When one considers the full spectrum of privacy or she must ask themselves if they are comfortable with all of their characteristics provided in the public domain. This includes their personal information, location, etc. We currently have over 10 million IOT devices that are storing our information, and according to Tech Insider, there is a prediction that there will be 34 billion devices connected to the internet by the year 2020. As big data grows, enterprises will attain robust data privacy and until then we will ask ourselves if we can trust our sources of big data, what information are they collecting, and so on.

It is generally known that these enterprises that attain our information are informing us how they are making use of our information in their privacy policy. However, many users approve a privacy policy without reading it, and many of these policies are vague guidelines where it is completely impossible for users to foresee the scope and content of their consent to the processing of their data. Consequently, users have no choice if they want to use their service though I am certain they would want ti know exactly what they are agreeing to when accepting a privacy policy.

A good way for fixing this issue is having automated software to analyze these privacy policies and to put the user at east when accepting a privacy policy. Parvaneh Shayegh is a PhD in Computer Science and she is developing a program to fix this issue. With the use of Natural Language Processing, she is developing a program which will allow the program to read a document and extract all the terms. These terms are the keywords, critical words in the document that will be significant in interpreting each sentence in the privacy policy. She is generally focusing on the noun and the adverbs. The program will then capture the relationships among terms using the ML methods. To do so, we consider each text document in the training documents that are important enough to be retained in the transaction. This is when it performs data classification, which is the process of organizing data into categories in such a way that classification error will be least. It is used to extract models that accurately define important data classes within the given database. If a sentence contains any keywords, they are then placed in the selective category, and if a sentence does not contain any keywords, then they are placed in the non-selective category.

The first step in her program is to extract all keywords; however, there are too many identical keywords when doing so thus it creates quite a problem when performing an automated extraction. It extracts too many of the keywords. For example, for 30 privacy policies, the program will extract 2000 words. Because of this we would have to do manual work on the output. We plan on making sure the output carries the base of the word. So if the program automatically extracts words, for example collect, collected, collecting, we remove the rest and keep the word “collect” as it is consider the base. So, to formally clarify, to reduce the output of the dataset the text documents need to be processed by.

1. We manually remove stop words and place them in the non-selective category, i.e., words that appear frequently in the document but have no essential meanings; and
2. Retaining only the root form of words by stemming their affixes as well as prefixes.

To complete the process, K-Nearest Neighbor (KNN), Naïve Bayes’ Support vector machines are the three ML methods

K-nearest-neighbors is a simple classifier that can be considered as an important building block of many machine learning algorithms. This algorithm classifies an unknown example with the most common class among *k* closest examples. This means that, using an increasing size of training set data, the algorithm will assign a class to a value depending on its k nearest training data points in Euclidean space, where k is some number chosen by the user. In the figure 1.1, it shows an illustration as to how this algorithm is performed. The chosen number by the user for “*k*”, in this case, is three. It looks for the closest neighbors to determine the class of ‘c’. In this case, the closest neighbors are two elements of ‘o’ and one element of ‘a.’ Since the number of ‘o’ elements is greater than the number of ‘a’ elements, the class of ‘c’ is identified to be ‘o.’ The algorithm is quite simple and powerful, and no training is actually involved as new training examples can be easily be added.

