Kenneth Allen

Jeffrey Young

Machine Learning

October 11, 2017

**A Bayesian Approach to News Headline Classification**

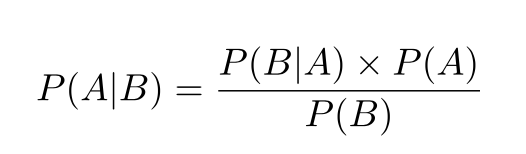
**Introduction:**

News article headlines have traditionally been a source of attraction without providing much content. Traditionally, the information value of headlines has been considered miniscule at best. This research effort tests these preconceived notions of headline information quality as well as the usability of minable headline data. The efforts of this study began with attaining a set of approximately one million headlines taken from the Australian Broadcasting Corporation. The data set was processed using a naïve Bayes approach to produce both test and training sets. The training set was used to make predictions of weekend, month, season, year, election time, first chronological half of the data set, prime minister’s affiliation, originating publication, and GDP growth. The study has concluded that quality information can be extracted and features can be predicted with an accuracy well above that of chance.

**Background:**

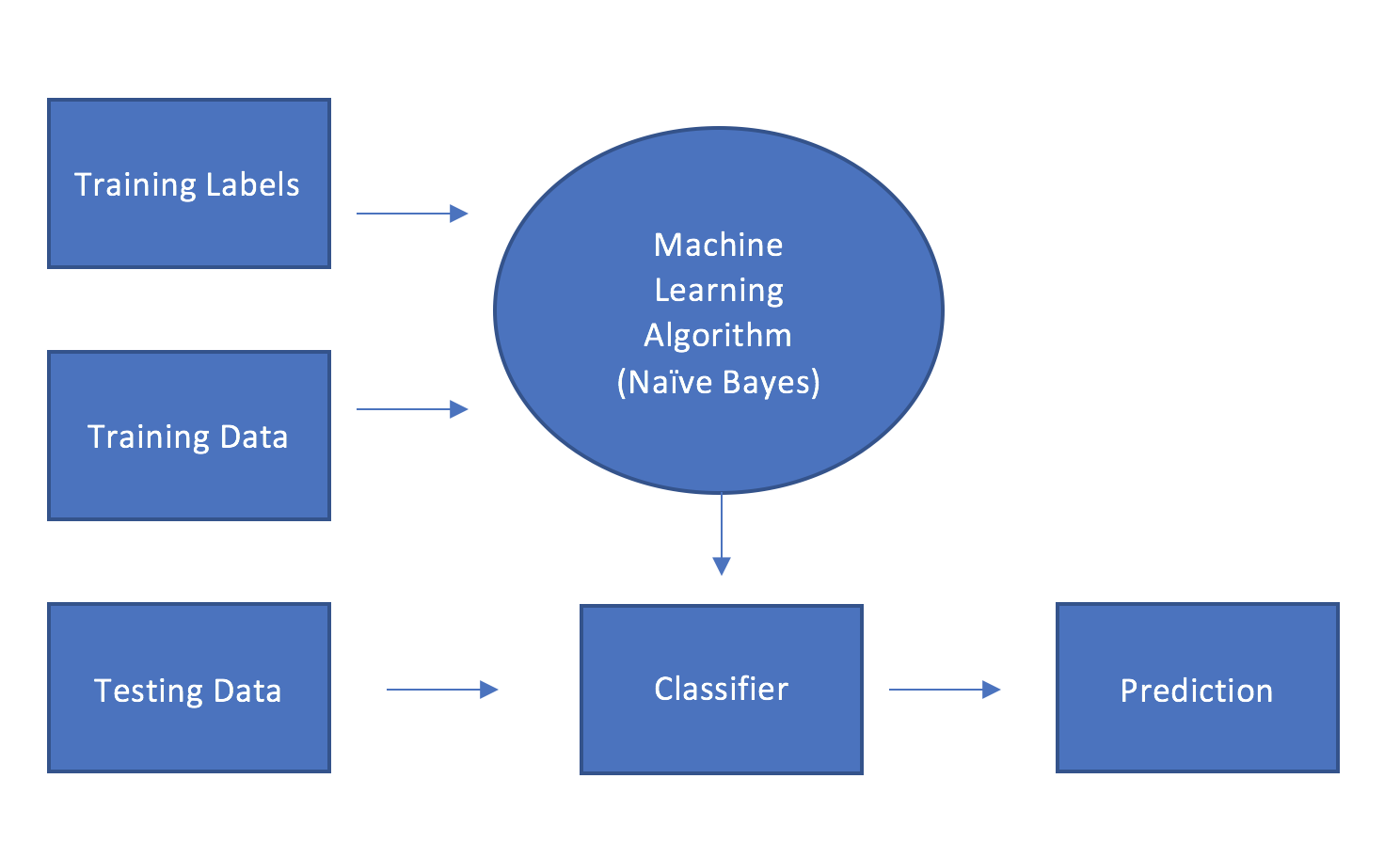
**Bayes' Theorem** is a simple mathematical formula used for calculating conditional probabilities. It figures prominently in *subjectivist* or *Bayesian* approaches to epistemology, statistics, and inductive logic. Subjectivists maintain that rational belief is governed by the laws of probability and lean heavily on conditional probabilities in their theories of evidence and their models of empirical learning. Bayes' Theorem is central to these initiatives both because it simplifies the calculation of conditional probabilities and because it clarifies significant features of subjectivist position. Indeed, the Theorem's central insight that a hypothesis is confirmed by any body of data that its truth renders probable is the cornerstone of all subjectivist methodology. [Joyce]

**Bayes' Theorem**:



**Naive Bayes** is a family of algorithms that take advantage of probability theory and Bayes’ Theorem to predict the category of a sample (like a piece of news or a customer review). They are probabilistic, which means that they calculate the probability of membership in each category for a given sample, and then output the category with the highest one. The way they get these probabilities is by using Bayes’ Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature. [article]

**Data Processing:** (insert block diagram)

****

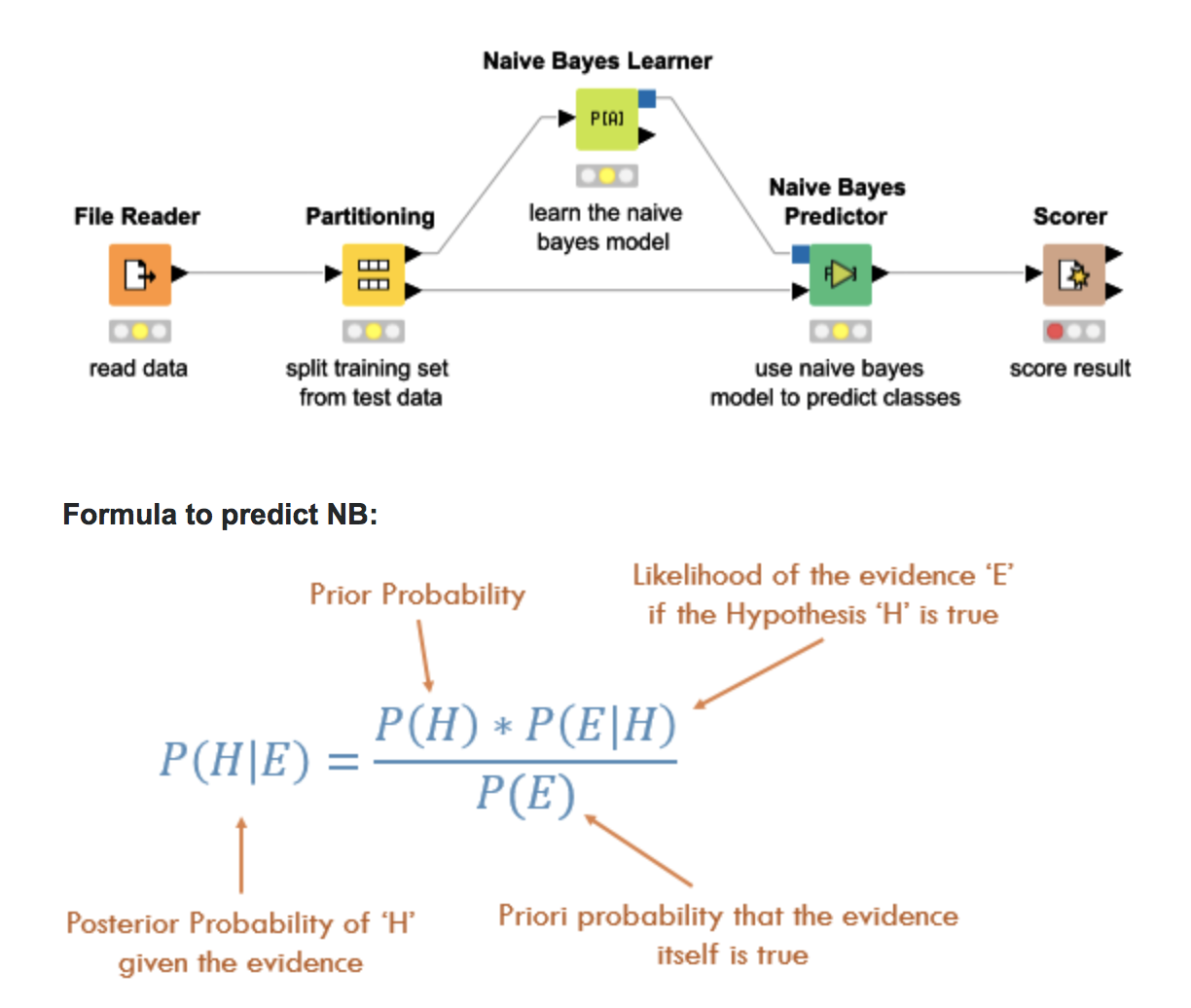
***Training Labels:***

Labeled corpora may be used to train and evaluate a wide range of learning algorithms. Assigning a label is considered a judgment task performed by a human (worker, judge, expert, annotator, etc.). Labels for this undertaking were determined exclusively by date and originating publication.

***Data:***

Data used in test and training sets was sourced from two sources. The first dataset was sourced from the Australian Broadcasting Corporation and consisted of approximately one million news headlines with corresponding dates. The second dataset was sourced from [www.examiner.com](http://www.examiner.com) and consisted of approximately three million headlines with corresponding dates. The first dataset was trained and used to extrapolate weekend, month, season, year, election time, first chronological half of the data set, prime minister’s affiliation and GDP growth. The second dataset was trained to determine if a headline held validity or should be considered clickbait. For both datasets, 10% of the data was randomly removed and used as a test set while the remaining 90% was used as a training set.

**Naïve Bayes Classifier:**

****

**Prediction:**

Successful prediction was measured with accuracy of prediction. The accuracy was measured using the formula:

accuracy = ((TP+ TN) / Total) : TP equals true positive and TN equals true negative.

**Results and Performance:** (Insert Table)

**Discussion:**

From the test results, it can be concluded that news headline data has a greater amount of informational value than previously assumed. This study has concluded that it is possible to extrapolate useable information in regard to a multitude of inquiries. Also, the approach outlined in this report can be used to predict the informational quality of the underlying article of which the headline represents.

**Sources:**

A practical explanation of a Naive Bayes classifier. (2017, October 03). Retrieved October 09, 2017, from https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/

Joyce, J. (2003, June 28). Bayes' Theorem. Retrieved October 09, 2017, from <https://plato.stanford.edu/entries/bayes-theorem/>

A simple explanation of Naive Bayes Classification. (n.d.). Retrieved October 10, 2017, from https://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification#20556654

Mitchell, T. M. (2015). *Machine learning*. Johanneshov: MTM.