

CS 410/510
Machine Learning
Winter, 2006

Homework 1: Feature Extraction

Due Tuesday, January 17.

For this homework you will design a set of features that you think will be useful in distinguishing spam e-mail from non-spam e-mail, and you will write code to extract these features from known spam and non-spam messages. You may use any programming language you want for this.

The 36 training examples of spam e-mail are given at
<http://www.cs.pdx.edu/~mm/MachineLearningWinter2006/spam-examples>

These examples have been adapted from examples given on an anti-spam web site at MIT.

The examples contain only the text of the main message; mail headers have been stripped off. For ease of reading, each message is separated from the previous one by a line of “%” characters. You can delete these lines of “%” if you wish before you run any code on this file.

For this assignment, you need to do the following:

1. Collect approximately 36 of your own non-spam e-mail messages (written by you or sent to you) to be the set of “negative” examples of spam (sometimes called “ham”).
2. Design six features that you can extract from these various messages, and which you think will be useful for learning algorithms to use in distinguishing spam from non-spam in general, not just for your own e-mail messages.
3. Write a program, in any high-level language, that extracts these features automatically. Make sure your program works for files of any length (not just limited to 36 messages). You will be using this code in later homework assignments.
4. Run your program on both the spam and non-spam training examples, and collect the output. **Note:** This assignment is *not* to write a spam classification problem. That will come later. This assignment is only to write a feature extractor.

Here is what you need to turn in, **hard copy**, in class on Jan. 17:

- A paragraph or two describing your six features, and why you think they will be useful in distinguishing spam from non-spam.

- Output from your program in the following format, one line per training example:

Message-number F1-value F2-value F3-value F4-value F5-value F6-value Class

where *F1-value* is the value of feature 1, etc., and *Class* is “P” if the training example is spam, “N” if the training example is non-spam.

Here is what you need to turn in **electronically**, by sending a file or tarball to me (mm@cs.pdx.edu) by class time on Jan. 17:

- A well-commented file or set of files containing your feature-extraction code. (**Please do not give me a hard copy of this.**)
- Any special directions for running your code.
- A file containing your non-spam examples. (**Please do not give me a hard copy of this either.**)

I will grade your homework based on (1) the clarity of your explanation of your six features and why they are likely to be useful for a general spam detector; (2) the correctness of your code in extracting these features; (3) the readability of your code.