Diego F. Rincon Santana
2/15/2016

Homework #0

Hash Joins

Firstly, the purpose of a "join" is to combine records from multiple tables by using values common to each. A Hash Join accomplishes this by hashing one data set (table) into memory (using join attribute value for the hashing) and reading the other data set(s) looking at the hash table for matches; for this reason, hash joins require an operator that can compare values from one table with the other (equijoin).

For example, suppose you have the following tables

A

Age	**Name**
18	Trillian
20	Arthur
25	Ford
19	Zaphod
90	Marvin

B

Name	**Major**
Arthur	English
Arthur	Computer Science
Zaphod	Mathematics
Trillian	Chemistry
Zaphod	Philosophy

The first step is to hash table A and then probe table B for matches. The result is:

Age	**Name**	**Major**
18	Trillian	Chemistry
20	Arthur	English
20	Arthur	Computer Science
19	Zaphod	Mathematics
19	Zaphod	Philosophy

The main difference between Hash Joins and a Natural SQL Join is that the latter depends on foreign keys and the former on equijoins. As the name implies, the former also requires the use of a hash table.

"The computational cost of a Hash Join depends on the cost of building the hash table"

The above statement describes the operational value of a Hash Join. The cost of a Hash Join will be relatively low if the hash table can be held entirely in memory; this means that the total cost of the join is basically the cost of reading the two tables. If the hash table is too big, then the cost rises considerably. Therefore, the cost of the Join will depend greatly on the cost of building the hash table. For this reason, the algorithm for a Hash Join always hashes the smallest table.

Twitter mood swings

The causality correlation algorithm used in the paper roughly relies on the principle that if variable A causes B, then we should observe changes in A to occur before changes in B thus showing a statistically significant correlation between A and B. As it is noted in the paper, this doesn't prove causation; however the purpose is not to test actual causation but rather that

whether by looking closely at A we can predict (or not) what will happen to B. As an illustration, let A be the amount of ice creams sold at beach stands in Australia and B the number of shark attacks (also in Australia). Of course the rise of ice cream sales does not *cause* the number of shark attacks to rise; however, we could imagine that these two variables are correlated in the sense that if we see a sudden rise of ice cream sales at beach stands, we can predict that the number of shark attacks will also rise (surely because the rise of ice cream sales is a good indicator of how many people are at the beach).

Flu Detection by Google

The overall architecture of the analytics behind Google trends and its prediction to Flue Outbreaks was based on combining data from CDC related to flu (number of doctor visits due to flu-like symptoms, number of patients that test positive for influenza, etc…) with the data coming from Google search queries related to influenza. The assumption was that there is a correlation between the number search queries related to influenza (which tend to be popular exactly when flu season is happening) with an incoming flu outbreak. Its purpose was to be able to predict a flu outbreak almost a week into the future in the U.S.

The reason this approach failed was because this correlation was done in a vacuum and did not take into account the unreliability of the Google search queries data. Google tended to exaggerate its predictions because search queries related to influenza are not causally connected with the flu. For instance, a whole generation of medical students might search on google for flu symptoms because of an upcoming exam; there could be a conference related to influenza that makes the influenza-queries spike; or even more ironically, Google Flu Trends announcement itself could prompt people to look up on Google for influenza symptoms.

Conceptual and High-Level Architectural Questions

1)

1. Big Data

Big Data refers to data that satisfies three thresholds: that of volume, that of variety and that of velocity. For example, data that is coming in terms of Gigabytes every minute with no particular homogeneity can be called big data (e.g. tweets related to Superbowl during the Superbowl).

2. Predictive Analytics

Predictive Analytics is concerned with predicting future trends and behaviors. It uses techniques from machine learning and data mining (among others) to analyze current and historical data to infer future events or the effects of future events. For example, insurance companies use predictive analytics (using age, gender, driving record, ethnicity) to issue policies.

3. Analytic

The attitude of breaking a concept, idea or *thing* into smaller chunks in order to have a better understanding of it. For example, a chef might have an analytic attitude if, in order to have a better understanding of a dish, she decides to first break it down into all the ingredients that were used.

4. Data Science

Data Science refers to applying the scientific method towards data. In other words, it is the discipline that, by means of observing, questioning, hypothesizing (interdisciplinary approach), and testing is able to process and gain insights from data.

5. Data Mining

The purpose of data mining is to extract (insightful) information from big data. The methods used to do this involve artificial intelligence, statistics, machine learning and others.

