CSC411: Assignment #3

Due on Monday, March 19, 2018

Lukas Zhornyak

March 13, 2018

Environment

Parts 1-6 were created with Python 2.7.14 with numpy 1.14.0, scipy 1.0.0, scikit-image 0.13.1, and matplotlib 2.1.1, as well as all associated dependencies.

Part 1

The headline dataset consists of two distinct parts: a set of ”real” headlines and a set of ”fake” headlines. There are 1968 real headlines, averaging 8.33 words per headline and 5.29 characters per word, and 1298 fake headlines, averaging 12.15 words per headline and 4.98 characters per word. Note that fake headlines seem to be longer but use smaller words, suggesting a difference in the language used in both sets. This gives credence to the feasibility of classifying a headline as fake or not based on the words used. Of course, just a headline is not very much information to go on, so it is likely that the accuracy will not be phenomenal.

Some words that might prove useful in identifying a certain headline as fake or not are ”says” (178 occurrences in the real data set vs 47 occurrences in the fake data set), ”donald” (829 vs 228), and ”hillary” (24 vs 159). The code used to obtain these words is given in the submitted code.

Part 2

To implement a Naive Bayes classifier, the ratio of the probability of a certain headline being fake to the ratio of it being real was examined when attempting to classify the headline:

$$\frac{P(y = c \mid x_1, x_2, ..., x_p)}{P(y = c' \mid x_1, x_2, ..., x_p)} = \frac{P(y = c)) \prod_{i=1}^{p} P(x_i \mid y = c)}{P(y = c') \prod_{i=1}^{p} P(x_i \mid y = c')}$$

where y is the label, x_i is a binary value indicating the presence of i-th keyword, p is the number of keywords, and c is the fake or real class, with c' being the opposite. If this value is larger then 1, $P(y = c \mid x_1, x_2, ..., x_p)$ is larger than $P(y = c' \mid x_1, x_2, ..., x_p)$ and thus the headline is classified as fake, and vice versa. Some minor preprocessing was done to convert the headlines into this one-hot feature representation.

The above equation involves the product of several thousand different probabilities, many quite small. To prevent the issue of arithmetic underflow, the log of this probability ratio was used instead:

$$\log\left(\frac{P(y = c \mid x_1, x_2, ..., x_p)}{P(y = c' \mid x_1, x_2, ..., x_p)}\right) = \log\left(\frac{P(y = c)}{P(y = c')}\right) + \sum_{i=1}^{p} \frac{\log P(x_i \mid y = c)}{\log P(x_i \mid y = c')}$$

as suggested in the handout. In this new formulation, a value greater than 0 denotes fake news. From this point, standard Naive Bayes was used to determine the probabilities.

To tune the prior m and $\hat{p}$, a basic gradient descent was performed on the validation set with initial values $\hat{p} = 0.5$ and $m = 10$. A secant approximation with small step size was used to approximate the gradient at each point. This resulted in small but consistent performance improvements of about 2 to 4 percent on the validation and testing set, but a loss in performance on the training set of about 1 to 3 percent. This suggests that this method is working properly to prevent over-fitting. The final accuracy achieved on the training and test sets was 0.9528 and 0.8408, respectively.

Part 3

3 (a)

To find the words that maximizes the probability that a given text is real or fake given the presence or absence of it, Bayes' rule can be applied:

$$\arg\max_i P(y = c \mid x_i = a) = \arg\max_i \frac{P(x_i = a \mid y = c)P(y = c)}{P(x_i = a)}$$

where a is either 0 or 1, depending on whether the presence or absence is desired. Unfortunately, the true value of $P(x_i = a)$ is not known very well and may often be 0 if only the dataset is used to determine it, so this formula cannot be applied as is. However, considering that only the ordering of the probabilities $P(y = c \mid x_i)$ is needed to find the maximum and knowing that $P(y = c' \mid x_i = a) = 1 - P(y = c \mid x_i = a)$, the maximizing x_i can also be found as

