

# Linear Classification

## I. Introduction

In this homework, we will implement logistic regression and linear discriminant analysis.  
**You shall submit a clearly written and commented report as well as your own code.**

## II. F\*!& Spams !

We want to build a classifier to filter spam emails. We will use a dataset which contains a training set and a test set of, respectively, 3065 and 1536 emails. The data is given in the file `spamData.zip`. Each email has been processed and a set of 57 features were extracted as follows:

- 48 features, in  $[0, 100]$ , giving the percentage of words in a given message which match a given word on a predefined list (called vocabulary). The list contains words such as "business", "free", "george", etc.
  - 6 features, in  $[0, 100]$ , giving the percentage of characters in the email that match a given character on the list. The characters are  `; ( [ ! $ #`.
  - Feature 55: the average length of an uninterrupted sequence of capital letters.
  - Feature 56: the length of the longest uninterrupted sequence of capital letters.
  - Feature 57: the sum of the lengths of uninterrupted sequence of capital letters.
1. What are the *max* and *mean* of the average length of uninterrupted sequences of capital letters in the training set?
  2. What are the *max* and *mean* of the lengths of the longest uninterrupted sequences of capital letters in the training set?
  3. Before training a classifier, we can apply several preprocessing methods to this data. We will try the following ones:
    - (a) Standardize the columns so they all have mean 0 and unit variance
    - (b) Transform the features using  $\log(x_{ij} + 0.1)$ , i.e. add 0.1 to each feature for every example and take the log. We add a small number to avoid taking log of zero !
    - (c) Binarize the features using  $\mathbb{I}(x_{ij} > 0)$ , i.e. make every feature vector a binary vector.

For each version of the data, i.e. using a different preprocessing, fit a logistic regression model.

- use cross validation to choose the regularization parameter.
  - report the mean error rate on the training and test sets.
  - what is the best preprocessing strategy? why?
4. Compare with the results of a Naive Bayes classifier.