6. Machine Learning

The goal of machine learning is to make a model infer the most adequate parameters of an algorithm based on data. In other words, Machine Learning uses training data to build a model that can make data-driven predictions or decisions. Its advantage over deterministic models is that the same program can dynamically change over time as data changes.

7. Statistics

The use of mathematical methods to interpret data; Statistics draws methods from Probability, Analysis, Differential Equations, and other fields of Mathematics to analyze and interpret data.

8. Business Intelligence

The ability to abstract the most fundamental aspects of the driving force behind a business; it encompasses researching data, analyzing data, interpreting data and synthesize it in order to be able to make informed, meaningful decisions.

9. Cross-validation

The method by means of which you use a data set for both training and testing; for example, 10-fold cross-validation divides the data set in 10 equal chunks and uses 9 for training and the remaining one for testing again and again until every chunk has been used as test.

10. Confusion Matrix

A Confusion Matrix is a table that presents the performance of a (usually) classification algorithm. The table is a good tool for visualizing how well the algorithm performs as it allows you to see the degree with which the system is properly labeling (or mislabeling) classes.

11. Unstructured Data

Unstructured data refers to information that doesn't have a homogenous format. An example of unstructured data would be a huge text document that contains numbers, quotes, dates, other languages, etc. The vast majority of raw data is unstructured data.

12. Structured Data

Structured data refers to information that is organized according to a specific model (e.g. tables, relational databases). This type of data is structured in fields and in well-defined formats; moreover it is usually the type of data that is used for data classification problems.

13. Semi-Structured Data

Similarly to Structured Data, Semi-Structured Data organizes data in fields and according to some type of hierarchy. However, it does not adhere to the usual models of Structured Data.

14. Data Clustering

Data Clustering refers to the technique of categorizing data in such a way to put together "similar" data. For instance, based on the data of all home sales in 2015, a clustering algorithm might be interested in categorizing those sales into different groups for analysis.

15. Stream Mining

Stream Mining refers to data mining of information that is not meant to be accessible for a long period of time. Examples of this type of information might be network traffic, web searches or sensor data.

16. Data Classification

Data Classification refers to the problem of assigning a label to a new piece of information based on historical data

17. Supervised Learning

Supervised Learning refers to the process of inferring a mapping from a set of training (labeled) data. Using this data set, supervised learning algorithms tries to infer relations for unseen observations. An example of Supervised Learning is Data Classification.

18. Correlations

A Correlation refers to the extent with which two variables have a relationship with each other. This, however, does not imply a causal dependence. Two variables may have a linear relationship with each other and still be causally independent.

19. Distributed Systems

A Distributed System is a system that seeks to increase its performance (speed, capacity, throughput, etc.) by pooling the resources of many servers. The goal of a distributed system is to have an *n*-times increase in performance by adding *n* machines to the system.

20. Unsupervised Learning

Contrary to Supervised Learning where you attempt to infer a function from a set of labeled data, Unsupervised Learning attempts to infer such function from unlabeled data. An example of Unsupervised Learning might be that of clustering.

21. Training Data

Training Data refers to the information used to build a predictive model. Training data, combined with Cross-Validation, can be a very powerful tool to build a successful model.

22. Test Data

Test Data refers to the information used to test a predictive model. Training data and Testing Data can be combined in Cross-Validation.

23. Recommender Systems

A Recommender system is a system that, based on a predictive model, is able to predict the ranking a user would give to an instance based on filtering, searching and machine learning techniques.

24. Trust Based Recommender Systems

A trust-based recommendation system, contrary to standard recommender systems, attempts to predict preferences by generating trust networks and aggregating rankings from such network. An example of such network could be a user's Facebook friends.

25. Biologically Inspired Data Mining

Biologically Inspired Data Mining refers to the task of extracting insights from data using known patterns in the biological sciences. An example of this is the use of bird flight-patterns to analyze political trends.

26. Knowledge Discovery

Knowledge Discovery is the broad field whose task is to extract insightful knowledge from data.

27. Class Label (in Data Classification)

The Class Label is that which the predictive model aims to predict. When looking at a new observation, the model will try to assign it a class label based on its training data.

28. Standard Deviation

The Standard Deviation of a set of numbers measures how spread-out the numbers are.

29. Variance

The Variance of a set of numbers is another way of measuring how spread-out the numbers are. The Standard Deviation is the square root of the variance.

30. ETL Jobs

ETL is the process of **E**xtracting data, **T**ransforming it into the proper structure of a database and **L**oading it into the database.

