- 13.18 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.
- 文本分类是根据所包含的文本将给定文档分配给一组固定的类别中的任何一个的任务。朴素贝叶斯模型经常用于此任务。在这些模型中，查询变量是文档类别，而“效果”变量是语言中每个单词的存在或不存在；假设单词独立出现在文档中，频率由文档类别决定。
  a.准确地解释如何构造这样的模型，给出一组被分配到类别的文档，作为“训练数据”。
  b.准确地解释如何对新文档进行分类。
  c.条件独立假设是合理的吗？讨论。

a. The model consists of the prior probability P(Category) and the conditional probabilities P(Wordi | Category). For each category c, P(Category =c) is estimated as the fraction of all documents that are of category c. Similarly, P(Wordi =true |Category =c) is estimated as the fraction of documents of category c that contain word i.
b. See the answer for 1. Here, every evidence variable is observed, since we can tell if any given word appears in a given document or not.
c. The independence assumption is clearly violated in practice. For example, the word pair “artificial intelligence” occurs more frequently in any given document category than would be suggested by multiplying the probabilities of “artificial” and “intelligence”.

该模型由先验概率 P（类别）和条件概率 P（Wordi | 类）组成。对于每个类别 c，P(Category=c)被估计为属于类别 c 的所有文档的分数。同样，P(Wordi=true|Category=c)被估计为包含单词 i 的类别 c 的文档的分数。

见 1 的答案。在这里，观察到每个证据变量，因为我们可以知道任何给定的单词是否出现在给定的文档中。

独立假设显然在实践中被违反了。例如，在任何给定的文档类别中，对“人工智能”一词的出现频率要比将“人工智能”和“智能”的概率相乘所建议的要高。