Problem 1 (20 pt) Regularized Logistic Regression

Let \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) be the training examples, where \( x_i \in \mathbb{R}^d \) and \( y_i \in \{-1, +1\} \). The negative log-likelihood function of the regularized logistic regression, denoted by \( \mathcal{L}(w) \), is written as

\[
\mathcal{L}(w) = \sum_{i=1}^{n} \ln \left(1 + \exp(-y_i x_i^\top w)\right) + \frac{\lambda}{2} \|w\|^2
\]

where \( \lambda \) is the regularization coefficient. The optimal solution for \( w \) is obtained by minimizing \( \mathcal{L}(w) \).

- Show if two attributes \( k \) and \( l \) are identical, i.e., \( x_{i,k} = x_{i,l} \) for any training example \( x_i \), we will have \( w_k = w_l \) for the optimal solution \( w \).

- Build and test a regularized logistic regression model using the heartstatlog dataset that can be downloaded from [http://www.cse.msu.edu/~cse847/assignments/heartstatlog.rar](http://www.cse.msu.edu/~cse847/assignments/heartstatlog.rar)

1. Implement the regularized logistic regression model by using the gradient descent method discussed in the class. The stepsize \( \eta_t \) of \( t \) iteration is set to be \( \eta_t = 1/\sqrt{t} \). Terminate the iterative algorithm when the norm of the gradient is less than 0.01. Submit your code.

2. Train the regularized logistic regression model over the training data specified in files `heartstatlog_trainSet.txt` and `heartstatlog_trainLabels.txt`. Use the cross validation method to decide the value of the regularization parameter \( \lambda \). Randomly select 80\% of training data to form the training set, and use the remaining 20\% of the training data for validation. Set the candidate values for \( \lambda \) to be \( 0.01, 0.05, 0.25, 1, 5, 25, \) and \( 100 \). For each \( \lambda \), report the training error, validation error, and test error. Email your code and results to TA.