Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2013/2014

Exercise 5 - Lesson 13 2 April 2014

Exercise : Bayesian Reasoning

In this exercise we will use this approach to classify a body of text based on the frequency of occurrence of words. This can be used by a text editor to recognize the category of document that a person is composing (for example personal letters, technical reports, or computer code) and propose appropriate formatting, grammar and spelling corrections. Histograms (or bags) of words can be used to estimate the required probabilities for $P(E|H)$ and $P(E)$. For this task, assume that you have a training corpus composed of $K=5$ classes of text, and that for each class you have a sample composed of M words.

1) Explain how the training corpus can be used to construct a table for the frequency of occurrence for each word in the training data for each class of text. What it the probability that word, W, will occur in a sample from Class K? What is the probability that W will occur anywhere in the corpus?

2) Propose a method to obtain an initial estimate for the probability that an unknown text (a probe) belongs to each class.

3) Explain how to update the estimate for the probability of each class as the user types each new word in the probe (the unknown text).

4) What happens if the probe contains a word that was not in the training corpus? What can you do to protect against this case? How do you update the class estimates?

5) Is it necessary to recompute the bags of words if the user decides to create a new class of document, and provides a new corpus for this class?