INFERENCE AND REPRESENTATION: HOMEWORK 1

STUDENT: MARGARITA BOYARSKAYA

1. Denote $A = \{\text{we have the disease}\}$, $B = \{\text{we test positive on having it}\}$.
We are given $P(A) = 25 * 10^{-6}, P(B|A) = 0.98$.
Then,

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) =$$

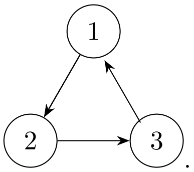
$$= 0.98 * 25 * 10^{-6} + 0.02 * (1 - 25 * 10^{-6})$$

Now, using Bayes' Theorem,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.98 * 25 * 10^{-6}}{0.98 * 25 * 10^{-6} + 0.02 * (1 - 25 * 10^{-6})} \approx 0.00122$$

The disease being rare is good news since as a result of this rarity the number of people (mis)diagnosed as having the disease among the healthy population greatly outweighs the number of the truly sick and diagnosed as such, bringing the fraction above to a minuscule value.

2. A cycle, if present in a graph, does not allow for a consistent assignment of probability distributions, since normalization of the sum of all values of JPD poses a challenge, as demonstrated in the following example of a directed graph:



Here each variable takes values from the set V, $|V| > 1$, and the value x_i for any node X_i, $i \in \{1, 2, 3\}$ is inherited from the parent node X_{pa}: $p(X_i = v|X_{pa} = v) = 1\ \forall v \in V$.
The "triplet" states $(v, v, v), v \in V$ (of which we have at least two) will each contribute a value of 1 to the JPD sum, which will thus become greater than 1, violating the definition of probability distribution.

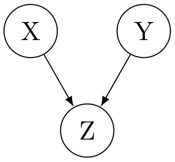
3. a. The marginal independence $X_i \perp\!\!\!\perp X_j$ holds for the following tuples of indices (i, j):
$(1, 2); (1, 3), (1, 5), (1, 7), (1, 8), (1, 9), (1, 10);$
$(2, 7), (2, 8);$
$(3, 4), (3, 7), (3, 8);$
$(4, 8);$
$(6, 7), (6, 8);$
$(7, 8), (7, 10);$
$(8, 10).$

b. The maximal set is $\{X_3, X_5, X_7, X_8, X_{10}\}$

4. Let's establish whether if the parametrization of the given distribution can be represented fully by a graph structure G.
If G is an I-map of $p(x, y, z)$, the conditional independencies of p (denote $I(p)$) must imply factorization according to graph G.

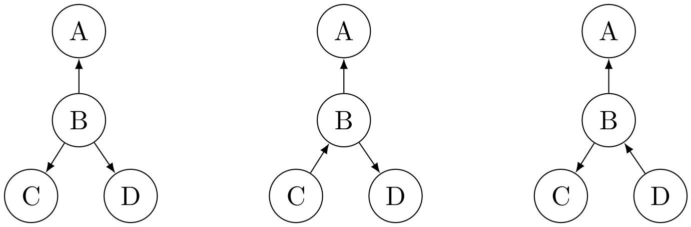
For the distribution given, the following independence statements hold:
1) given no evidence, any two variables are independent of each other;
2) given any one variable, the other two are dependent on each other.

For a single chosen pair X, Y let's examine a set of constraints:
- $X \perp\!\!\!\perp Y \in I(p)$
- $Y \perp\!\!\!\perp Z|X \notin I(p)$
- $X \perp\!\!\!\perp Z|Y \notin I(p)$
The only structure that these constraints admit is a v-shaped graph G:



However, $I_{d-sep}(G) \not\ni \{X \perp\!\!\!\perp Z\} \in I(p)$. Thus, G does not realize $I(p)$.

5. Two networks are I-equivalent if and only if they have the same skeleton and the same set of immoralities.
a. The only modification of this network that preserves the skeleton and immorality $A \rightarrow B \leftarrow C$ is reversing the direction of edge (B, D). However, this created new v-structures, violating the condition of the I-equivalency theorem.
b. There exist three networks that are I-equivalent to the given network:



6. [the .py file is attached separately]

The key step of the code is the construction of a dictionary that contains unique words found in all e-mails together with the number of their respective occurrences in all messages, as well as separately in messages labeled as ’SPAM’ and ’HAM’.
The conditional probabilities are computed as thus: $p(w_j|spam) = \frac{n_c+mp}{n+m}$,
where $m = \alpha|V|$, α is the smoothing coefficient.

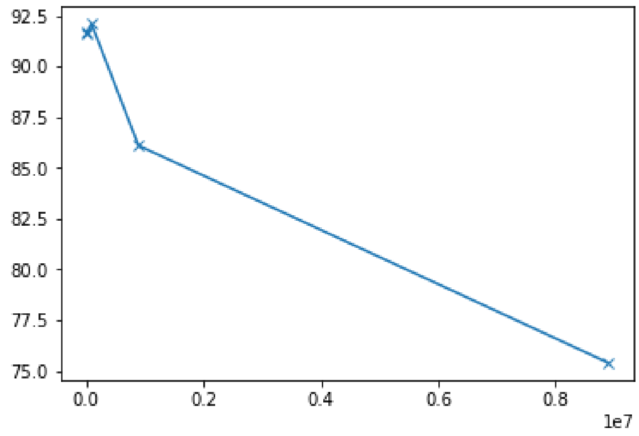
The code implements a prompt asking user if the set of ”stop words” is to be used.
When the stop words are excluded, the results are as follows:

p(spam)= 0.5736666666666667

Size of vocabulary $|V|$= 890
most likely words given SPAM are: [’http’, ’1’, ’b’, ’corp’, ’enron’]
most likely words given HAM are: [’content’, ’1’, ’corp’, ’enron’,
’aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa’]

Given the values of m: [890, 8900, 89000, 890000, 8900000]
Accuracies are: [91.7, 91.60000000000001, 92.10000000000001, 86.1, 75.4]

Accuracy plotted against m :

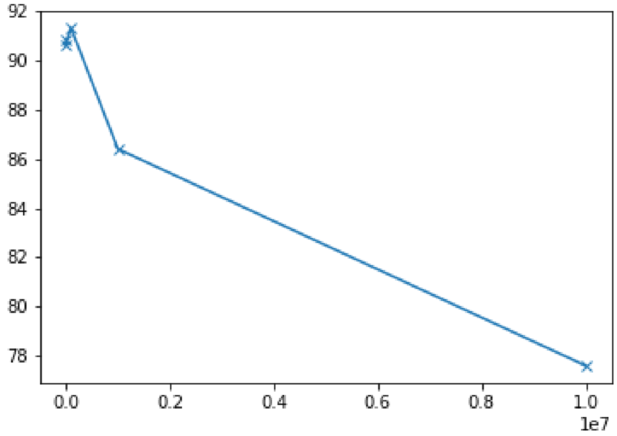


When the stop words are not excluded, the results are as follows:
p(spam)= 0.5736666666666667

Size of vocabulary $|V|$= 1000
call ”wSp” to see conditional probabilities of voc words given SPAM
call ”wH” to see conditional probabilities of voc words given HAM
most likely words given SPAM are: [’to’, ’the’, ’corp’, ’a’, ’enron’]
most likely words given HAM are: [’a’, ’to’, ’the’, ’enron’,
’aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa’]

Given the values of m: [1000, 10000, 100000, 1000000, 10000000]
Accuracies are: [90.60000000000001, 90.8, 91.3, 86.4, 77.60000000000001]

Accuracy plotted against m :



(e) To understand the behavior of accuracy it is useful to think of the m as a product $m = \alpha|V|$. Typically, the smoothing parameter is either chosen to be equl to 1 or < 1. The graphs indicate that as the smoothing parameter α becomes absurdly large, the model “oversmooths”, and the accuracy plummets.

(f) If I were a spammer, I would use images instead of symbols to represent sensitive words. (Rasterized images of typographic symbols should work fine, but using pictograms and even words encoded as a rebus seems attractive).