Visual Analytics

Ryan Greenup

May 13, 2020

Contents

Tutorial

- PDF Version
- Office Document

Question 1

> What are visual analytics? And why we are Visual Analytics important in Big Data analytics?

Visual analytics is concerned with visual representations of data to facilite reasoning in an analytic fashion. Visual Analytics is usually implemented through a combination of automated analysis techniques relating to data mining, filtering out noise, compressing out unnecessary features and statistical alaysis.

Visual analytics are important in big data analytics because large amounts of data can be unwieldly and difficult to draw insights from without first processing the data. For example a data set with 1, 000, 000 observations over 1000 features will be difficult to interpret despite the fact that there is a sufficient amount of data.

Question 2

> What are the strengths of a Computer versus a Human?

Computers can rapidly process very large amounts of information with repeatable results and make high quality graphics (both static and live). Computers cannot however think, it's similar to choreagraphing a chain of dominoes, you can change the layout of the dominoes but you cannot expect the dominoes to alter there own layout.

Humans can solve problems that are difficult to break down into simple steps, simply visualise things that are difficult to compute (e.g. clustering), draw on past knowledge and invent new means to solve a problem.

Ideally the flexibility and inventiveness of a human should be paired with the computer's capacity to efficiently reach repeatable results.

Question 3

> Explore and research commercial visual analytics tools, such as Splunk, Datameer, Jaspersoft, Tableau, Karmasphere / Fico, Pentaho, etc. You are then required to provide a brief description about your 3 favorite platforms as well as their pros and cons.

Splunk

1. Description Splunk is a visualisation platform used for interpreting big data through a web-style interface, designed to allow access to visualisations across a variety of devices.
2. Pros
   - (a) 1. Platform Agnostic *Splunk* implements web technologies to make there software work on a variety of devices, the advantage to this is that you can use a mobile OS to browse visualisations, it can be far more immersive to browse visualisations on an *iPad* than a Desktop for instance.
   - (b) 2. Modern Interface The interface to use splunk is a modern web platform which makes for a good user experience.
   - (c) 3. Accessible *Splunk* is designed to allow people with out experience in programming to draw insights from visualisations, meaning that anybody can interpret data regardless of there background.
3. Cons
   - (a) Slow Interface The problem with using a web interface is that it will be slower and less responsive than a native program, this takes away from the user experience.
   - (b) Complex Architecture *Splunk* has a lot of features which means that Implementing it can be quite complex, potentially requiring on-going staff to keep it effective.
   - (c) Learning Curve Splunk has a high learning curve, difficult to get started with.

Tableau

1. Description *Tableau* is a commercial visualisation software which can turn data into visualisations through a *GUI*.

2. Pros

    (a) High Performance The unser interface is robust and stable and is renowned to operate fast even on big data.

    (b) Large User Base There is a large user base for Tableau which can make finding help easier.

    (c) Ease of Use *Tableau* has an easier learning curve than a programming language owing to it's *GUI* interface, this does however have the trade off that there is less room to grow.

3. Cons

    (a) High Cost and Training Tableau is commercial software and so it comes attached with a high cost, this means that if it is adopted any visualisations are tethered to the software and that training people to use it will also tether people to that software.

    This also means it can be difficult to find people with the training to produce high-quality Tableau Visualisations because individuals may not be willing to undertake training, as of there own volition, for a propriatery piece of software as opposed to something like *GGPlot*, *Plotly* or *D3* which could be used anywhere or any place of emplyment.

    (b) Security Issues Tableau is closed-source, this means that the community has had no opportunity to inspect the codebase for vulnerabilities.

    (c) Integration Issues It can be difficult to easily embed Tableau into a business's products because of it's proprietary nature, compare this to something like Shiny whcih in contrast can be simpler to implement.

Pentaho

1. Description

2. Pros

    (a) Many Features *Pentaho* has a vast array of features that users can implement.

    (b) Open Source Components of *Pentaho* are open source, this means that there is a higher capacity for software integration and for community review of software security.

    (c) Compatability *Pentaho* is compatable with many database sources including *MySQL* as opposed to for example *Tableau* which requires driver selectors.

3. Cons

    (a) Cryptic Error Messages Bugs encountered when using this software are very difficult to fix and the error messages are usually not descriptive of a pathway to rectify the issue.

    (b) Inconsistent Products in the *Pentaho* suite are implemented in an inconsistent fashion, making manipulating the software difficult and tiring.

    (c) Poor Documentation This platform assumes that users are familiar with scripting and the documentation assumes that, so this software can be very challenging to users not experienced with computer programming.

Question 4

> Explore and research open source packages and/or libraries. You are then required to provide a brief description about your 3 favorite platforms as well as their pros and cons.

Generally the main advantage of an Free Open Source library is that individuals can learn the software and implement the software with no fear of vendor lock in.

GGPlot

GGPlot is a graphing library that implements the grammar of graphics.

1. Pros

   (a) Layers The benefit of using ggplot2 (as opposed to for example `base-r` graphing libraries) is that plots can be built from the bottom-up by adding layers on top of each other.

   (b) Quality *GGPlot* Produces some of the nicest static plots of any Data Viz platform.

   - By merging many plots into a `gif` by using a `for` loop and `imagemagic` some degree of dynamic behaviour can be somewhat implemented.

2. Cons There are non really other than the learning curve. There is a steep learning curve, you have to learn how to use ***R*** and then how to coerce data into a tidy data format.

   The necessity to coerce data into a tidy data format isn't really an impediment though because it can actually make analysis easier generally.

Plotly

Plotly is an implementation of *D3* that is available in *Python* and ***R***.

1. Pros

   (a) Interactive Plotly allows for fully interactive plots to be created and embedded into HTML files via *JavaScript* (without actually needing to write any and just writing in *Python* and ***R***).

   (b) Complements ggplot2 *ggplot2* objects can be *passed* to plotly using the `plotly::ggplotly()` wrapper.

2. Cons

   (a) Documentation Documentation treats ***R*** users like second class citizens but it's still fairly useful.

   (b) Performance For large Data sets (like > 1000) the performance of this library is an issue. I ran into this problem when trying to analyse the *Marvel* dataset in my assignment, ideally it should be used on datasets of around 200 observations (this could also be an issue with the ***R*** memory management)

Shiny

Shiny is a platform to produce dynamic interactive HTML pages from inside ***R***.

1. Pros This package produces very nice pages and dashboards without needing to leave ***R*** the convenience of this is that analysis and publishing can be done all in one place.

   Combined with a package to produce static sites from within ***R*** like `bookdown` (or even the python package *MkDocs* in conjuction with `knitr`) this means that documentation and publication of statistical analysis can all be done in one place.

2. Cons The syntax is a little different from ordinary ***R*** because you need to specify variables in odd ways for shiny components, this can be difficult for a data scientist not experienced in web development or programming.

Question 5

> What is Predictive Analytics?

Predictive analytics are more concerned with statistics and data mining, the primary interest being trying to understand relationships between features and output variables.

Predictive Analytics are very dependent on the analysis and models implemented.

Question 6

> Explore and research predictive tools for big data, such as Revolution Analytics R, Spotfire Miner, Apache Mahout, RapidMiner, Oracle Data Miner, SAS Enterprise Miner, IBM SPSS Modeler etc. You are than required to provide a brief description about your 3 favorite platforms as well as their pros and cons (if have).

R

R is an open source

1. Pros The statistical packages implemented in ***R*** are often more rigorously implemented and this in part is one of the reasons that it is larger in academia than *Python*.

2. Cons

   (a) System ***R*** is inferior to python for system level scripting, this makes it less versatile and for this reason it is not used in spheres like *Astronomy* or even in industry generally.

   (b) Neural Networks *Google* writes all there machine learning in *Python*, so, if you want to use *Neural Networks* you'll have to use python anyway.

   (c) Syntax ***R*** is very *lispy* and nowhere nearly as readable as *Python* (although I like {bracket[s]...})

Apache Mahout

Apache Mahout is linear alebra framework to allow data scientists and mathematicians to rapidly implement algorithms.

1. Pros Mahout is efficient in that it optimizes operations given to it into a mathematically equivalent although computationally easier methodology.

2. Cons Mahout isn't well documented and because of this many people implement *MapReduce* jobs until the size of the data set has been reduced enough for working in ***R***.

Oracle Data Miner

The oracle Data Miner allows people to directly interact with data inside a database in an interactive fashion.

1. Pros This Allows an easy representation for data that must be *trapped* in a database for performance concerns to be analysed in a relatively painless way.

   As an example performing text mining in the context of say a *GitHub* wiki is very simple, clone the wiki and use `grep` or `recoll`, but trying to analyse *Wikipedia* is quite difficult because `mediawiki` uses a *MySQL* database. This tool allows an easy methodology to take advantage of the structured database to draw insites.

2. Cons

   - Requires an Oracle Dev Account to Download
   - Only Provides an `rpm` package for linux
     - This is a nuisance because `deb` is very common and an `AppImage` or `snap` would be cross platform, this means that users of *debian* have to either use `alien` to get the right format or install various ***C*** libraries in order to compile from source.