Biological Data Analysis (CSE 182) : Assignment 5

Logistics

See assignment notes on the course home page. Please cite all sources, and while you can collaborate and discuss, you must write the assignment yourself.

Supervised Classification (Gene Expression Matrix)

1. Given a hyperplane L defined by the vector $\beta = [-0.15, 0.9, 0.05, -0.02]^T$, and $\beta_0 = 0.88$, calculate the distance of each point in the input file 'expression.txt' from the hyperplane. Find the point that is closest to the hyperplane. [25 pts]

2. For the hyperplane L in (1), calculate the number of misclassified points. What is the classification error (defined as the sum of the absolute distances of the misclassified points from the hyperplane)? [10 pts]

3. We want to find the hyperplane that minimizes the classification error. A commonly used approach for optimization is *grid search* or an exhaustive search over the space of possible solutions. In a grid search, we discretize possible assignments to β, and test the optimization function over all discrete assignments. Implement a grid search based strategy to find the best hyperplane defined by $\beta = [\beta_1, \beta_2, \beta_3, \beta_4]^T$ and β_0. Grid search can be slow for multiple parameters, therefore, use the following to reduce the computation time:

   - Maintain $||\beta||_2 = 1$. Therefore $-1 \leq \beta_i \leq 1$ for $1 \leq i \leq 4$.
   - For the grid search, use a step size of 0.1. After changing β, re-normalize by dividing by $||\beta||_2$. (**Note.** This procedure does not give you equally spaced grid-points on the unit hypersphere, but will suffice for this assignment).
   - assume that $\beta_0 \leq 1$

   [45 pts]

4. Given a vector $\beta = [-0.4, 0.45, 0.01, 0]^T$, calculate the F function (Fisher's LDA) using the data in "expression.txt". Repeat the calculation for $\beta = [-0.15, 1.0, 0.1, 0.2]^T$. Which of these two vectors is a better choice for clustering the points? [20 pts]

Input files

- "expression.txt": Gene expression matrix for four genes (columns) and 53 individuals (31 individuals with label '-' and 22 individuals with label '+').