Assignment 5

AUTHOR
Torin White - 657467127

A document can be represented as a 4-dimensional vector $x = (F, D, S, G)$, where F, D, S, G are all binary variables indicating whether the corresponding word from a vocabulary V appears in the document or not. For example: consider the following tiny vocabulary: $V = football, defence, strategy, goal.$

Then, a sentence "Adam from UIC scored two goals in a community football game." is represented as $x = (F = 1, D = 0, S = 0, G = 1)$, since it contains only the words from V: football and goal. We do not care about the order of the words, nor the words that are not in the vocabulary. We want to classify documents as being about sports $(C = 1)$ or not $(C = 0)$. A simple model for $P(F, D, S, G|C)$ is Naïve Bayes:

$P(F, D, S, G|C) = P(F|C) \cdot P(D|C) \cdot P(S|C) \cdot P(G|C).$

Q1. [17 pts] State in natural language what is the conditional independence relationship assumed by Naïve Bayes model.

Naive Bayes model states that F, D, S, G are conditionally independent given C. So the appearance of the words football, defence, strategy and goal are independent of one another given that the document they are found in is a sports article. If not conditioned on C, it is highly likely that F, D, S and G are not independent and if one or more of the words appear in a document, the others are far more likely to appear, aka are marginally correlated. However, conditioned on the condition C (sports article) the probability of seeing one of the words does not affect the probability of seeing another.

Q2. [30 pts] Assume the conditional distribution tables (CPTs) are given by

$P(C=1) = 0.5$	$P(C=0) = 0.5$
$P(F = 1\|C=1) = 0.8$	$P(F = 1\|C=0) = 0.1$
$P(D = 1\|C=1) = 0.7$	$P(D = 1\|C=0) = 0.7$
$P(S = 1\|C=1) = 0.2$	$P(S = 1\|C=0) = 0.8$
$P(G = 1\|C=1) = 0.7$	$P(G = 1\|C=0) = 0.3$

Now a new document arrives and it is described by x = (F=0, D=1, S=1, G=1). Assuming a naïve Bayes model, what is the probability of this document being about sports? That is, compute P(C=1 | x). To enable partial grading, you may write out the formula using symbols (e.g. P(F=1|C=0)), then plug in the numbers, and then calculate the final value.

$$P(Y|X_{1:d}) = \frac{P(X_{1:d}|Y)P(Y)}{P(X_{1:d})}$$

where $P(x_{1:d}|C) = \prod_{j=1:d} P(x_j|y)$ and $P(X_{1:d}) = \Sigma_{Y'} P(X_{1:d}|Y')P(Y')$

So, for the example

$$P(C = 1|F = 0, D = 1, S = 1, G = 1) = \frac{P(F = 0, D = 1, S = 1, G = 1|C = 1)P(C = 1)}{P(F = 0, D = 1, S = 1, G = 1)}$$

$P(F = 0, D = 1, S = 1, G = 1|C = 1) = P(F = 0|C = 1)P(D = 1|C = 1)P(S = 1|C = 1)P(G = 1|C = 1) = .2 * .7 * .2 * .7 = .196$

$P(C = 1)$ is given as .5

and
$P(F = 0, D = 1, S = 1, G = 1) = P(F = 0|C = 1)P(C = 1) + P(D = 1|C = 1)P(C = 1) + P(S = 1|C = 1)P(C = 1) + P(G = 1|C = 1)P(C = 1) = .2 * .5 + .7 * .5 + .2 * .5 + .7 * .5 = .1 + .35 +$

So $P(C = 1|F = 0, D = 1, S = 1, G = 1) = \frac{.196 * .5}{.9} = .109$

Q3. [30 pts] Suppose for each $P(F=1|C), P(D=1|C), P(S=1|C)$, and $P(G=1|C)$, we assign a $Beta(2,2)$ prior. Recall $Beta(\alpha,\beta)$ distribution has density function $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Its mean is $\frac{\alpha}{\alpha+\beta}$ and its mode is $\frac{\alpha-1}{\alpha+\beta-2}$.

Now suppose we observed the following five data points (documents):

F	D	S	G	C
1	0	1	0	1
0	1	0	1	1
1	1	0	0	0
1	0	1	0	0
0	1	0	1	0

What is the MAP estimate of $P(F=1|C=1)$ and $P(D=1|C=0)$?

$\theta_{MAP} = argmax_\theta p(\theta|X) = argmax_\theta p(X|\theta)p(\theta)$

the posterior of the Beta function is

$p(\theta|A,N,\alpha,\beta) \propto \theta^{A+\alpha-1}(1-p)^{N-A+\beta-1}$

where $A = \Sigma_t x^t$

$\alpha' = A + \alpha$ and $\beta' = N - A + \beta$

for $P(F=1|C=1)$

A = 1 N = 2

$\alpha' = 1 + 2 = 3$

$\beta' = 2 - 1 + 2 = 3$

$\theta_{MAP} = \frac{\alpha'-1}{\alpha'+\beta'-2}$

$\theta_{MAP} = \frac{3-1}{3+3-2}$

$\theta_{MAP} = 1/2$

for $P(D=1|C=0)$

A = 2 N = 3

$\alpha' = 2 + 2 = 4$

$\beta' = 3 - 2 + 2 = 3$

$\theta_{MAP} = \frac{4-1}{4+3-2} = 3/5$

Q4. [23 pts] Suppose we also assign a Beta(2, 2) prior on P(C=1). Using the data table in Q3, what is the probability that the next document will be about sports (C=1)?

we can use the MAP (or the Bayes estimator) also to predict the next document:

$\theta_{MAP} = \frac{\alpha' - 1}{\alpha' + \beta' - 2}$

A = 2

N = 5

$\alpha' = 2 + 2 = 4$

$\beta' = 5 - 2 + 2 = 5$

$\theta_{MAP} = \frac{4-1}{4+5-2} = 3/7$

There's a 3/7th chance the next document will be about sports with a Beta(2,2) prior.