Machine Learning

Homework 1

Due on Oct. 30, 2025.

Problem 1: cross-validation, feature selection, and classification

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.

Please use Bayesian decision models for feature selection and classification with the assumption that all features are jointly distributed according to a multivariate Gaussian distribution. In this homework, the discriminant function for each class in a Bayesian decision classier is modeled as a multivariate Gaussian function. For performance assessment, 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.

For each fold, i.e., for each test data, use forward selection to select a subset of features. Use the selected features to construct your multivariate Gaussian Bayesian classifier based on the training data and assess its performance using the test data.

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,

- Describe the multivariate Gaussian models used for each class and how you estimate the model parameters.
- Describe your forward selection procedure, your cost function for feature selection (i.e., what you try to optimize ?) and how you determine the number of features to be selected.
- List the names of the features selected by your approach (note that each data may have a different set of selected features).
- 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.
- Report the top 2 most frequently selected features, i.e. the two features with the highest number of times selected in the 103 folds. If there is a tie among *k* features, i.e., they are selected with the same number of times, please propose a strategy to resolve the tie preferably with an argument to justify your strategy. Use **ALL** data and the top 2 features to construct a bivariate Gaussian Bayesian decision model. Illustrate and provide a figure on which
    - plot the scatter plot using these top 2 features with classes 1 and 0 represented 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.

Problem 2. Proof of Bayesian estimator.

Suppose $x^t \sim N(\theta, \sigma^2)$ and $\theta \sim N(u_0, \sigma_0^2)$, where $u_0, \sigma_0^2, \sigma^2$are known. That is

$$\begin{aligned} p(X|\theta) &= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{\sum_t (x^t - \theta)^2}{2\sigma^2}\right] \\ p(\theta) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right] \end{aligned}$$

Please show that

$$E[\theta|X] = \frac{N/\sigma^2}{N/\sigma^2 + 1/\sigma_0^2} m + \frac{1/\sigma_0^2}{N/\sigma^2 + 1/\sigma_0^2} \mu_0$$

where *m* is the maximum likelihood estimator of the sample mean.