13.22 Text categorization is the task of assigning a given document to one of a fixed set of categories on the basis of the text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence or absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by the document category.
a. Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories.
b. Explain precisely how to categorize a new document.
c. Is the conditional independence assumption reasonable? Discuss.

a. Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories.

Example:

- **Task**: e-mail classification
- **Training Data**: a bulk of emails already classified in categories
- **Categories**: spam, personal, work
- **Frequency of categories**: helps to calc. P(Category=spam), P(Category=personal) and P(Category=work)
- **Vocabulary**: parse these emails to select the most 100 frequent words
- **Frequency of words per category**: parse these emails using the selected words and count the frequency of each word per category. The idea is to calculate P(Word=word_i | Category=category_j)

$P(Word = viagra \mid Category = spam)$: *gives the frequency of the word in category spam.*
$P(Word = viagra \mid Category = personal)$: *gives the frequency of the word in category personal.*
$P(Word = viagra \mid Category = work)$: *gives the frequency of the word in category work.*

It cool be useful to avoid setting the frequency value to zero. Think about what happen when we use the Naive Bayes to calculate the joint distribution. Once we calculate all of these values, we are ready to classify new emails with Naive Bayes.

$P(Word = viagra \mid Category = spam) = 0.1$ (*viagra is present* $1/10$ *spam emails*).
$P(Word = viagra \mid Category = personal) = 0.001$
$P(Word = viagra \mid Category = work) = 0.05$ (*the owner of the email could be a doctor*)

The joint distribution can be calculated with Naive Bayes:

$$P(Query,\ Word_1,\ Word_2,\ ...,\ Word_n) = P(Query) \times \prod_{i=1}^{n} P(Word_i \mid Query)$$

Observation: bigger the number of words, lower the probability of each word per category, and then it is lower the joint distribution value calculated using the Naive Bayes method.

b. Explain precisely how to categorize a new document.

Example:

1) The doctor receives a new email.
2) The program parse the email looking for the presence of absence of our vocabulary and set the values of the $Word_1$ *to* $Word_n$.
3) The program makes use of Naive Bayes to estimate the joint distribution for each category.

$$P(Query = spam,\ Word_1 = word_1, Word_2 = \neg word_2, Word_3 = word_3,\ ..., Word_n = word_n)$$
$$P(Query = personal,\ Word_1 = word_1, Word_2 = \neg word_2, Word_3 = word_3,\ ..., Word_n = word_n)$$
$$P(Query = work,\ Word_1 = word_1, Word_2 = \neg word_2, Word_3 = word_3,\ ..., Word_n = word_n)$$

We observe which was the category (of these joint distributions) with the greatest joint distribution probability to identify the category to be assigned to the new processed email.

c. Is the conditional independence assumption reasonable? Discuss.

The conditional independence is not reasonable from the point of view of the probabilities generated using Naive Bayes (assuming independence of the events). But from the point of view of the classification or the ordering of the results, Naive Bayes tends to do it well. Here I share the link to a very good paper that tries to explain why Naive Bayes performs well even when working with not independent variables:

http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf