Machine Learning

Homework 2

Due on Nov. 27, 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.

Use a Bayesian decision model as the classifier for all questions in this homework. As in Homework 1, leave-one-out cross-validation is used for performance assessment.

Reports:

You are supposed to submit a written report as well as your code. Make sure that your code can be executed by TAs to reproduce all results mentioned in your written report. In your written report,

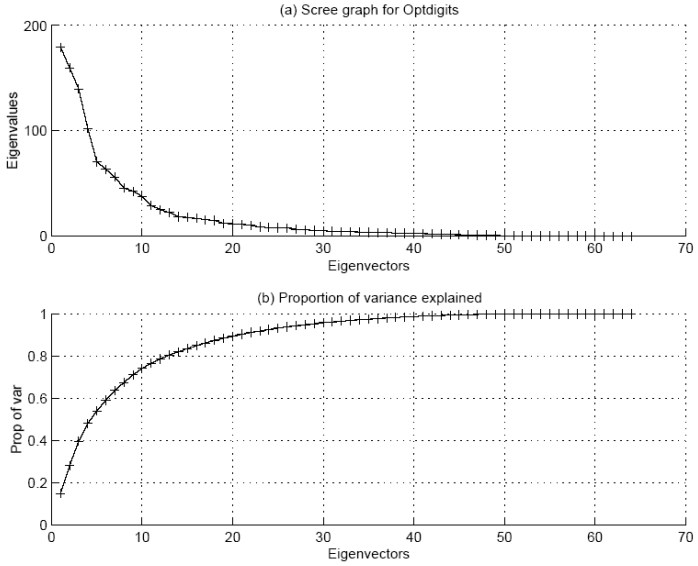
1. **For each fold of leave-one-out cross-validation**, use PCA to find the eigenvectors of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, where $\boldsymbol{X}$ denotes a $102 \times d$ matrix of the training data, in which each row of $\boldsymbol{X}$ contains the features of a training subject and d is the number of features. Use the eigenvectors of the largest k eigenvalues of the training data to form a k-dimension feature space for dimension reduction. Project the feature vectors of the training data on to the k-dimension feature space and use the projected feature vectors of the training data to construct your Bayesian decision model. Assess the performance of the Bayesian decision model using the projected test data. (Remember to project the original test data to the k-dimension feature space before performance assessment.)

    a. Report the numbers of eigenvalues selected for each of the 103 folds and clearly describe how you determined these 103 k's.

    b. 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. **Use ALL 103 data** as the training data to perform PCA on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and accomplish the following:

   a. Plot the eigenvalues in the descending order as shown below. Also, plot the proportion of the variance explained by the first m eigenvectors as shown below, where $m = 1, \ldots, d$.



   b. List the largest five eigenvalues of your PCA results.

   c. Project all data on to the 2D space spanned by the eigenvectors of the largest two eigenvalues, denoted by PC1 and PC2. Use **ALL** data projected on PC1 and PC2 to construct a bivariate Gaussian Bayesian decision model. Illustrate and provide a figure on which
      - illustrate the 2D scatter plot of the projected data with PC1 and PC2 as the horizontal and vertical axes, respectively, and denote the projected data of classes 1 and 0 by two different symbols, e.g., plus signs and circles,
      - for each class, plot the contour map, i.e., each contour line joins the points of the same probability values, and
      - plot the decision boundary.