Exercise: Spam classification using naive Bayes

We will re-examine the dataset from Exercise “Spam classification using logistic regression”.

1. Use naiveBayesFit.m and naiveBayesPredict.m on the binarized spam data. What is the training and test error? (You can try different settings of the pseudocount α if you like (this corresponds to the $\mathrm{Beta}(\alpha, \alpha)$ prior each θ_{jc}), although the default of $\alpha = 1$ is probably fine.) Turn in your error rates.

2. Modify the code so it can handle real-valued features. Use a Gaussian density for each feature; fit it with maximum likelihood. What are the training and test error rates on the standardized data and the log transformed data? Turn in your 4 error rates and code.