Artificial Intelligence assignment 2: Spam Filtering Using a Naive Bayes Text Classifier

Ilse Barf (S3500306), Galina Lesnic (SS3367398), Nathan van Beelen (S3392961)

September 30, 2019

3.3 Example runs

What happens if you train and test on the same data?

The accuracy will be higher. This is due to the fact that it can calculate correlations that are apparent in this particular set of the data but do not define the difference between regular and spam mails. In other words, you can safely overfit on the train data and get a high accuracy score. Additionally, we used the vocabulary of the train data, which will be exactly the same in the test if we use the same data for testing. This is not the case if we use a different data set for testing.

5 Final Questions

The data used in this assignment contains only e-mails in the English language. What happens if an e-mail in Dutch is given to your spam filter trained with English messages? How will the Dutch message be classified? Assume that there are no common words in English and Dutch. Explain your answer.

- Nathan van Beelen: The messages will be classified according to the number of regular and spam messages used in the training data. This is because the classifier works by calculating the likelihood that a certain word is from a regular or spam mail. However, in Dutch it will not recognise any words. So this chance will be equal for both classifications. The only thing that differs is the chance that a mail is regular or spam based on the number of regular or spam mails. In the case this is also the same, the chances are equal. Due to the way we classify mails this means that mails will be classified as spam since the chance that it is regular needs to be bigger in order for it to be classified as regular.

- Ilse Barf: In this case, if there are an equal number of spam and regular messages (which makes the probability for encoutering one equal), the probability that the message is regular will be equal to the probability that the message is spam (it will be a probability very close to 0). Since

the classifier will only classify an e-mail as regular if the probability that it is regular is strictly larger than the probability that it is spam, it will classify the e-mail as spam.

- Galina Lesnic:

The Naive Bayes assumption is that the attributes (or features) are independent. Are the words in a message really independent? And what can you say about the independence between and within bigrams? Explain your answer in 250 words.

- Nathan van Beelen: No, the individual words are not independent. This is because language has a certain structure. An adjective needs a noun for example. Another example would be semantics. The adjective 'new' would make sense in combination with the noun 'concept', but the adjective 'green' does not. This means that 'new' and 'concept' have a semantic dependence, but 'green' and 'concept' do not. Both in the case of between bigrams and within bigrams there is a dependence. The dependence within bigrams can be explained in a similar way as I did for the individual words. Another example would be the determiner, which will always be accompanied by either an adjective or a noun on which it is dependent. In the case of the dependence between bigrams it is mostly due to the dependence of words to form a sentence (although the dependence can also be similar to within bigrams if the words are split correctly). To illustrate the dependence between bigrams, consider the sentence 'Bob is playing soundly.' This will be split in: 'Bob is' and 'playing soundly.' The bigram 'Bob is' doesn't make sense in and of itself. It needs a verb in order to make sense. Therefore it is dependent on the bigram containg the verb.

- Ilse Barf: The words in a message are not likely to be independent. This is caused firstly by the fact that a message typically has a certain subject. Within the context of a subject certain words might be either more or less common and either more or less likely to appear together in the same message.

  Another factor which causes dependece is the grammar of a language, which forces words to appear together and sometimes even in a specific order.

  These arguments also hold for indepency between bigrams, since the words that are next to each other in a message are not necessarily dependent, so the dependency between bigrams is very to the dependency between words.

  Within bigrams, it is even more likely that the words are dependent, since words like adjectives, adverbs, and prepositions that go with certain verbs, are often found in close proximity of the word they go with. However, this is not always the case, and since we do not include words with less than four letters, the dependency will be smaller than initially expected.

- Galina Lesnic: