Indicative answers of Supervision questions: set 1

February 12, 2017

1 Sentiment lexicon

2. On the whole, *not* alters the polarity of following words, but there are lots of counter-examples. Doing this properly requires parsing.

4. The point is the number of significant figures. 0.796 is a reasonable answer, though I would probably go for 0.80, but any more digits is wrong.

5. Suppose the class split were 95% A to 5% B. Always guessing A gives 95% accuracy even though the classifier is doing nothing.

2 Naive Bayes

1. a. Create a table with F1, not F1, F2, not F2, F3, not F3 with class A and B

   $P(A|F1) = 5/10$ etc

   $P(F1|A) = 5/50$ etc

   b. Naive Bayes: Probability of class given document is proportional to P(class) x product of probabilities of feature given class. Better not to use logs here, because we can work this out more straightforwardly (good practice for exams). We are not working out probabilities (we can't really estimate probability of the document) but relative probabilities. So the answer will not sum to 1.

   $$P(A|d) \propto P(A) \times P(F_1|A) \times P(notF_2|A) \times P(F_3|A) = .5 \times 5/50 \times 50/50 \times 3/50$$

   $$P(B|d) \propto P(B) \times P(F_1|B) \times P(notF_2|B) \times P(F_3|B) = .5 \times 5/50 \times 40/50 \times 27/50$$

   c. Same as b apart from initial element in product.

   d. F1 is never informative. F2 always gives the correct decision when it is present, but it will only occur 10% of the time. F3 is less reliable, but is useful for more documents, so is overall more useful.

e. Essentially this is about feature selection. We are not covering how this is done formally, but experimentally, one can remove features and see what happens to the performance. Obviously this won't work with a huge number of features. Rather than "using up" test data with multiple trials, one could also generate artificial data with NB and experiment.

2. The argument there, which seems plausible, is that using tokens does not model burstiness. e.g., once you've seen "Bond" in a document, you're more likely to see "Bond" again in the same document. This is therefore a case where the failure of the independence assumption hurts performance. Burstiness doesn't apply to very frequent words (e.g., closed class words like the), but these are rarely good indicators of sentiment.

3 Statistical properties of language

1. Several things involved here. The first point is (yet another) ambiguity in the word *word*. Dictionaries don't usually contain most proper names, though this is just a convention. Most (but definitely not all) of the very rare words encountered when processing a large corpus will be proper names. People who say "that's not a word" are implicitly talking about a dictionary word (plus word endings).

One reason for intuitions about possible words is that languages have different phonological properties. Pferd is a perfectly good word in German but the initial pf doesn't correspond to anything in English. Some German words (e.g., pfennig) do nevertheless make it into English dictionaries. In some cases, it is exceedingly difficult for an adult English speaker to mimic the phonology of a language: this is the case for the so-called "click" languages, of which Kx'a is an example. Kx'a itself is a word which violates normal orthographic conventions of English.

Finally, abtruce could be a word of English but looks like a spelling mistake.