

Use MATLAB for question 2, and hand in the code along with your results. Note that the assignment is due IN CLASS.

1. Suppose that we have three colored boxes  $r$  (red),  $b$  (blue), and  $g$  (green). Box  $r$  contains 3 apples, 4 oranges, and 3 limes; box  $b$  contains 1 apple, 1 orange, and 0 limes; box  $g$  contains 3 apples, 3 oranges, and 4 limes. Boxes are chosen at random with probabilities  $p(r) = 0.2$ ,  $p(b) = 0.2$ ,  $p(g) = 0.6$ , and there is an equal probability of selecting any of the items in a box. Compute the contingency table for this example. What is the probability of selecting an apple? If we observe that the selected fruit is an orange, what is the probability that it came from the green box? What is the KL-divergence between the distribution of choosing a box given that an apple has been chosen, and the distribution of choosing a box given that an orange has been chosen? What is the mutual information  $I(B; F)$ , where  $B$  corresponds to the boxes and  $F$  to the fruits?
2. This exercise will examine differences between ridge regression and lasso regression. Download an implementation of lasso from [www2.imm.dtu.dk/pubdb/views/publication\\_details.php?id=3897](http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897). The function `lars.m` performs lasso regression over varying  $\lambda$ . To run on a data set  $X$  with labels  $y$ , run the command `beta = lars(normalize(X), center(y), 'lasso')`.

The script `lars_test.m` produces plots similar to Figures 3.7 and 3.9. Download the prostate cancer data set from [www-stat.stanford.edu/ElemStatLearn](http://www-stat.stanford.edu/ElemStatLearn) and modify `lars_test.m` to reproduce a plot similar to Figure 3.9 for the training data of the prostate data set. Note that in the prostate data, columns 2 to 9 are the data variables, column 10 is the output variable, and column 11 indicates whether the data point is training or test.

Repeat this experiment with ridge regression. Write a function `ridgeregression.m` that produces beta vectors for `lambda = [100 90 80 70 60 50 40 30 20 10 1 .1 .01 0]` (do not use Matlab's existing ridge regression function). Plot the coefficients and compare the results to the lasso results. Also compute the RSS error on the test set for both regression methods.

Also run this experiment on the South African heart disease data set and the waveform training data set, produce coefficient plots for these data sets, and discuss the results. For waveform, download the test data and compare prediction accuracy on the test set between the two methods.