Assignment 2 – Solution

Machine Learning
MSc Business Analytics

Wolfram Wiesemann

1 Individual Assignment

1.1 Bayes' Theorem: Dining Safely

Using C for 'clean restaurant' (and the complementary event $\overline{C}$ for 'dirty restaurant') and S for 'falling sick' (as well as the complementary event $\overline{S}$ for 'not falling sick'), the prior probabilities are $\mathbb{P}(C) = 0.99$ and $\mathbb{P}(\overline{C}) = 0.01$. From the text we also know that $\mathbb{P}(S|C) = 0.01$ and $\mathbb{P}(S|\overline{C}) = 0.5$.

1. You eat at a restaurant in Kentish Town and you get sick. What is the probability that the restaurant practices good hygiene?

From the law of total probability, we obtain

$$\mathbb{P}(S) = \mathbb{P}(S|C) \cdot \mathbb{P}(C) + \mathbb{P}(S|\overline{C}) \cdot \mathbb{P}(\overline{C}) = 0.01 \cdot 0.99 + 0.5 \cdot 0.01 = 0.0149.$$

We then obtain that

$$\mathbb{P}(C|S) = \frac{\mathbb{P}(C)}{\mathbb{P}(S)} \cdot \mathbb{P}(S|C) = \frac{0.99}{0.0149} \cdot 0.01 = 66.44 \cdot 0.01 = 0.66$$

as well as

$$\mathbb{P}(\overline{C}|S) = \frac{\mathbb{P}(\overline{C})}{\mathbb{P}(S)} \cdot \mathbb{P}(S|\overline{C}) = \frac{0.01}{0.0149} \cdot 0.5 = 66.44 \cdot 0.5 = 0.34.$$

Thus, it is more likely that the restaurant practices clean hygiene.

2. You go to the same restaurant for a second time, and you get sick again. What is the probability of the restaurant practicing good hygiene now?

We denote by S_1 and S_2 the two events 'sick for the first time' and 'sick for the second time', respectively. From the law of total probability, we obtain

$$\begin{aligned}
\mathbb{P}(S_1 \wedge S_2) &= \mathbb{P}(S_1 \wedge S_2|C) \cdot \mathbb{P}(C) + \mathbb{P}(S_1 \wedge S_2|\overline{C}) \cdot \mathbb{P}(\overline{C}) \\
&= \mathbb{P}(S_1|C) \cdot \mathbb{P}(S_2|C) \cdot \mathbb{P}(C) + \mathbb{P}(S_1|\overline{C}) \cdot \mathbb{P}(S_2|\overline{C}) \cdot \mathbb{P}(\overline{C}) \\
&= 0.01^2 \cdot 0.99 + 0.5^2 \cdot 0.01 \;=\; 0.0026.
\end{aligned}$$

We then obtain that

$$\begin{aligned}\mathbb{P}(C|S_1 \wedge S_2) &= \frac{\mathbb{P}(C)}{\mathbb{P}(S_1 \wedge S_2)} \cdot \mathbb{P}(S_1 \wedge S_2|C) \\ &= \frac{\mathbb{P}(C)}{\mathbb{P}(S_1 \wedge S_2)} \cdot \mathbb{P}(S_1|C) \cdot \mathbb{P}(S_2|C) \\ &= \frac{0.99}{0.0026} \cdot 0.01^2 \;=\; 0.0381\end{aligned}$$

as well as

$$\begin{aligned}\mathbb{P}(\overline{C}|S_1 \wedge S_2) &= \frac{\mathbb{P}(\overline{C})}{\mathbb{P}(S_1 \wedge S_2)} \cdot \mathbb{P}(S_1 \wedge S_2|\overline{C}) \\ &= \frac{\mathbb{P}(\overline{C})}{\mathbb{P}(S_1 \wedge S_2)} \cdot \mathbb{P}(S_1|\overline{C}) \cdot \mathbb{P}(S_2|\overline{C}) \\ &= \frac{0.01}{0.0026} \cdot 0.5^2 \;=\; 0.9619.\end{aligned}$$

Thus, with 96.2% probability the restaurant does *not* practice clean hygiene!