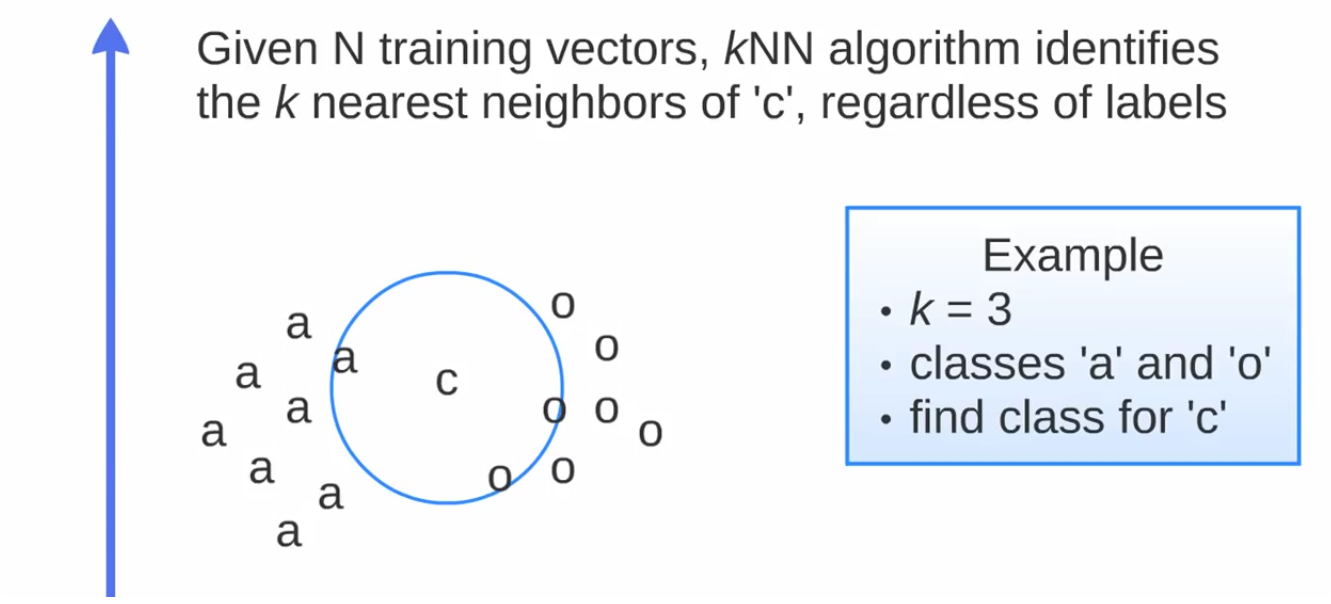


Figure 1.1

Support Vector Machine (SVM) is another supervised machine learning algorithm used for classification and regression practices by a separated hyperplane. This means that it collects a training set and the program categorizes new examples from that training set by an optimal hyperplane; which the program outputs. In the figure 2.1, it shows an illustration as to how this algorithm is performed. The training set consists of a set of circle and square classes. The green and red lines are known to be the hyper-plane and it segregates the two classes through distance. The distance from the hyper-plane and the element is known to be the margin. The goal of using SVM is it to find an optimal boundary between the possible classes provided**.** A hyperplane that passes too close to the classes is not is not recommended because it will be noise sensitive and will not generalize correctly. Therefore, the hyperplane of the line is best if it is passed between the classes as far as possible.

