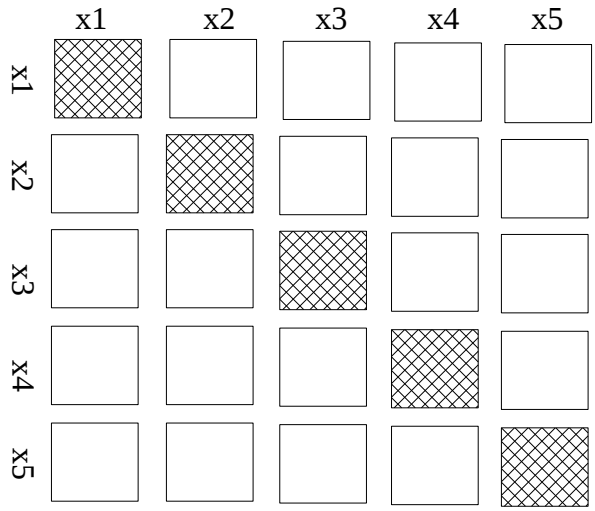
Assignment 2

Problem 1: Find the expectation, the variance and the standard deviation for the random variable X taking instances as in file: problem1.dat

Problem 2: The data in file problem2.dat represents a training set of size 300 samples with feature vectors $\in R^{20}$. Find the covariance matrix of the set's features. Note: every 300 consecutive readings from the file represent one row in X. (Hint: build a 20x300 matrix, X, from the data given then find the covariance matrix for it).

Problem 3: A sample of size 300 with feature vectors $\in R^5$. Plot a scatter diagram for each pair of the features (i.e. (x1, x2), (x1, x3), (x1,x4), (x1,x5), (x2,x3), (x2,x4),... and so on (total 10). Then find the correlation coefficient rho for each pair. In your opinion, which pair, of the 10 pairs, can be represented by a line?
Notes:
(1) the corresponding data is found in file problem3.dat. Every 300 consecutive readings from the file represent one row in X. (Hint: build a 5x300 matrix, X, from the data given).
(2) For better readability of the scatter diagram you should plot them in a tiled layout as:



Problem 4: Using the data in problem3, we now assign the vector v in file problem4.dat to each vector of the matrix X (from problem3). Vector v contains the label of the class assigned to each vector of X (we have only two classes R, B).
(1) Re-plot the scatter diagram (of problem3) using color code to represent the class labels.
(2) run QDA on each pair to find the decision surface (i.e. based on two features) (This means that you will get 10 different decision surfaces).
(3) Now run QDA on all features (5 features) and come up with only one decision surface.

Problem 5: (a) Show that Bayes classifier is the best classifier that could be used. What is the problem with this classifier?
(b) File problem5.dat contains 5 test vectors t1 to t5 $\in R^5$ (same as before; every 5 consecutive readings from the file represent one row in matrix T. (Hint: build a 5x5 matrix, T, from the data given and use its 5 vectors as t1, t2, t3, t4, t5). According to the decision surface found in problem4 – (3), use the decision surface to classify each of the vectors t1 to t5 to one of the classes {R,B}.