Exercise Session 1 Solutions

CS 233 - Introduction to Machine Learning

February 20, 2019

1 Probability Exercises

1.1 Spam filter 1

You have just installed a system to detect junk email. The system can identify junk messages in 99% of cases. Nevertheless, the system says that a message is junk when it isn't in 2% of cases. Given that 10% of received emails are junk:

(a) What is the probability that a message is not junk?

(b) Given that a message is junk, what is the probability that this message is detected as junk ?

(c) Given that a message is junk, what is the probability that this message is wrongly detected as not junk?

(d) What is the probability that a message is detected as junk?

(e) (Challenging question) What is the probability that a message truly is junk when the system says it is?

Solution: Let I denote the event "the message is junk" and D the event "the message is detected as junk". From the wording, $P(D|I) = 0.99, P(D|I^c) = 0.02, P(I) = 0.1$, therefore:

(a) $P(I^c) = 0.9$
(b) $P(D|I) = 0.99$
(c) $P(D^c|I) = 0.01$
(d) $P(D) = P(D|I)P(I) + P(D|I^c)P(I^c) = 0.117$
(e) $P(I|D) = \frac{P(D|I)P(I)}{P(D)} = 0.8462$

1.2 Spam filter 2

Assume you have collected a dataset of emails out of which 432 are considered to be **spam** and 2170 to be **legit** emails. Furthermore, you have selected three words, namely *exercise*, *fun* and *viagra* and counted how many emails contained each of them:

word	appearances in spam	appearances in legit
exercise	6	39
fun	59	9
viagra	39	19

You intend to use a naive Bayes classifier to classify new emails. Assume that you have received an email which contains only the words *exercise* and *fun* but not the word *viagra*. Do you classify it as **spam** or **legit**?

Solution
Let H denote a hypothesis which can be one of S (spam) or L (legit). Let W_e, W_f, W_v denote binary random variables corresponding to occurrence of the word *exercise*, *fun* and *viagra* respectively. Let $\#S, \#L$ denote number of spam and legit emails respectively.

Prior class probabilities:

$$p(H = S) = \frac{432}{432 + 2170} = 0.166$$
$$p(H = L) = 1 - p(S) = 0.834$$

Probabilities of words conditioned on the class:

$$p(W_e = 1|H = S) = \frac{6}{\#S} = 0.014, \quad p(W_e = 1|H = L) = \frac{39}{\#L} = 0.018$$
$$p(W_f = 1|H = S) = \frac{59}{\#S} = 0.137, \quad p(W_f = 1|H = L) = \frac{9}{\#L} = 0.004$$
$$p(W_v = 1|H = S) = \frac{39}{\#S} = 0.090, \quad p(W_v = 1|H = L) = \frac{19}{\#L} = 0.009$$

Posteriors (note that $p(A, B|C) = p(A|C)p(B|C)$ since we assume Naive Bayes classifier):

$$\begin{aligned}
p(H = S|W_e = 1, W_f = 1, W_v = 0) &\propto p(W_e = 1, W_f = 1, W_v = 0|H = S)p(H = S) \\
&= p(W_e = 1|H = S)p(W_f = 1|H = S)p(W_v = 0|H = S)p(H = S) \\
&= 0.014 \cdot 0.137 \cdot (1 - 0.090) \cdot 0.166 \\
&= 0.0002897 \\
p(H = L|W_e = 1, W_f = 1, W_v = 0) &\propto p(W_e = 1, W_f = 1, W_v = 0|H = L)p(H = L) \\
&= p(W_e = 1|H = L)p(W_f = 1|H = L)p(W_v = 0|H = L)p(H = L) \\
&= 0.018 \cdot 0.004 \cdot (1 - 0.009) \cdot 0.834 \\
&= 0.0000595
\end{aligned}$$

The email will be classified as **spam** since $p(S|W_e, W_f, \neg W_v) > p(L|W_e, W_f, \neg W_v)$