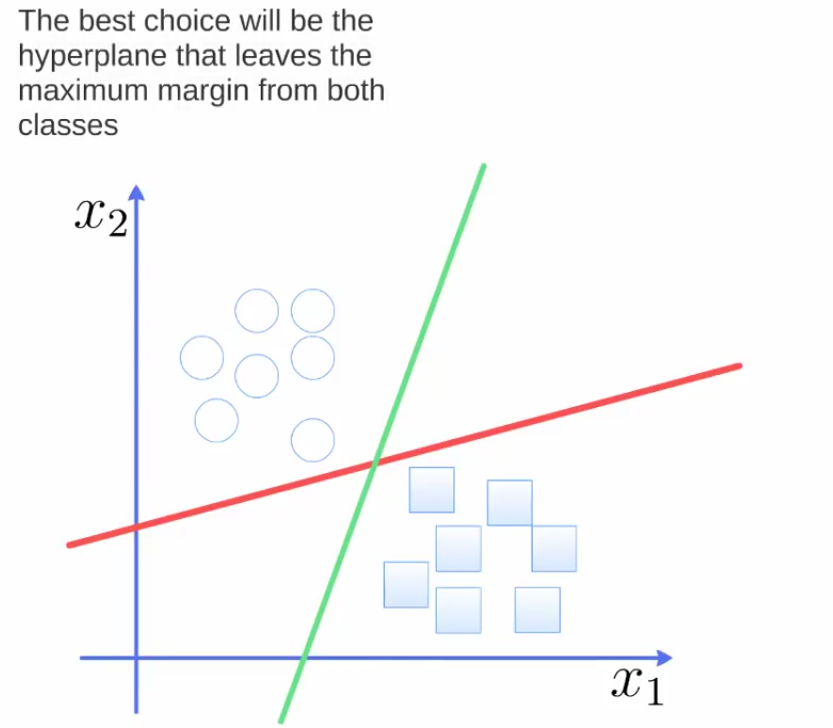


Figure 2.1

Naive Bayes is another supervised machine learning algorithm that makes use of the Bayes Theorem. The algorithm is performed by predicting the membership probabilities for each class. This involves the probability that given record or data point belongs to, in a particular class. In figure 3.1, it presents the formula used in this algorithm. The formula functions on how the features of each class are “generated.” The algorithm uses the training sets, and to implement the training set into the formula, it consists of the target value (hypothesis) and each attribute value at of each datum instance.

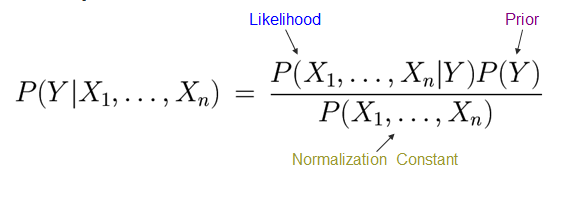


Figure 3.1

Maximum A Posteriori (MAP) is finding the class with the highest probability. This is considered to be the best class. All features in the Naive Bayes algorithm is considered independent; which is known to be the Naïve Bayes assumption. In figure 3.2, the classifier is often represented as the graph presented. In the graph, the direction of the arrows clarify that each class causes certain features, with a certain probability.

Figure 3.2

Parvaneh used a tool she found online where she can upload the privacy policy document. In it, it would create an output of each word in a sentence. Her goal is to determine whether each sentence is considered sensitive or non-sensitive.

She wanted to classify each sentence based on the presence or absence of each word. Again, her goal is to determine whether each sentence is considered sensitive or non-sensitive. For example, there was a sentence in one of the documents which included the terms “access” “store” “information.” The program will output a file in excel where each word is arranged in a vector with either a number of a zero or a one. In this case, the terms, “access” “store” “information,” will have a one at the bottom of the cell as it clarifies that it is a sensitive word. Other terms, for example “process,” was not considered a sensitive word and thus will have a zero at the bottom of the cell.

Using the three algorithms perspectives, each vector of 1 and 0’s will then calculate the distance between each vector. It then checks if they are similar; either sensitive or non-sensitive. For example, one sentence had ten neighbors (words). Seven of them are sensitive while three of them are non-sensitive. Thus, the result is that the sentence is sensitive. Using the algorithm, it calculated that it is 0.7 sensitive and 0.3 non - sensitive; making it a sensitive sentence.

The next step is to break each privacy policy documents, based on its topic. Our goal here is to pick each word as keywords that once the program notices, will be able to determine what topic the sentence is about. In figure 4.1, an excel file is shown to have words in various privacy policies in each vector. If a word has a number below, it tells as to which topic it belongs to. For words that have no numbers, belong to no topic. Below is what each number is represented:

1. Information
2. Collection
3. Sharing
4. Permission
5. Technology

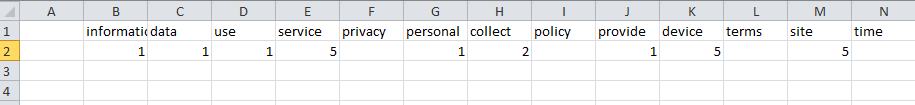


Figure 4.1