Homework 3, due September 27th, 11:59pm

September 21, 2023

1. Implement the Logistic Regression learning by gradient ascent as described in class. Before using logistic regression, be sure to normalize the variables of the training set to have zero mean and standard deviation 1, and to use the exact same transformation on the test set, using the mean and standard deviation of the training set.

a) Using the `Gisette` data, train a logistic regressor on the training set, starting with $\mathbf{w}^{(0)} = 0$, with 300 gradient ascent iterations and shrinkage $\lambda = 0.0001$ in the update equation:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\lambda\mathbf{w}^{(t)} + \frac{\eta}{N}\frac{\partial}{\partial \mathbf{w}}L(\mathbf{w}^{(t)})$$

where $L(\mathbf{w}^{(t)})$ is the log likelihood from page 7 of the Logistic Regression slides. Observe that there is an extra factor of $1/N$ in the loss term compared to the class notes.

Find a good learning rate η such that the log-likelihood converges in at most 300 iterations and is monotonically increasing. Plot the training log-likelihood vs iteration number. Report in a table the misclassification error on the training and test set. On the same graph, plot the Receiver Operating Characteristic (ROC) curve of the obtained model on the training and test set. (2 points)

b) Repeat point a) on the `hill-valley` dataset, where you might need as many as 10,000 iterations for it to converge. (2 points)

c) Repeat point a) on the `dexter` dataset. (2 points)

2. For the `Gisette` data, minimize by gradient descent the L_1-penalized logistic loss:

$$C(\mathbf{w}) = -\frac{1}{N}L(\mathbf{w}) + \lambda\sum_{i=1}^{p}|w_i|,$$

by gradient descent starting from $\mathbf{w}^{(0)} = 0$, where $\lambda = 0.01$. Here $L(\mathbf{w})$ is again the log likelihood from page 7 of the Logistic Regression slides.

Find a good learning rate η such that the loss converges in at most 300 iterations and is monotonically decreasing. Again, be sure to normalize the variables of the

training set to have zero mean and standard deviation 1, and to use the exact same transformation on the test set, using the mean and standard deviation of the training set.

a) Plot the training loss vs iteration number. Report in a table the misclassification error on the training and test set. (1 point)

b) On the same graph, plot the Receiver Operating Characteristic (ROC) curve of the obtained model on the training and test set. (1 point)

c) How many nonzero entries are in $\mathbf{w}$? How many values in $\mathbf{w}$ satisfy $|w_i| > \lambda$? (1 point)