$$\begin{aligned} \arg\max_i P(y = c \mid x_i = a) &= \arg\max_i \frac{P(y = c \mid x_i = a)}{P(y = c' \mid x_i = a)} \\ &= \arg\max_i \frac{P(x_i = a \mid y = c)P(y = c)}{P(x_i = a \mid y = c')P(y = c')} \\ &= \arg\max_i \frac{P(x_i = a \mid y = c)}{P(x_i = a \mid y = c')} \\ &= \arg\max_i \log\left(\frac{P(x_i = a \mid y = c)}{P(x_i = a \mid y = c')}\right) \end{aligned}$$

where the log was added to assist in interpreting the value. If $\log(\cdots)$ is greater than zero, it means that $x_i = a$ predicts that the text has the label c and vice versa. Larger magnitudes indicate a greater probability. The top ten words and associated log probability ratios are shown in table 1. Note that the strength of the presence of a word in determining whether a headline is real or fake is significantly larger than the strength of its absence. This makes intuitive sense since a typical headline will not contain the majority of the words, regardless of it being fake or real.

3 (b)

The top ten words, excluding stop words, and associated log probability ratios are shown in table 2.

ban	4.8711	trump	1.6880	breaking	4.6573	donald	0.3265
korea	4.7235	the	0.2619	3	4.4378	trumps	0.1204
travel	4.5503	hillary	0.1040	won	4.2343	us	0.0950
turnbull	4.3693	a	0.0906	soros	4.2343	says	0.0507
australia	4.1123	in	0.0869	u	4.0712	ban	0.0425
tax	3.7654	to	0.0843	woman	4.0712	north	0.0391
paris	3.6560	and	0.0787	because	4.0712	korea	0.0365
james	3.5331	is	0.0742	homeless	4.0712	travel	0.0305
trumps	3.4768	clinton	0.0735	liberty	3.9785	turnbull	0.0254
debate	3.4654	for	0.0592	reason	3.9785	australia	0.0195

(a) Presence suggests real. (b) Absence suggests real. (c) Presence suggests fake. (d) Absence suggests fake.

Table 1: Words most likely to denote a headline as real or fake based on its presence or absence with associated importance.

ban	4.8711
korea	4.7235
travel	4.5503
turnbull	4.3693
australia	4.1123
tax	3.7654
paris	3.6560
james	3.5331
trumps	3.4768
debate	3.4654

(a) Presence suggests real.

trump	1.6880
hillary	0.1040
clinton	0.0735
just	0.0532
america	0.0392
watch	0.0287
voter	0.0279
new	0.0265
victory	0.0257
voting	0.0249

(b) Absence suggests real.

breaking	4.6573
3	4.4378
won	4.2343
soros	4.2343
u	4.0712
woman	4.0712
homeless	4.0712
liberty	3.9785
reason	3.9785
7	3.8763

(c) Presence suggests fake.

donald	0.3265
trumps	0.1204
says	0.0507
ban	0.0425
north	0.0391
korea	0.0365
travel	0.0305
turnbull	0.0254
australia	0.0195
house	0.0188

(d) Absence suggests fake.

Table 2: Words most likely to denote a headline as real or fake based on its presence or absence with associated importance, excluding stop words.

3 (c)

Stop words are present in all written works and normally do not carry with them any particular viewpoint or biases. Additionally, since there is a large number of them and since they are relatively common, slight differences in their prevalence in the trained data set may be exploited, potentially overfitting the data.

From a different perspective, if there is a consistent difference in how stop words are used in real and fake headlines, then excluding these words causes a loss in information. Consider the difference in words present in tables 1b and 2b. Almost all of the words in the original list are stop words and are not present in the second. This confirms an earlier observation made in part 1: fake headlines seem to use more but shorter words, suggesting a greater prevalence of linking words.

Part 4

Part 5

Part 6

Part 7

Part 8