COL341 Machine Learning Assignment - 1 (Part 2)

Logistic Regression

Part(a)

We have to perform logistic regression using gradient descent method to learn the optimum decision surface. The log-likelihood for logistic regression can be written as:

$$L(\theta) = \sum_{i=1}^{m} t^{(i)} \log(h_\theta(x^{(i)})) + (1 - t^{(i)}) \log(1 - h_\theta(x^{(i)})) + \frac{\lambda}{2}||w||^2$$ and $$h_\theta(x) = \frac{1}{1+e^{-w^T x+b}}$$.

The loss function can be written in the matrix form as:

$$\frac{1}{m}(-(y^T \log(\phi(X\theta))) - (1 - y^T)(\log(1 - \phi(X\theta))) + \frac{\lambda}{2m}\theta^{1T}\theta$$

The gradient in the matrix form can be written as:

$$\frac{1}{m}((\phi(X\theta) - y)^T X)^T + \frac{\lambda}{m}\theta^1$$

The equation to learn the decision surface is as follows:

$$W' = W - rate * gradient$$

(i) We learn the decision surface with constant learning rate.

(ii) We learn the decision surface with adaptive learning rate. $n^t = n^0/\sqrt{t}$, where t is the iteration number and n^0 denotes initial learning rate and n^t denotes learning rate after ith iteration.

(iii) We learn the decision surface with adaptive learning rate using adaptive line search algorithm.

Part(b)

In this part we use stochastic gradient descent with a batch size of 128. This method works same as the last but we modify the w_0 using the gradient of one batch and then modify w_0 using all the batches one by one.

(i) We learn the decision surface with constant learning rate.

(ii) We learn the decision surface with adaptive learning rate. $n^t = n^0/\sqrt{t}$, where t is the iteration number and n^0 denotes initial learning rate and n^t denotes learning rate after ith iteration.

(iii) We learn the decision surface with adaptive learning rate using adaptive line search algorithm.

Part(c)
https://courses.cs.washington.edu/courses/cse599c1/13wi/slides/l2-regularization-online-perceptron.pdf
The above link helped me in finding the stopping criteria and how to find it. When the difference between two consecutive log-likelihood is less than normalized gradient then we stop iteration. This is the stopping criteria.

$$\ell(\mathbf{w}^*) - \ell(\mathbf{w}) \leq \frac{1}{2\lambda}||\nabla\ell(\mathbf{w})||_2^2$$