Prof. Wolfgang Nejdl
M.Sc. Ambreen Zaina, M.Sc. Jeff Reimer
Department of Knowledge-based Systems (KBS)

January 13, 2020

Artificial Intelligence II

Exercise 8

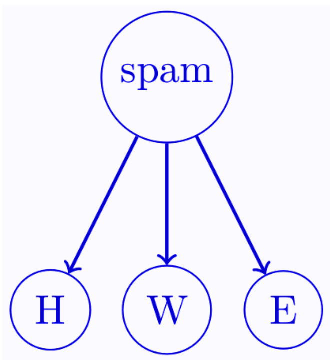
Q1. Naive Bayes

Your friend claims that he can write an effective Naive Bayes spam detector with only three features: the hour of the day that the email was received ($H \in 1, 2, ..., 24$), whether it contains the word 'viagra' ($W \in yes, no$), and whether the email address of the sender is Known in his address book, Seen before in his inbox, or Unseen before ($E \in K, S, U$).

(a) Flesh out the following information about this Bayes net:

Graph structure

Solution:



Parameters:

Solution:

$\theta_{spam}, \theta_{H,i,c}, \theta_{W,c}, \theta_{E,j,c}, i \in \{1, ..., 23\}, j \in \{K, S\}, c \in \{spam, ham\}$ is a correct minimal parameterization. Note that the sum-to-one constraint on distributions results in one fewer parameter than the number of settings of a variable. For instance, θ_{spam} suffices because $\theta_{ham} = 1 - \theta_{spam}$. Aside: a non-minimal but correct parameterization was also accepted since the question did not ask for minimal parameters.

Size of the set of parameters:

Solution:

1 + 23.2 + 1.2 + 2.2
The size of the set is the sum of parameter sizes. Every parameter has size = number of values x number of settings of its parents. For instance, $\theta_{\mathrm{H,i,c}}$ has 23 values of hour H and its parent c, the class has 2.

Suppose now that you labeled three of the emails in your mailbox to test this idea:

spam or ham?	H	W	E
spam	3	yes	S
ham	14	no	K
ham	15	no	K

(b) Use the three instances to estimate the maximum likelihood parameters.

Solution:

The maximum likelihood estimates are the sample proportions.
$\theta_{spam} = 1/3, \theta_{H,3,spam} = 1, \theta_{H,14,ham} = 1/2, \theta_{H,15,ham} = 1/2, \theta_{W,spam} = 1.0, \theta_{E,S,spam} = 1,$
$\theta_{E,K,ham} = 1$

(c) Using the maximum likelihood parameters, find the predicted class of a new datapoint with $H = 3, W = no, E = U$.

Solution:

No prediction can be made. Since $E = U$ is never observed, it has zero likelihood under both classes.

(d) Now use the three to estimate the parameters using Laplace smoothing and $k = 2$. Do not forget to smooth both the class prior parameters and the feature values parameters.

Solution:

The Laplace smoothed estimate for a categorical variable X with parameters $\theta_{1,...,d}$ for the $\{1, ..., d\}$ values of X is $\theta_i = \frac{x_i+k}{N+kd}$ where x_i is the number of times value i is observed, N is the total number of observations, and d is the number of values of X.

$\theta_{spam} = 3/7, \theta_{H,3,spam} = 3/49, \theta_{H,other,spam} = 2/49, \theta_{H,14,ham} = 3/50, \theta_{H,15,ham} = 3/50,$
$\theta_{H,other,ham} = 2/50, \theta_{W,spam} = 3/5, \theta_{W,ham} = 2/6, \theta_{E,S,spam} = 3/7, \theta_{E,other,spam} = 2/7,$
$\theta_{E,K,ham} = 4/8, \theta_{E,other,ham} = 2/8.$

(e) Using the parameters obtained with Laplace smoothing, find the predicted class of a new datapoint with $H = 3, W = no, E = U$.

Solution:

Ham. The probability under the model for each class is computed as the product of the class prior and the feature conditonals:
$p(ham)\alpha(1-\theta_{spam})(\theta_{H,other,ham})(1-\theta_{W,ham})(1-\theta_{E,K,ham}-\theta_{E,other,ham})$ and
$p(spam)\alpha(\theta_{spam})(\theta_{H,3,spam})(1-\theta_{W,spam})(1-\theta_{E,S,spam}-\theta_{E,other,spam})$
where both are proportional because the distribution has not been normalized.

(f) You observe that you tend to receive spam emails in batches. In particular, if you receive one spam message, the next message is more likely to be a spam message as well. Explain a new graphical model which most naturally captures this phenomena.

Graph structure

Solution:

The structure is the same as an HMM except each hidden state node has three observation child nodes.

Parameters:

Solution:

Add 2 parameters: transition to spam from spam and from ham.

Size of the set of parameters: Add 2 to the expression in the first question.