Data Mining Theoretical Questions

Q1: **What is the difference between *Data Processing* and *Data Mining*?**

Answer

- **Data Processing** is the process of *collecting* and *manipulating* the items of a dataset to produce meaningful information.
- **Data Mining** is a process of "mining" the data to uncover patterns. The term *data mining* is a misnomer (a wrong or inaccurate name or designation), because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction of data itself.

Source: en.wikipedia.org

Q2: **What is the difference between *Data Mining*, *Statistics*, *Machine Learning* and *AI*?**

Problem

What exactly do they have in common and where do they differ? If there is some kind of hierarchy between them, what would it be?

Answer

In short

- **Statistics** *studies* probability
- **Data Mining** *explains* patterns
- **Machine Learning** *predicts* with models
- **Artificial Intelligence** *behaves* and *reasons*

More detailed:

- **Statistics** is concerned with probabilistic models, specifically inference on these models using data.
- **Data Mining** is about using **Statistics** as well as other programming methods to find patterns hidden in the data so that you can explain some phenomenon. Data Mining builds intuition about what is really happening in some data and is still little more towards math than programming, but uses both.
- **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes. Math is the basis for many of the algorithms, but this is more towards programming.
- **Artificial Intelligence** uses models built by **Machine Learning** and other ways to *reason* about the world and give rise to intelligent *behavior* whether this is playing a game or driving a robotcar. Artificial Intelligence has some goal to achieve by predicting how actions will affect the model of the world and chooses the actions that will best achieve that goal. Very programming based.

Source: stats.stackexchange.com

Q3: **Briefly describe the process of *Data Mining***

Answer

The workflow of a typical data mining application contains the following steps:

1. **Data collection:** Data collection may require the use of specialized hardware such as a sensor network, manual labor such as the collection of user surveys, or software tools such as a Web document crawling engine to collect documents. After the *collection phase*, the data are often stored in a database, or, more generally, a *data warehouse* for processing.
2. **Feature extraction and data cleaning:** When data is collected, they are often not in a form that is suitable for processing. In many cases, different types of data may be arbitrarily mixed in a free-form document. To make the data suitable for processing, it is essential to transform them into a format that is friendly to data mining *algorithms*, such as multidimensional, time series, or semi-structured format. The *multi-dimensional* format is the most common one, in which different *fields* of the data correspond to the different measured properties that are referred to as *features*, *attributes*, or *dimensions*.
3. **Analytical processing and algorithms:** The final part of the mining process is to design effective analytical methods from the processed data. It may not be possible to directly use a standard data mining problem for the application at hand.

Source: Data Mining: The Textbook by Charu C. Aggarwal

Q4: **What are some different types of *Data Mining Techniques*?**

Answer

Data mining works by using various algorithms and techniques to turn large volumes of data into useful information. Here are some of the most common ones:

- **Association rules:** An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.
- **Neural networks:** Neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold), and an output. If that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. When the cost function is at or near zero, we can be confident in the model's accuracy to yield the correct answer.
- **Decision tree:** This data mining technique uses classification or regression methods to classify or predict potential outcomes based on a set of decisions. It uses a *tree-like* visual to represent the potential outcomes of these decisions.
- **K-nearest neighbor:** It is also known as the **KNN algorithm**. It is a non-parametric algorithm that classifies data points based on their proximity and association to other available data. This algorithm assumes that similar data points can be found near each other. As a result, it seeks to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category or average.

Source: www.ibm.com

Q5: **What are some applications for *Data Mining*?**

Answer

- **Financial Analysis:** The banking and finance industry relies on high-quality, reliable data. In loan markets, financial and user data can be used for a variety of purposes, like predicting loan payments and determining credit ratings. And data mining methods make such tasks more manageable. *Classification* techniques facilitate the separation of crucial factors that influence customers' banking decisions from irrelevant ones. Further, *multidimensional clustering techniques* allow the identification of customers with similar loan payment behaviors. Data analysis and mining can also help detect money laundering and other financial crimes.
- **Intrusion Detection:** Global connectivity in today's technology-driven economy has presented security challenges for network administration. Network resources can face threats and actions that intrude on their confidentiality or integrity. Therefore, detection of intrusion has emerged as a crucial data mining practice. I. It encompasses *association* and *correlation analysis*, *aggregation techniques*, visualization, and query tools, which can effectively detect any anomalies or deviations from normal behavior.
- **Energy Industry:** Big Data is available even in the energy sector nowadays, which points to the need for appropriate data mining techniques. Decision tree models and support vector machine learning are among the most popular approaches in the industry, providing feasible solutions for decision-making and management. Additionally, *data mining* can also achieve productive gains by predicting power outputs and the clearing price of electricity.

Source: www.upgrad.com

Q6: **How does *High Dimensionality* affect *Distance-Based Mining Applications*?**

Answer

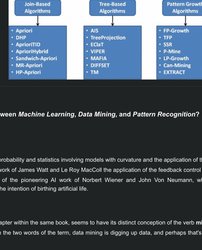
- Many *distance-based* data mining applications **lose their effectiveness** as the dimensionality of the data increases.
- For example, a distance-based clustering algorithm may group unrelated data points because the distance function may poorly reflect the *intrinsic semantic distances* between data points with increasing dimensionality.
- As a result, distance-based models of clustering, classification, and outlier detection are often qualitatively ineffective. This phenomenon is referred to as the *curse of dimensionality*.

Source: www.amazon.com

Q7: **What is *Frequent Pattern Mining*?**

Answer

- **Frequent pattern** is a pattern that appears frequently in a dataset.
- By identifying frequent patterns we can observe strongly correlated items together and easily identify similar characteristics, associations among them. By doing frequent pattern mining, it leads to further analysis like clustering, classification, and other data mining tasks.



Source: towardsdatascience.com

Q8: **What are the differences between *Machine Learning*, *Data Mining*, and *Pattern Recognition*?**

Answer

Machine Learning

- Some see it as a branch of applied probability and statistics involving models with curvature and the application of those principles in digital computing.
- Some see it as an extension of the work of James Watt and Le Roy MacColl the application of the feedback control concept to digital control.
- Some see it as the natural result of the pioneering AI work of Norbert Wiener and John Von Neumann, where the adaptive qualities of nature, including neurochemistry, are simulated with the intention of birthing artificial life.

Data Mining

- Each book, and sometimes each chapter within the same book, seems to have its distinct conception of the verb **mining**.
- Unfolding the metaphor contained in the two words of the term, data mining is digging up data, and perhaps that's a satisfactory definition for the most general use of the term.

Pattern Recognition

- The term pattern recognition is perhaps the most ambiguous because neither of the two words arose in a scientific context.
- Pattern recognition is the use of machine learning algorithms to identify patterns. It classifies data based on statistical information or knowledge gained from patterns and their representation.

Overlap and Associations

- When mining data, we may be looking for a particular kind of structure in a sea of data and have a particular search strategy to narrow the search and make it manageable for computing resources available. The test used during the search may be called pattern recognition.
- In machine learning, we may train a network of artificial cells to assist in locating data or features in data that are meaningful to the stakeholders in the project. That would be using ML for data mining projects.

- Although they share some similarities, it would be difficult to declare any two of the three to be synonymous. The three arose out of different kinds of research and from different orientations.

Source: ai.stackexchange.com

Q9: **How do you measure *Similarity* in Data Mining?**

Answer

- There is a measure of **similarity** in data mining and machine learning, and it falls under **Metric Learning**.
- It aims to automatically construct task-specific **distance metrics** from (weakly) supervised data, in a machine learning manner. The learned distance metric can then be used to perform various tasks. So,e distance metrics examples:
  - Euclidean distance
  - Mahalanobis distance
  - Pearson correlation
  - Cosine similarity
  - Jaccard similarity

Source: stackoverflow.com

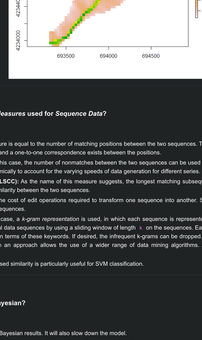
Q10: **What is the difference between *Rectangular Data* and *Non-Rectangular Data*?**

Answer

- **Rectangular Data** is the general term for a two-dimensional matrix with rows indicating records and columns indicating features. The data does not always start in this form: unstructured data must be processed and manipulated so that it can be represented as a set of features in the rectangular data.

	match_id	bowler	level_2	wickets	runs_conceded	runs_per_wicket
0	616	A Zampa	6	6	19	3.166667
1	608	AD Russell	6	6	25	4.166667
2	440	DJG Sammy	6	6	23	3.833333
3	83	Sohail Tanvir	6	6	15	2.500000
4	11310	A Joseph	6	6	18	3.000000
5	7932	TG Southee	5	5	33	6.600000
6	357	SP Narine	5	5	23	4.600000
7	119	A Kumble	5	5	6	1.200000
8	43	JD Unadkat	5	5	30	6.000000
9	90	CRD Fernando	5	5	18	3.600000

- **Non-rectangular Data** is a type of data that can not be represented with a table. *Time-series data* records successive measurements of the same variable. It is the raw material for *statistical forecasting methods*, and it is also a key component of the data produced by *devices*. *Spatial data structures*, which are used in mapping and location analytics, are more complex and varied than rectangular data structures. *Graph (or network) data* structures are used to represent physical, social, and abstract relationships.



Source: www.amazon.com

Q11: **What are some *Similarity Measures* used for *Sequence Data*?**

Answer

- **Match-based measure:** This measure is equal to the number of matching positions between the two sequences. This can be meaningfully computed only when the two sequences are of equal length, and a one-to-one correspondence exists between the positions.
- **Dynamic time warping (DTW):** In this case, the number of nonmatches between the two sequences can be used with dynamic time warping. The idea is to stretch and shrink the time dimension dynamically to account for the varying speeds of data generation for different series.
- **Longest common subsequence (LSCO):** As the name of this measure suggests, the longest matching subsequence between the two sequences is computed. This is then used to measure the similarity between the two sequences.
- **Edit distance:** This is defined as the cost of edit operations required to transform one sequence into another. Several alignment methods, such as BLAST, are specifically designed for biological sequences.
- **Keyword-based similarity:** In this case, a *k-gram representation* is used, in which each sequence is represented by a bag of segments of length k. These k-grams are extracted from the original data sequences by using a sliding window of length k on the sequences. Each such k-gram represents a new "keyword," and a 0-1df representation can be used in terms of these keywords. If desired, the infrequent k-grams can be dropped. Since the ordering of the segments is no longer used after the transformation, such an approach allows the use of a wider range of data mining algorithms. Any text mining algorithm can be used on this transformation.
- **Kernel-based similarity:** Kernel-based similarity is particularly useful for SVM classification.

Source: www.wiley.com

Q12: **Does *Noisy Data* benefit Bayesian?**

Answer

- Adding *noise* reduces the quality of Bayesian results. It will also slow down the model.
- Increases in *variability* do improve Bayesian methods if it identifies signal rather than noise.

Imagine a training set that only had green and brown-eyed individuals. How would it handle its first blue-eyed person outside the training set? By having a blue-eyed person in the data set, this increase in natural variability improves the degree to which the model matches reality. This will speed up processing speed. It will narrow the variability.

Source: stats.stackexchange.com

Q13: **How is *Computational Complexity* measured in *Ensemble Learning*?**

Answer

Computational complexity is the amount of CPU consumed by each inducer. It is convenient to differentiate between three metrics of computational complexity:

- **Computational complexity for generating a new classifier:** This is the most important metric, especially when there is a need to scale the data mining algorithm to massive datasets. Because most of the algorithms have computational complexity that is worse than linear in the number of tuples, mining massive datasets might be prohibitively expensive.
- **Computational complexity for updating a classifier:** Given new data, what is the computational complexity required for updating the current classifier such that the new classifier reflects the new data?
- **Computational complexity for classifying a new instance:** Generally, this type of metric is neglected, because it is relatively small. However, in certain methods (like k-nearest neighbors) or some real-time applications (like anti-missile applications), this type of metric can be critical.

Smaller ensembles require less memory to store their members. Moreover, small ensembles have a higher classification speed. This is particularly crucial in several near-real-time applications, such as worm detection.

Source: www.amazon.com