Big Data

- **Financial markets**: Equity, swaps, futures, options, and other derivatives.
- **Businesses**: Financial statements, credit card purchases, and commercial transactions.
- **Governments**: Payroll, economic, trade, employment data, etc.
- **Individuals**: Product reviews, credit card purchases, social media posts, etc.
- **Sensors**: Shipping cargo information, traffic data, and satellite imagery.
- **The Internet of Things**: data generated by ‘smart’ buildings through fittings such as CCT V cameras, vehicles, home appliances, etc.

Professional investors, particularly quantitative ones, use alternative data sources in their financial analysis and decision-making. These sources significantly influence how they conduct their processes. They use alternative data to support data-driven investment models and decisions.

The following are the top three alternative data sources:

- **People-generated data**: This data is unstructured and is primarily accessed through website clicks and page visits
- **Commercial operations data**: This includes data on credit cards and corporate exhaust. It includes information from business transactions like point-of-sale records and banking activities. This data is typically structured.

- **Data produced by sensors**: This data is typically unstructured and is gathered through satellites, smartphones, cameras, RFID chips, and webcams.

Investment professionals must consider legal and ethical aspects when they use non-public information. Web data scraping can gather personal data that might be legally protected or disclosed without people's knowledge or consent.

Big Data Challenges

- **Quality**: Important questions include, but are not limited to: Does the dataset contain selection bias, missing data, or outliers?
- **Volume**: Is the quantity of data gathered adequate?
- **Appropriateness:** Is the dataset suitable for the chosen analysis method?

Experts have created artificial intelligence (AI) and machine learning methods to handle large and intricate alternative datasets. These technologies help in understanding and evaluating this vast and complex data.

Artificial Intelligence (AI) and Machine Learning (ML)

Artificial Intelligence

In broad terms, **artificial intelligence** refers to machines that can perform tasks in “intelligent” ways. It has much to do with developing computer systems that exhibit cognitive and decision-making abilities comparable to or superior to humans. It is the broader concept of machines being able to carry out tasks in a way that we would consider “smart.”

Early AI took the shape of expert systems, using "if-then" computer programming to mimic human knowledge and analysis. Neural networks, another early form, mimicked human brain functions in learning and processing information.

Machine Learning

Machine learning is a current application of AI that revolves around the idea that we should really just give machines access to data and let them learn by themselves without making any assumptions about the underlying probability distribution.

The idea is that when exposed to more data, machines can make changes on their own and come up with solutions to problems without reliance on human expertise – find and apply the pattern.

In the context of investment, machine learning requires big data for training. The growth of big data has enabled AI algorithms to improve modeling and predictive accuracy.

In machine learning (ML), a computer algorithm receives inputs, which can be datasets or variables, as well as outputs, representing the target data. The algorithm then learns how to model inputs into outputs or describe a data structure effectively. It learns by identifying data relationships and using this knowledge to enhance its learning process.

The ML divides the dataset into three unique types: a **training dataset**, a **validation dataset**, and a **test dataset**. A training dataset allows the algorithm to identify the link between inputs and outputs based on the historical pattern in the data. These relationships are then validated, and the model is adjusted using the validation dataset.

As the name suggests, the test dataset is used to test the model's strength in predicting well on the new data. Note that machine learning still needs human intervention to understand the underlying data and choose suitable techniques for data analysis. In other words, before data is utilized, it must be cleaned and free from bias and spurious data.

Causes of Errors in Machine Learning

Overfitting the Data

The model overfits the data when it discovers “false” associations or “unsubstantiated” patterns that cause prediction errors and wrong forecasts. In other words, overfitting happens when the ML model is overtrained on the data and considers the noise in the data as true parameters.

Underfitting the Data

Underfitting of data occurs when the model considers the true parameters as noise and is unable to identify the relationship within the training data. In other words, the model is too simple to recognize patterns in the data.

Black Box Problem

Machine learning models don't use explicit rules like traditional software. They learn from lots of data during training. This makes ML models, such as black boxes, sometimes give results that are hard to understand or describe.

Types of Machine Learning

Supervised Learning

Under supervised learning, computers learn to model data based on labeled training data containing inputs and the desired outputs. After “learning” how best to model the relationships for the labeled data, the algorithms are employed to predict the results for the new datasets.

Unsupervised Learning

In unsupervised learning, computers get input data without labels and have to describe it, often by grouping data points. They learn from unlabeled data and react based on commonalities. For example, grouping companies based on their financial, not geographical or industrial, characteristics is unsupervised learning.

Deep Learning

Deep learning involves computers using neural networks to process data in multiple stages, identifying complex patterns. It employs both supervised and unsupervised machine learning methods.

Question

Machine learning refers to one of the following:

A. Autonomous acquisition of knowledge through the use of computer programs.
B. Ability of machines to execute coded instructions.
C. Selective acquisition of knowledge through the use of computer programs.

Solution

The correct answer is A.

Machine learning means computers independently acquire knowledge through programs, enabling them to solve problems without human input. It's about computers learning and making decisions on their own.

LOS 11c: Describe applications of Big Data and Data Science to investment management

Data science is an interdisciplinary field that uses developments in computer science, statistics, and other fields to extract information from Big Data or data in general.

Data Processing Methods

Data analysts and scientists in big data analysis use different data management approaches. They consist of capture, curation, storage, search, and transfer.

- **Capture:** describes the method by which data is gathered and put into a form that may be used by the analytical process.
- **Curation:** By undertaking a data cleaning activity, data curation ensures the quality and accuracy of the data. Data inaccuracies are found in this procedure, and any missing data is made up for.
- **Storage:** Process of recording, archiving, and accessing data, as well as the fundamental structure of the underlying database:
- **Search:** Involves querying data to locate specific information. With big data, sophisticated techniques are necessary to efficiently retrieve the requested data content.
- **Transfer:** Describes the process of transferring data from the underlying data source or storage place to the underlying analytical instrument.

Data Visualization

Visualization encompasses data formatting, display, and summarization through graphical representations. For traditional structured data, tables, charts, and trends are commonly used, while non-traditional unstructured data demand innovative techniques like interactive three-

dimensional (3D) graphics, tag clouds, and mind maps.

Fintech is applied in investment management, including text analytics and natural language processing, risk assessment, and algorithmic trading.

Text Analytics and Natural Language

Text analytics employs computer programs to analyze and extract insights, primarily from unstructured text- or voice-based datasets like company filings, written reports, quarterly earnings calls, and social media content. Text analytics can be utilized in predictive analysis to identify potential indicators of future performance, such as consumer sentiment.

Natural language processing (NLP) is an area of study that involves creating computer programs to decipher and analyze human language. Essentially, NLP combines computer science, AI, and linguistics.

Translation, speech recognition, text mining, sentiment analysis, and topic analysis are examples of automated tasks that use NLP. Annual reports, call transcripts, news articles, social media posts, and other text- and audio-based data may all be analyzed using natural language processing (NLP), allowing NLP to discover trends more quickly and accurately than is humanly possible.

Using natural language processing data, earnings projections for a company's near-term prospects can be created. Twitter sentiments have also been used to gauge an initial public offering (IPO) success.

Python, R, and Excel VBA are frequently used programming languages, whereas SQL, SQLite, and NoSQL are prominent database systems.

Question

Which of the five data processing methods refers to the process of ensuring data quality and accuracy through a data cleaning exercise?

A. Data search
B. Data storage
C. Data curation

The correct answer is C.

Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise. It involves uncovering data errors and adjusting for missing data.

A is incorrect. Data search refers to how to query data. Big data requires advanced techniques to locate requested data content.

B is incorrect. Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design.

LOS 11a: describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data

Fintech refers to technological innovation in designing and delivering financial services and products. At its core, fintech has helped companies, business owners, and investment managers to manage their operations better thanks to specialized software and algorithms.

Note that the term fintech is commonly used to refer to companies that develop new technologies and their applications and also the business sector that encompasses such companies.

Fintech in Gathering and Analyzing Financial Data

Initially, financial innovation was limited to simple tasks such as data processing and automation of routine tasks. Today, fintech encompasses more advanced systems that can analyze information and make decisions based on machine-learning logic. Machines have been developed to “learn” how to perform tasks over time. Using such systems has brought about high levels of efficiency that surpass human capabilities. Fintech covers a broader range of services and applications. As such, services and applications of fintech relevant to the investment industry include:

1. **Analysis of Large Datasets**Apart from traditional data such as corporate financial statements and economic indicators, fintech development has helped integrate alternative data, such as social media, into investment decision-making.
2. **Analytical Tools**Artificial Intelligence (AI) can identify complex, non-linear relationships compared to traditional quantitative methods by enabling different data analysis techniques. Diverse approaches to data analysis are now possible because of advancements in AI-based methodologies. As an illustration, analysts use AI to sift through the vast volumes of data from corporate filings and annual reports to produce insights.

Question

A characteristic of fintech is that it is:

A. at its most advanced state, using systems that follow specified rules and instructions.

B. limited to simple tasks such as automating routine processes and data processing.

C. primarily driven by the increased availability of data and technological advancement.

The correct answer is C.

The availability of vast amounts of data and technological advancements have been the primary drivers of fintech's expansion. The rapid growth in data, including diverse types, large quantities, and improved quality, has provided valuable insights for financial institutions and fintech companies to develop data-driven solutions. Technological advancements, such as artificial intelligence and big data analytics, have made it possible to analyze and interpret this data effectively, creating innovative financial products and services.

A is incorrect. While this may be true for some aspects of fintech, it doesn't directly address the two most important reasons behind its growth - the rapid growth in data and technological advancements. The advanced state of fintech is more a result of leveraging data and technological innovations to develop sophisticated and efficient financial solutions.

B is incorrect. Fintech is not limited to simple tasks; it has expanded to encompass various complex financial activities. While it does automate routine processes and data processing, it does so through advanced technologies that enable the handling of massive amounts of data efficiently.