Machine Learning

Homework 3

Due Date: Dec. 31, 2025

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[2]. 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.

In this homework, 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. Note that "Gender" is not considered as a feature in this homework.

To assess the performance of each classification model, you need to carry out leave-one-out cross-validation (equivalently, 103-fold cross-validation), in which each data takes turn to serve as the test data and the others data as the training data. The leave-one-out cross-validation method will produce a probability output in each fold. As a result, you will have 103 probability values, each for a test data.

1. Use Logistic Discrimination Function, ANN and SVM to construct the classification models. For each classification model, you need to (75%)
   a. Describe what features you used for the classification models and how you derived these features. For example, if you used forward selection, state clearly how the forward selection was realized.
   b. Describe how you determined the model parameters if applicable. For example, you need to describe how you determined the number of layers and the number of neurons in each layer for the ANN.
   c. Use the 103 probability values to report the test performance figures, including accuracy, sensitivity, specificity, and the area under curve (AUC) of your ROC curve. Plot the ROC curve.
2. Discuss and compare the performances achieved by the methods implemented in Homeworks 1, 2 & 3. (25%)

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