Sentiment analysis

(document classification)

Document classification

- language identification
- authorship attribution
- spam detection

Goals

- objective vs. subjective
- positive vs. negative
- more complex…
    - Jurafsky and Martin, ch. 21

Approaches

- lexicon-based
- supervised machine learning

Lexicon-based

happy, joy, great… vs. bad, sadness, horrible…

- “This phone is great.”
- "I don't like it, I *love* it!"

not always right, just heuristic
maybe we need like love—5 point , like — 3 point

no matter how long the document is, we just count and look.
e.g. 50 positive, 20 negative — 30 positive

Problems

- "not"
- implication
- sarcasm
- document context
- cultural/personal context
- ...

"Not"

"This phone is not great." - unfavorable product review

Implication

"At least the food was good, when it finally arrived." - unfavorable restaurant review

Sarcasm

"Parking at Duke is so much fun!" - no one ever

Vocal tone, facial cues, and body language are also critical for conveying sentiment.

Cultural/personal context

- "This phone has a 2-second battery life."
- "This phone is the greatest paperweight ever."
- "Bless her heart."

Document context

"I ran into my ex yesterday. It was terrifying."

-vs.-

"I went to a haunted house yesterday. It was terrifying."

Supervised classification

predictor(s) → label(s)

generative/discriminative

linear/nonlinear

if I can present to you a probability that this string came from Class i and the probability that this document is of that class, we can infer the posterior probability that this document is of that class

Naïve Bayes classification

probability of class is proportional to the probability of having observed this set of words given that this is the class

generative, linear

indicator of class: positive negative neutral

$$p(c_i | w_1, w_2, \ldots, w_n) \propto p(c_i) p(w_1, w_2, \ldots w_n | c_i)$$

the model is generating w… based on the topic or t… sentiment that you asked…

we want to understand the probability of a class(Ci) given a set of observations (words or counts of words)

$$\approx p(c_i) \prod_j p(w_j | c_i)$$

Assumption: the probability of any word given the class or the topic or the sentiments is independent of the other word

The multinomial distribution has this "naïveté" property:

$$p(w_1, w_2, \ldots, w_n | c_i) = \frac{n!}{w_1! w_2! \ldots w_n!} p_{i,1}^{w_1} p_{i,2}^{w_2} \ldots p_{i,n}^{w_n}$$

'bag of words assumption':
all the orders lost, but the word is still there

Inference

If we draw a word for a document with class c_i, draw w_j with probability $p_{i,j}$.

e.g. the number of word good in the positive documents

$$p_{i,j} \approx \frac{count(w_j|c_i)}{\sum_k count(w_k|c_i)}$$

i.e. the fraction of all words in c_i documents that are w_j.

→ This is much simpler than LDA and has an analytically tractable solution.

analogous to the rare transition problem where we may have a token that's known like terrifying
and simply find we never saw terrifying in a positive document
we haven't seen it doesn't mean it's impossible — so we smoothing

Add-one smoothing

$$p_{i,j} \approx \frac{count(w_j|c_i) + 1}{\sum_k (count(w_k|c_i) + 1)}$$

Example

Jurafsky and Martin, ch. 4.3

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

The prior $P(c)$ for the two classes is computed via Eq. 4.11 as $\frac{N_c}{N_{doc}}$:

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

The word *with* doesn't occur in the training set, so we drop it completely (as mentioned above, we don't use unknown word models for naive Bayes). The likelihoods from the training set for the remaining three words "predictable", "no", and "fun", are as follows, from Eq. 4.14 (computing the probabilities for the remainder of the words in the training set is left as an exercise for the reader):

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20} \qquad P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$
$$P(\text{"no"}|-) = \frac{1+1}{14+20} \qquad P(\text{"no"}|+) = \frac{0+1}{9+20}$$
$$P(\text{"fun"}|-) = \frac{0+1}{14+20} \qquad P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

For the test sentence S = "predictable with no fun", after removing the word 'with', the chosen class, via Eq. 4.9, is therefore computed as follows:

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$
$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

Caveat

Humans don't agree on sentiment analysis!

Only about 70-80% of the time.

[Wiebe et al. 2005]

This is effectively a cap on sentiment analysis performance.

Resources

- http://ai.stanford.edu/~amaas/data/sentiment/