CS 567 Project Documentation
Sparsh Agarwal
9075905142

The project consists of 5 sections:

1. Setup
   This section is used for placing the folder paths, reading the image files and labels. This also defines the number of folds of cross validation the user wants to use (It should be a divisor of 18 to keep the labels balanced in the folds). 3 is found to be the optimal number of folds because it creates enough folds to get diversity in the data while maintaining the minimum optimal number of elements in each segment to get generic applicable result. With 3 fold validation each fold has at least 12 elements.
2. Image Processing
   a. Distributed Data
   b. NotDistributed Data (Hidden Data)

   The process used to process both the sets of data is same. It is as follows:
   - Only green channel is used from the three channels because it has the maximum useful information to extract the lesions.
   - A mask is made from the image to find the probable area where the lesions may occur. This is chosen to be the area where the intensity value is greater than 15% of the maximum intensity in image. Then this mask is further eroded functions strel (which creates a disk of radius 35) and imerode which uses the disk created by strel to erode the edges of the mask so that we can avoid any discrepancies in features at the edges of eyes.
   - Then a median filter of 9 by 9 is used to remove the gradient in images. This is the optimal size of filter because given the data, it removes the gradient without adversely affecting the required lesions. This is done using medfilt2 which takes in a filter size as argument and performs median filtering on the specified image
   - Next to binarize the image 73.5% of max intensity in processed image is found out to be the best value to retain the lesions and remove any unwanted content
   - The now found segments may still by very small and may also belong to the same lesion. Dilation is used to join the small segmented lesion points to create combined lesions.
   - The isolated outliers which are not lesions may also get dialated, to remove them we use erosion with a disk radius of 1 more than the disk used for dilation. This will not remove the combined lesions but will remove the outliers.
   - The previous two steps collectively remove the outliers and unwanted erroneous lesions while protecting the lesion segments which might be too small to survive the erosion, by combining them with their neighbors.

- To further amplify the probable lesions and combine any close enough neighbors, we use dilation. 9 is used as the optimal radius for this dilation because if a larger radius is used, it may join the different lesions which are supposed to be separate otherwise.
- After extracting the lesions, the mask created before is used to ensure that the obtained lesions are in the expected region of eye and are not outside the eye. By multiplying the mask, any lesions which are outside the required are will be eliminated.
- This entire process gives us the probable lesions and removes any outliers as well.
- The two features generated from these lesions are:
    - Number of lesion segments in each image
      Bwconncomp function is used to calculate the number of connected segments in the image. It returns an attribute NumObjects
      Which specifies the number of segments present in the image.
    - The total area percentage occupied by lesions in each image
      SInce image is binarized, this feature can be calculated by finding the percentage area occupied by the indexes where image intensity is greater than zero.

3. Classification (Only on distributed data using cross validation)
    a. Classifier 1: KNN
       After comparing results from 3,5,7,9 nearest neighbor knn classification, 7 is found to provide an optimal accuracy for classifying the testing fold.
    b. Classifier 2: Logistic Regression
       It uses the same process as we learned in class

   Cross validation is used to create training and testing sets from the same distributed data. CV requires fold creations. Well balanced and randomized folds are created by permuting the folds in both label section and then combing the sections. This generates a well randomized and balanced fold set which can be used in the classifiers.
   Further for classifiers, we need to normalize the features. For proper scaling, both the training and testing data are scaled according to mean and standard deviation of the training data. To randomize the classification in case of a tie, permutation of the possible classification values is used.
   Both the classifiers use the process we learned in class.
4. Classification (Of Hidden Data by using distributed data as training data)
    a. Classifier 1: KNN
       After comparing the result obtained from all possible odd numbers nearest neighbors, 29 neighbors is found to be optimal when testing the distributed data by a knn classifier trained from distributed data.

b. Classifier 2: Logistic Regression
   It uses the same process as we learned in class

No cross validation is required for this section but we still need to normalize the features and randomization in case of a tie. Here

5. Result
   Even though each classifier prints the results. This section can be used to print any extra required final results. The current code provides the inclass accuracies of both classes for both classifiers and for both sections(distributed data, both distributed and hidden data).

   After averaging over results obtained from 5 iterations of the program, below are the obtained accuracies

   Distributed Class 0 (Healthy retina) accuracy using KNN(k = 7)= 83.332%.
   Distributed Class 1 (Unhealthy retina) accuracy using KNN(k = 7)= 65.558%.
   Distributed Class 0 (Healthy retina) accuracy using Logistic regression= 78.89%.
   Distributed Class 1 (Unhealthy retina) accuracy using Logistic regression= 63.334%.
   NotDistributed Class 0 (Healthy retina) accuracy using KNN(k = 29)= 100.000000%.
   NotDistributed Class 1 (Unhealthy retina) accuracy using KNN(k = 29)= 77.78%.
   NotDistributed Class 0 (Healthy retina) accuracy using Logistic regression= 77.78%.
   NotDistributed Class 1 (Unhealthy retina) accuracy using Logistic regression= 66.67%.