Machine Learning

Homework 3

Due Date: Dec. 29, 2022

Acromegaly results in a 72% increase in all-cause mortality compared to the general population, which is due to over-secretion of growth hormone (GH) stimulating production of the insulin-like growth factor-1 (IGF-1) from the liver. Over 95% of acromegaly results from a GH-secreting pituitary adenoma composed of somatotroph cells. Manifestations are caused by central compression effects, leading to headache, visual defect, and peripheral actions to exhibit soft tissue growth and metabolic dysfunction, including large fleshy lips and nose, spade-like hands, frontal skull bossing, enlarged tongue, bone, thyroid, heart, liver, and spleen, diabetes mellitus (DM), hypertension, and heart failure. These changes are so slow and insidious that acromegaly usually has a delayed diagnosis after about 6-10 years. Better clinical, economic, and health-related quality of life may be attained if acromegaly can be diagnosed and controlled. Several computer-aided diagnosis (CAD) approaches using 2D photographs or 3D stereo-photographs have been shown to be promising in differentiating acromegaly patients from normal ones.

As in Homework 2, a set of features extracted from the 3D stereo-photographs of 103 subjects' faces were listed in "AcromegalyFeatureSet.xlsx", including 39 males (gender: 1) and 64 females (gender: 2), and 41 positives (GroundTruth: 1) and 62 negatives (GroundTruth: 0). Suppose $\boldsymbol{X}$ denotes a $103 \times d$ matrix, in which each row of $\boldsymbol{X}$ contains the features of a subject and d is the number of features.

In this homework, you are required to develop three classification models to differentiate acromegaly positive from acromegaly negative cases based on logistic regression function (LRF), artificial neural network (ANN) and support vector machine (SVM), respectively. To assess the performance of each classification model, you need to carry out a stratified five-fold cross-validation method for each model. More specifically, all 103 cases are to be divided into 5 groups such that all groups have approximately the same number of positive cases (i.e., 8 or 9 cases) as well as the same number of negative cases (i.e. 12 or 13 cases). Then, each group takes turn to serve as the test data and the other four groups as the training data.

1. Feature Selection/Dimension Reduction:

   For each classification model, you need to describe your feature selection or dimension reduction process, including

   a. Feature selection or dimension reduction methods
   b. How you determine the number of features
   c. The number of features used in each of the five folds.

2. Classification:

   For each classification model, you need to

   a. Describe how you determine the model parameters if applicable. For example, you need to describe how you determine the number of layers and the number of neurons in each layer for the ANN.
   b. Report the accuracies, sensitivities and specificities for each of the 5 folds and their mean$\pm$std.

3. Discussions:

   Discuss the methods you have employed, e.g., feature selection/dimension reduction, classification, and the performances of these methods.

Note. You need to submit your codes along with the answers. Your codes should be executable by TAs to reproduce your results. List all performances in a Table for ease of comparison.