Adrien Xie 451852 xie.z

Problem 1 Big Data Characteristics

(a) Access Log Data
1. Amount – huge: the webserver constantly logs the all the records over the duration of five years.
2. Size – small: each data point is rather small, each on is a short patterned string
3. Infinity – yes: as long as the webpage is on, each new visitors is creates a new log
4. Structure – structured/semi-structured: the data point is a short patterned string
5. Complexity – simple: data records are independent of each other

(b) Wikipedia articles
1. Amount – huge: there are millions of articles on Wikipedia
2. Size – big: many single articles contains thousands of words and links
3. Infinity – no: new terminology created or updates occur less than often
4. Structure – unstructured: most articles are long unstructured text
5. Complexity – complex: there are links and dependencies between data records

(c) Chemical compounds
1. Amount – huge: there are millions of compounds recorded
2. Size – small: most chemical compounds is simple and therefore small.
3. Infinity – no: each day there are around 10.000 new compounds being added to the dataset; however they can be processed.
4. Structure – structured: the records is stored as bonds and its type, atoms and its types and compound property
5. Complexity – simple: data records are independent of each other

(d) (a) and (b) data set difference
1. Each record in the access log dataset is small and semi-structured (patterned string) while each record in the Wikipedia article dataset is large and unstructured.
2. The access log dataset is ever growing – it is infinity – while the Wikipedia article dataset remains rather stable in dataset size.

Implication on data analysis
1. It is easier to analyze access log data because the data points are semi-structured and requires little data preprocessing. On the contrary, the Wikipedia article data requires different data presentation based depending on the problem – therefore it takes much more efforts to preprocess the article data.

(e) One data point in access log data is a visiting record containing time, visiting IP address, visiting file, file transfer time and a number at the end. It is a semi-structured string that can be parsed into a 5-tuple table for programmatic convenience.

One data point in Wikipedia articles data is an article containing headings and contents and reference links. It is an unstructured string. For programmatic representation, we can use word frequency to represent the documents content to perform semantic analysis.

One data point in chemical compound data is an entry that contains atoms, bonds and property. They are stored as integer pairs. It is structured and can be presented as four tuple entry in the table and the first three tuple is vector of integer pairs and the last tuple is a category (0 or 1) feature.

Problem 2 Bonferroni’s Principle

The probability of p people decide to visit a hotel on any given day is $0.01^p = 10^{-2p}$. The chance that they will visit the same hotel is this probability divided by $1 \times 10^{5(p-1)}$. The chance that p people will visit the same hotel on d different days is the probability 10^{-7p+5} to the power of $d - 1$, $10^{-7pd+5d}$.

The number of groups of people is $\binom{10^9}{p}$ and the number of day periods is $\binom{1000}{d}$.

The expected number would be $\binom{10^9}{p}\binom{1000}{d} 10^{-7dp+d} = \frac{(10^9)^p(10^3)^d}{p!d!} 10^{-7dp+5d} = \frac{10^{9p-7pd+8d}}{d!p!}$

Problem 3 the Unreasonable Effectiveness of Data

(a) Unlabeled data is ubiquitous – there is more unlabeled data (in the wild) than the labeled data.

The price of (obtaining) annotating data can be very expensive and time-consuming. The examples listed in the article such as document classification, part-of-speech tagging, named-entity recognition, and parsing also face the problem of possible ambiguity and difficulty for the experts to agree on.

(b) In data-based approach, usually three steps are required – choose a representation, encode a mode and perform inference. In the natural language processing, a good policy for speech recognition is to use memorization if there is sufficient training data to build elaborate probabilistic models (in today's translation models, general rules are only introduced when they improve translation over mere memorizing particular phrases). In many cases there appears to be a threshold above which the data is considered sufficient.

(c) There is a threshold that sets whether the data amount is sufficient. The data-approach works poorly when the number of data points is below the threshold. The main difficulty is that how to represent the features of the raw data and how skilled a researcher has to be in order to be able to extract the information out of the raw data. Last but not least, some raw data may be inaccurate and yet they cannot be simply excluded.