CS/ECE/ME532 Period 18 Activity

Estimated time: 15 mins for P1, 20 mins for P2, 15 mins for P3, 20 mins for P4

1. A breast cancer gene database has approximately 8000 genes from 100 subjects. The label y_i is the disease state of the ith subject (+1 if no cancer, -1 if breast cancer). Suppose we build a linear classifier that combines the 8000 genes, say $\boldsymbol{g}_i, i = 1, 2, \ldots, 100$ to predict whether a subject has cancer $\hat{y}_i = \text{sign}\{\boldsymbol{g}_i^T \boldsymbol{w}\}$. Note that here $\boldsymbol{g}_i$ and $\boldsymbol{w}$ are 8000-by-1 vectors. You recall from the previous period that the least-squares problem for finding classifier weights has no unique solution.

Your hypothesis is that a relatively small number of the 8000 genes are predictive of the cancer state. Identify a regularization strategy consistent with this hypothesis and justify your choice.

2. Consider the least-squares problem $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$ where $\boldsymbol{y} = 4$ and $\boldsymbol{X} = \begin{bmatrix} 2 & 1 \end{bmatrix}$.

a) Does this problem have a unique solution? Why or why not?

b) Sketch the contours of the cost function $f(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$ in the $w_1 - w_2$ plane.

c) Now consider the LASSO $\min_{\boldsymbol{w}} ||\boldsymbol{w}||_1$ subject to $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 < 1$. Find the solution using the following steps

i. Repeat your sketch from part **b)**.

ii. Add a sketch of $||\boldsymbol{w}||_1 = c$

iii. Find the $\boldsymbol{w}$ that satisfies $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 = 1$ with the minimum possible value of $||\boldsymbol{w}||_1$.

d) Use your insight from the previous part to sketch the set of solutions to the problem $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1$ for $0 < \lambda < \infty$.

3. The script provided has a function that will compute a specified number of iterations of the proximal gradient descent algorithm for solving the ℓ_1-regularized least-squares problem

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1$$

The script will get you started displaying the path taken by the weights in the proximal gradient descent iteration superimposed on a contour plot of the squared error surface for the cost function defined in problem **2.** part **b)** starting from $\boldsymbol{w}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The script assumes $\lambda = 4$ and $\tau = 1/4$.

Include the plots you generate below with your submission.

a) How many iterations does it take for the algorithm to converge to the solution? Use the data cursor to find the converged value for $\boldsymbol{w}$.

b) Change to $\lambda = 2$. How many iterations does it take for the algorithm to converge to the solution? Use the data cursor to find the converged value for $\boldsymbol{w}$.

c) Explain what happens to the weights in the regularization step.

4. Use the proximal gradient algorithm to solve $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + 4||\boldsymbol{w}||_1$ for the parameters defined in problem **2.**

a) What is the maximum value for the step size in the negative gradient direction, τ?

b) Suppose $\tau = 0.1$ and you start at $\boldsymbol{w}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Calculate the first two complete iterations of the proximal gradient algorithm and depict $\boldsymbol{w}^{(0)}, \boldsymbol{z}^{(1)}, \boldsymbol{w}^{(1)}, \boldsymbol{z}^{(2)}$ and $\boldsymbol{w}^{(2)}$ on a sketch of the cost function identical to the one you created in problem **2.b)**.