Sharik Purkar

Machine Learning

Assignment 3

Question 1

a. After coding an extra column of ones in the x matrix, as well as without, I can report no observable changes in the accuracy. This is likely due to the columns' lack of an effect on Naïve Bayes since the mean of the ones' column will be one and it will lack a standard deviation (0) therefore should not have any positive boosts for the accuracy of the model.
b. In code
c. In code
d. Here are a few observations from running the script a few times
    a. Neural Networks were almost always the most accurate algorithm
    b. Surprisingly enough, Naïve Bayes was thoroughly outperformed by Linear Regression
    c. A combination of negative and positive floats performed better
    d. Neural Nets and Logistic Regression benefited largely for larger datasets.

Question 2

$$P(y=1 \mid x_i, w) = \frac{1}{2}\left(1 + \frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)$$

$$P(y=0 \mid x_i, w) = 1 - \frac{1}{2}\left(1 + \frac{w^T x_i}{\sqrt{1+(w^T x_i)^2}}\right)$$

If,

$$\sin\theta = \frac{w^T}{\sqrt{1+(w^T x_i)^2}}$$

Likelihood estimation:

$$\prod_{i=1}^{n} \left[\left(\frac{1}{2}(1+\sin\theta)\right)^{y_i}\left(1-\frac{1}{2}(1+\sin\theta)\right)\right]^{1-y_i}$$

$$\prod_{i=1}^{n} \left[\left(\frac{1}{2}(1+\sin\theta)\right]^{y_i}\left(\frac{1}{2}[1-\sin\theta]\right)\right)^{1-y_i}$$

$$l(w) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1+\sin\theta}{2} \right) + (1-y) \log \frac{(1-\sin\theta)}{2} \right]$$

$$\tan\theta = w^T x_i$$

$$\sec^2\theta \frac{\partial\theta}{\partial w_j} = x_{ij}$$

$$\frac{\partial\theta}{\partial w_j} = x_{ij} \cos^2\theta$$

$$\frac{\partial l\, l(w)}{\partial w_j} = \sum_{i=1}^{n} x_{ij} \left[y \cos\theta (1-\sin\theta) - (1-y) \cos\theta (1+\sin\theta) \right]$$

$$= \sum \cos\theta \left(2y - (1+\sin\theta) \right)$$

$$\frac{\cos^2\theta}{1+\sin\theta} = 1 - \sin\theta \quad \& \quad \frac{\cos^2\theta}{1-\sin\theta} = 1 + \sin\theta$$

$$= \sum_{i=1}^{n} x_{ij} \left[2y_i - \left(1 + \frac{x^T w}{\sqrt{1 + (x^T w)^2}} \right) \right] \frac{1}{\sqrt{1 + (x^T w)^2}}$$

Question 3

a. L1 Regularizer: lambda*|w|
   a. Lamba is constant
   b. |w|: Absolute value of the weights
   c. Adding it to the loss function will provide the L1 regularized loss penalty

New Derivative when added regularizer to logistic regression equation

Delta = X^T(Y-p) + lambda * (sign of w)

Implementing Lambda for L1 regularizer

a. Initialize weights randomly
b. Apply gradient descent
c. Observe fluctuating weight value
d. Decrease step size accordingly
e. Divide value of derivative by the Xtrain size
f. Multiply weight value by average value of training data
g. Run the function for different lamba values

b. Proceeding with L Infinity as the regularizer, add max of weights as the regularizer to the weight matrix.

Equation: **W = W + alpha * (XTW(Y-p)) + lambda * max(W))/ (size of X)**

Implementing Lambda for L3 regularizer

a. Initialize weights randomly
b. Apply gradient descent
c. Observe fluctuating weight value
d. Decrease step size accordingly
e. Divide value of derivative by the Xtrain size
f. Multiply weight value by average value of training data
g. Run the function for different lambda values

c. L1 regularization while tremendously useful does have clear limitations. Such as it’s inability to be universally applied as the second derivative of L1 turns out to be 0. Unlike L2, which can be performed with Newton Raphson, L1 cannot because of hessian terms providing matrix errors.