MLDM Assignment

Theoretical question (no computers necessary):

1. (25%) Show that the maximum likelihood estimates for a univariate Normal distribution with unknown mean and variance are given by:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})\,(x_i - \hat{\mu})^T$$

Question with data and software

2. (75%) Load data files: M1, M2, Sigma1, and Sigma2 (all are .csv files). These files give data for two 6-d normal distributions $(\mu_1, \Sigma_1, \mu_2, \Sigma_2)$ with $P_1 = 0.35, P_2 = 0.65$.
    a. Generate 10,000 observations from the two distributions, proportionate to the *a priori* probabilities, which will be the training set.
    b. Compute the MLE estimators for each of the class conditional parameters. Compare them to the true values.
    c. Generate another set, with 2,000 observations (this will serve as validation set).
    d. Fit a random forest to the data. Use the validation set to compare a number of forest configurations and choose the best performing one. Then use CV-10 over the training set to estimate the model accuracy and generalization error. (You may not use existing functions for the cross-validation for optimization and estimation part but write your own).
    e. Fit a neural network to the data. Use the validation set to determine the number of neurons to use for the network. After choosing the number of neurons, use CV-10 over the training set to estimate the classification accuracy.
    f. Choose one of the two models above. We will now consider the overfitting phenomenon as a function of training set size. Fit the model with training sets of size $N = 10, 20, 30, \ldots, 1{,}000$. Plot the test and training error as a function of N. *For estimating test error, use the validation set.