1.2 Building a Spam Filter

1. Compute the prior probabilities of a new SMS message being 'spam' or 'ham'. (These are just the empirical probabilities from the four messages above.)

The priors are $\mathbb{P}(\text{spam}) = \frac{2}{4} = 0.5$ and $\mathbb{P}(\text{ham}) = \frac{2}{4} = 0.5$.

2. For each de-capitalised keyword that appears in your training set (that is, 'i', 'am', 'not', 'coming', 'good', 'work', 'do', 'you', 'need', 'viagra', 'win', 'an' and 'imac'), build a frequency table that records the likelihoods $\mathbb{P}(W|\text{ham})$, $\mathbb{P}(\neg W|\text{ham})$, $\mathbb{P}(W|\text{spam})$ and $\mathbb{P}(\neg W|\text{spam})$.

The likelihoods (truncated to the $[0.05, 0.95]$ interval) are given in the following table:

W	$\mathbb{P}(W\|\text{ham})$	$\mathbb{P}(W\|\text{spam})$	$\mathbb{P}(\neg W\|\text{ham})$	$\mathbb{P}(\neg W\|\text{spam})$
i	0.5	0.05	0.5	0.95
am	0.5	0.05	0.5	0.95
not	0.5	0.05	0.5	0.95
coming	0.5	0.05	0.5	0.95
good	0.5	0.05	0.5	0.95
work	0.5	0.05	0.5	0.95
do	0.05	0.5	0.95	0.5
you	0.05	0.5	0.95	0.5
need	0.05	0.5	0.95	0.5
viagra	0.05	0.5	0.95	0.5
win	0.05	0.5	0.95	0.5
an	0.05	0.5	0.95	0.5
imac	0.05	0.5	0.95	0.5

Now we can make predictions. For 'Coming home', we first calculate

$$\begin{aligned}
&\mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \neg\text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \neg\text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}|\text{ham}) \cdot \mathbb{P}(\text{ham}) + \\
&\quad \mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \neg\text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}|\text{spam}) \cdot \mathbb{P}(\text{spam}) \\
&= \mathbb{P}(\neg\text{i}|\text{ham}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{ham}) \cdot \mathbb{P}(\text{ham}) + \\
&\quad \mathbb{P}(\neg\text{i}|\text{spam}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{spam}) \cdot \mathbb{P}(\text{spam}) \\
&= 0.5^6 \cdot 0.95^7 \cdot 0.5 + 0.95^5 \cdot 0.5^7 \cdot 0.05 \cdot 0.5 \;=\; 0.00561.
\end{aligned}$$

We now calculate

$$\begin{aligned}
&\mathbb{P}(\text{ham}|\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \neg\text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \frac{\mathbb{P}(\text{ham})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac}|\text{ham}) \\
&= \frac{\mathbb{P}(\text{ham})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i}|\text{ham}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{ham}) \\
&= \frac{0.5}{0.00561} \cdot 0.5^6 \cdot 0.95^7 \;=\; 0.97305
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(\text{spam}|\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \neg\text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \frac{\mathbb{P}(\text{spam})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac}|\text{spam}) \\
&= \frac{\mathbb{P}(\text{spam})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i}|\text{spam}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{spam}) \\
&= \frac{0.5}{0.00561} \cdot 0.95^5 \cdot 0.5^7 \cdot 0.05 \;=\; 0.02695.
\end{aligned}$$

Thus, the message should considered to be 'ham'.

