

**CS 410/510**  
**Machine Learning**  
**Spring, 2007**

**Homework 1: Feature Extraction**

Due Tuesday, April 10.

For this homework you will design a set of features that you think will be useful in distinguishing spam e-mail from non-spam (“ham”) 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.

For this assignment, you need to do the following:

1. Go to the website

`http://spamassassin.apache.org/publiccorpus`

This is a public corpus of spam and ham messages for use in testing spam classification methods.

2. Read the “readme.html” file for a description of the data.
3. Create a subdirectory or folder for the training data.
4. Download the following files into this subdirectory or folder:

`20030228_easy_ham.tar.bz2`  
`20030228_hard_ham.tar.bz2`  
`20030228_spam.tar.bz2`

These are tarballs containing the spam and ham files.

5. Untar these files. The total is about 20 MB. If you don’t have enough room in your directory for these files, let me know.
6. Design 10-20 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.
7. Write a program, in any high-level language, that extracts these features automatically. You will be using this code in later homework assignments.

8. Run your program on both the spam and ham (easy and hard) 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 April 10:

- A paragraph or two describing your various features, and why you think they will be useful in distinguishing spam from non-spam.
- Example output from your program, on 10 spam and 10 ham messages, in the following format, one line per training example:

```
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 ham.

**Don’t give me the entire output files, just short samples so I can see what your program does.**

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 April 10:

- 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.

I will grade your homework based on (1) the clarity of your explanation of your 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.