

CS 441/541
Artificial Intelligence
Fall, 2006

Homework 5:

Knowledge Representation, Information Retrieval, and Probabilistic Reasoning

Due Tuesday, November 14.

1. Suppose you are a doctor and a patient comes to you complaining of stomach pains. You know that the prior probability of stomach cancer (its incidence in the population) is 0.0008. There is a blood test for stomach cancer which has the following probabilities associated with it: If the patient has stomach cancer, the test is positive 80% of the time and (falsely) negative 20% of the time. If the patient does not have stomach cancer, the test is (falsely) positive 2% of the time and negative 98% of the time. You prescribe the blood test, which comes back positive. Given this data, what is the probability that the patient has stomach cancer?
2. For problem 1, give the full joint probability distribution of the two random variables Cancer ($\in \{\text{true}, \text{false}\}$) and BloodTest ($\in \{\text{positive}, \text{negative}\}$).
3. Suppose your problem has m random variables, each of which can take on n possible discrete values. How many entries are there in the full joint probability distribution of these m random variables?
4. Suppose you are training a naïve Bayes classifier to classify e-mail messages as “spam” or “not spam”. Suppose your classifier uses the presence (1) or absence (0) of the following keywords as its features: “stock”, “\$”, “offer”.

Suppose your training data is the following:

Instance	“stock”	“\$”	“offer”	Class
D1	0	0	1	not spam
D2	1	1	0	spam
D3	0	0	0	not spam
D4	1	1	1	spam
D5	0	1	1	spam
D6	0	1	0	not spam

Now suppose you encounter a new message, with the following feature values:

Instance	“stock”	“\$”	“offer”	Class
D1	1	0	1	?

Show how your naïve Bayes classifier will classify this message. Use m -estimate of probability, as described in class, with $p = 1/2$ and $m = 2$.

5. Suppose you have a corpus of 200 documents. You give your search engine a query Q . There are actually 10 documents relevant to your query in the corpus. The search engine returns 5 of these, plus 5 others that are not relevant. What is the precision and recall of this search engine on this query?