31. SQL

SQL (Structured Query Language) is a programming language designed to manage data in the context of a relational database.

32. NoSQL

NoSQL is a programming language designed to manage data that doesn't necessary fit into the relational database structure.

33. RDBMS

A relational database management system is a database system that is based on the relational model. You use SQL-type language to access and modify data.

Why is Data Science considered a Science?

Data Science is considered a science, as the name implies, because the way in which Data is analyzed and processed follows a specific methodology that is closely follows the scientific method which includes observing, hypothesizing and critically evaluating instances. Usually the *scientific* part of Data Science is ignored as people often simply follow a recipe instead of developing a critical approach towards data.

In many ways, all sciences are a type of data science. In Biology, Chemistry and Physics (among many), scientists often attempt to manipulate objects in order to extract useful data which can be transformed, evaluated and analyze in order to prove or disprove previously developed hypothesis; and this is indeed the business of data science.

2) Analytics Project Lifecycle

The meetings with the marketing team provided a good understanding of the business at hand. Now the data team needs to do the research into the kind of data that they will use to make a good analysis. The goal is to be able to classify a new customer after a few transactions so that the marketing team can choose the appropriate strategy for him/her.

Given that the business has collected labeled Historical Data of several customers, which amount to all their transactions, we can start by understanding what this structured data means. This part of the data analysis will involve diagnosing which data is useful and which one is not. For instance, age, gender, zip code will all be very valuable, but perhaps other data such as name, last name, email address, credit card info, and other such personal information can be reduced to one field in order to reduce the dimension of the data. We could use PCA for this task. Other problems that may occur would be those of missing data. This, however, for an e-commerce site may not prove to be such a big deal as you can force customers to fill in appropriate and meaningful data (e.g. cannot have empty fields) for vital data fields. Then again, if the data is unclean, we could always use an appropriate method to clean it, such as using mean values or most probable values (using, for example, decision trees).

Another very important part of understanding the data is not to think of it in a vacuum. The data team needs to do research into the type of real-life events that may have affected the sales. It may be that there was a weather catastrophe in a specific area that made certain customer's purchasing habits change for a significant period of time. Odd events like these may be hard to integrate into the data understanding process and may contribute into what chunk of the historical data we take into account for modeling (perhaps we can ignore all transactions that happened during the catastrophe).

Once we have selected the data that will go into the analysis. We must prepare it for modeling. Depending on which types of algorithms we will use, we could break the data into training and testing sets; or choose to do cross-validation. It is also important to have a clear idea of what algorithms we will consider and the pros-cons of each. It may be that naïve Bayes is the faster and cheapest algorithm but may prove to be somewhat irrelevant. Other more sophisticated methods such as ensemble algorithms may give us more accurate insights but may be more expensive to run. It is at this point that we should also revisit our hypothesis – that is to think about what exactly we mean by loyal customer or discount-driven customer or wandering customer. If the constraints are too tight, then we might fall into overfitting. If the constraints are too loose, we might fail to distinguish between categories. At this point we must also take into

account our budget and redefine the goals of the analysis based on the data at hand (if it turns out that the usable dataset is very small, we may have to include that fact into our analysis).

At this point we are in position to do the first pass at modeling. From the analysis describe above, the modeling may give us good, bad or incomplete results. It may have a very good accuracy (very unlikely at the beginning) or very poor accuracy or it may have some weird behavior. For instance, the algorithms may predict with very high accuracy if a customer is loyal or discount-driven, but may fail to distinguish from the others. In that case we must revisit our data and perhaps change the labels to make a clearer distinction between the types of customers. Even if the first algorithm proves to have a good performance, it may be very useful to run the rest of the algorithms in our pool of methods in order to have good insights not only about this dataset but also to learn more about their usefulness in specific circumstances. Perhaps we can then gain insights that will be useful in later data analysis projects. At this point it is very important to document all of our efforts. In the end, perhaps most of the team finds other jobs and new data scientists come in; it is better to have everything documented so that the handoff is efficient.

When evaluating the results of the modeling stage, it is important to compare these with our hypothesis. If need be, we should also review the definitions of the types of customers. The most important part of the final stage is to revisit the model with fresh data. This means that we should revisit the model once new data has come (perhaps wait 2 – 3 months). At this point we should review whether our predictions were truthful or incorrect. It may be that a new type of customer must be considered. At this point we need to maintain the model and tweak it (or overhaul it) if necessary. Model maintenance and persistence is a vital part of the analytics lifecycle.