For 'Get Viagra now', we first calculate

$$\begin{aligned}
&\mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \neg\text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \neg\text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}|\text{ham}) \cdot \mathbb{P}(\text{ham}) + \\
&\quad \mathbb{P}(\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \neg\text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}|\text{spam}) \cdot \mathbb{P}(\text{spam}) \\
&= \mathbb{P}(\neg\text{i}|\text{ham}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{ham}) \cdot \mathbb{P}(\text{ham}) + \\
&\quad \mathbb{P}(\neg\text{i}|\text{spam}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{spam}) \cdot \mathbb{P}(\text{spam}) \\
&= 0.5^6 \cdot 0.95^6 \cdot 0.05 \cdot 0.5 + 0.95^6 \cdot 0.5^7 \cdot 0.5 \;=\; 0.00316.
\end{aligned}$$

We now calculate

$$\begin{aligned}
&\mathbb{P}(\text{ham}|\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \neg\text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \frac{\mathbb{P}(\text{ham})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac}|\text{ham}) \\
&= \frac{\mathbb{P}(\text{ham})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i}|\text{ham}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{ham}) \\
&= \frac{0.5}{0.00316} \cdot 0.5^6 \cdot 0.95^6 \cdot 0.05 \;=\; 0.0909
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(\text{spam}|\neg\text{i} \wedge \neg\text{am} \wedge \neg\text{not} \wedge \neg\text{coming} \wedge \neg\text{good} \wedge \neg\text{work} \wedge \neg\text{do} \wedge \neg\text{you} \wedge \neg\text{need} \\
&\qquad \wedge \text{viagra} \wedge \neg\text{win} \wedge \neg\text{an} \wedge \neg\text{imac}) \\
&= \frac{\mathbb{P}(\text{spam})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac}|\text{spam}) \\
&= \frac{\mathbb{P}(\text{spam})}{\mathbb{P}(\neg\text{i} \wedge \ldots \wedge \neg\text{imac})} \cdot \mathbb{P}(\neg\text{i}|\text{spam}) \cdot \ldots \cdot \mathbb{P}(\neg\text{imac}|\text{spam}) \\
&= \frac{0.5}{0.00316} \cdot 0.95^6 \cdot 0.5^7 \;=\; 0.9091.
\end{aligned}$$

Thus, the message should considered to be 'spam'.

2 Group Assignment

We proceed as follows:

1. Load the data into a Python data frame.

This can be done with the following commands:

```
import pandas as pd
messages = pd.read_csv ('SMSSpamcollection', sep = '\t', names = ["label", "
    sms"])
```

2. Pre-process the SMS messages: Remove all punctuation and numbers from the SMS messages, and change all messages to lower case. (Please provide the Python code that achieves this!)

You can create new columns by operating on the existing ones. For example, you can create a new column with the length of each messages as follows:

```
messages['length'] = messages.sms.apply (lambda x : len (x))
```

Lambda expressions provide a convenient way to create ad-hoc functions in Python. The 'apply' method applies a function to every element of a DataFrame column. We can now pre-process the SMS messages as follows:

```
whitelist = set ('abcdefghijklmnopqrstuvwxy ')
messages['sms'] = messages.sms.apply (lambda x : ''.join (filter (whitelist.
    __contains__, x.lower())))
```

3. Shuffle the messages and split them into a training set (2,500 messages), a validation set (1,000 messages) and a test set (all remaining messages).

This can be done as follows:

```
# shuffling with pandas
messages = messages.sample (frac=1).reset_index (drop = True)
# splitting
msgs = list (messages.sms)
lbls =list (messages.label)
trainingMsgs = msgs[:2500]
trainingLbls = lbls[:2500]
valMsgs = msgs[2500:3500]
valLbls = lbls[2500:3500]
testingMsgs = msgs[3500:]
testingLbls = lbls[3500:]
```

5. Explain the code: What is the purpose of each function? What do 'train' and 'train2' do, and what is the difference between them? Where in the code is Bayes' Theorem being applied?

The functions 'train' and 'train2' calculate and store the prior probabilities and likelihoods. In Naïve Bayes this is all the training phase does. The 'predict' function repeatedly applies Bayes' Theorem to every word in the constructed dictionary, and based on the posterior probability it classifies the message as 'spam' or 'ham'. The 'score' function calls 'predict'

for multiple messages and compares the outcomes with the supplied 'ground truth' labels and thus evaluates the classifier. It also computes and returns a confusion matrix.

The difference between 'train' and 'train2' is that the latter function only takes into account words that are 20 times more likely to appear in a spam message than a ham message. Not all words are equally informative. Using 'train2' will decrease the size of the dictionary significantly, thus making the classifier more efficient. There are multiple other ideas that one can use to construct more informative dictionaries. For example, you can treat words with the same root ('go', 'goes', 'went', 'gone', 'going') as the same word.

6. Use your training set to train the classifiers 'train' and 'train2'. Note that the interfaces of our classifiers require you to pass the ham and spam messages separately.

The training functions require the ham and spam messages to be passed on separately:

```
hammsgs = [m for (m, l) in zip (trainingMsgs, trainingLbls) if 'ham' in l]
spammsgs = [m for (m, l) in zip (trainingMsgs, trainingLbls) if 'spam' in l]
print len (hammsgs)
print len (spammsgs)
```

We have 2,180 ham messages and 320 spam messages in our training set. Now we can create a classifier with

```
clf = NaiveBayesForSpam()
```

and train it via

```
clf.train (hammsgs, spammsgs)
```

or

```
clf.train2 (hammsgs, spammsgs)
```

7. Using the validation set, explore how each of the two classifiers performs out of sample.

We can do this via:

```
score,confusion = clf.score (valMsgs, valLbls)
print score
print confusion
```

If we use the 'train' function, the confusion matrix is

	ham	**spam**
predicted ham	841	21
predicted spam	14	124

with an overall performance of 0.965.

Using 'train2', we obtain the confusion matrix

	ham	**spam**
predict ham	853	30
predict spam	2	115

with an overall performance of 0.968.

Our data is not equally divided into the two classes. As a baseline, let's see what the success rate would be if we always guessed 'ham':

```
print 'basescore', len([1 for l in valLbls if 'ham' in l]) / float (len (
    valLbls))
```

This gives a performance of 0.855, which is inferior. We can also calculate our in sample error:

```
score, confusion = clf.score (trainingMsgs, trainingLbls)
print score
print confusion
```

If we use the 'train' function (this might take a long time!), the confusion matrix is

	ham	**spam**
predict ham	2,124	28
predict spam	56	292

with an overall performance of 0.9764. Using 'train2', which is much faster, we obtain

	ham	**spam**
predict ham	2,178	53
predict spam	2	267

with an overall performance of 0.978.

8. Why is the 'train2' classifier faster? Why does it yield a better accuracy both on the training and the alidation set?

We have answered this question already under point 5: The difference between 'train' and 'train2' is that the latter function only takes into account words that are 20 times more likely to appear in a spam message than a ham message. This makes the dictionary much smaller and hence speeds up the algorithm.

However, why would a simpler model (less words) give a better performance — not just out of sample (in the validation set), where overfitting may deteriorate the quality of more sophisticated methods, but also in sample? The reason is the conditional independence assumption of Naïve Bayes. With more words in our dictionary, the number of correlated words is larger and therefore the violation of our theoretical assumption stronger. It appears that 'train2' mediates this effect.

9. How many false positives (ham messages classified as spam messages) did you get in your validation set? How would you change the code to reduce false positives at the expense of possibly having more false negatives (spam messages classified as ham messages)?

We can achieve this by changing the following line in the 'predict' function

```
if (posteriors[0] > 0.5):
        return ['ham', posteriors[0]]
```

to something like

```
if (posteriors[0] > 0.4):
        return ['ham', posteriors[0]]
```

There does not seem to be any difference. How about:

```
if (posteriors[0] > 0.001):
        return ['ham', posteriors[0]]
```

Using 'train2', we obtain the following confusion matrix for the validation set:

	ham	**spam**
predict ham	880	34
predict spam	1	85

with an overall performance of 0.965. With these settings we lose only one email in our validation set.

10. Run the 'train2' classifier on the test set and report its performance using a confusion matrix.

We can do so with the following piece of code:

```
score, confusion = clf.score (testingMsgs, testingLbls)
print score
print confusion
```

We obtain the following confusion matrix:

	ham	**spam**
predict ham	1178	93
predict spam	2	199

with an overall performance of 0.954. We identified correctly more than 2 thirds of the spam messages, and only lost 2 ham emails.