CS189/289A – Spring 2017 — Homework 4

Yicheng Chen, SID 26943685

1. Logistic Regression with Newton's Method

(a) The gradient of the cost function is:

$$\nabla_w J(w) = 2\lambda w - \sum_{i=1}^{n}(X_i y_i(1-s_i) - X_i(1-y_i)s_i) = 2\lambda w - X^\top(y-s)$$

(b) The Hessian is:

$$\nabla_w^2 J(w) = 2\lambda I + X^\top \Omega X$$

Where Ω is the diagnoal matrix of $s(i)$.

(c) The update equation for Newton's method:

$$e \leftarrow \text{solution to } (2\lambda I + X^\top \Omega X)e = 2\lambda w - X^\top(y-s)$$

$$w \leftarrow w + e$$

(d) (1) $s^{(0)} = [0.9546, 0.7311, 0.7311, 0.2689]$

(2) $w^{(1)} = [-2.7496, 0.5472, 1.5374]$

(3) $s^{(1)} = [0.9600, 0.6057, 0.8894, 0.3396]$

(4) $w^{(2)} = [-3.8843, -0.1456, 3.7334]$

2. l_1- and l_2-Regularization

(a) Since $X^\top X = nI$, The cost function:

$$J(w) = w^\top X^\top X w + y^\top y - 2y^\top X w + \lambda |w|_1 = w^\top w + y^\top y - 2y^\top X w + \lambda |w|_1$$

Let $g(y) = y^\top y$, then

$$f(X_{*i}, w_i, y, \lambda) = w_i^2 - 2y^\top X_{*i} w_i + \lambda |w_i|_1$$

Therefore $J(w)$ can be written as the form of g and f.

(b) If $w_i^* > 0$, then

$$w_i^* = \frac{2y^\top X_{*i} - \ lambda}{2}$$

(c) If $w_i^* < 0$, then

$$w_i^* = \frac{2y^\top X_{*i} + lambda}{2}$$

(d) For w_i^* to be zero, $\lambda = 0$ and $y^\top X_{*i} = 0$

(e) Similar to (a),

$$f(X_{*i}, w_i, y, \lambda) = w_i^2 - 2y^\top X_{*i} w_i + \lambda w_i^2$$

$$w_i^* = \frac{2y^\top X_{*i}}{2(\lambda + 1)}$$

Therefore, when $y^\top X_{*i} = 0$ and $\lambda \neq 1$, $w_i^* = 0$.

3. Regression and Dual Solutions

(a)

$$\frac{\partial}{\partial w_i}|w|^4 = 4(\sum w_i^2)w_i$$

$$\nabla|w|^4 = 4w^\top ww$$

Then

$$\nabla|Xw-y|^4 = 4(Xw-y)^\top(Xw-y)X^\top(Xw-y)$$

(b) From (a), we know that

$$\nabla J(w) = 4(Xw-y)^\top(Xw-y)(Xw-y)^\top X + 2\lambda w$$

Setting it to be zero,

$$4(Xw^*-y)^\top(Xw^*-y)(Xw-y)^\top X + 2\lambda w^* = 0$$

Therefore,

$$w^* = -\frac{2}{\lambda}(Xw^*-y)^\top(Xw^*-y)(Xw-y)^\top X$$

So vector $a = -\frac{2}{\lambda}(Xw^*-y)^\top(Xw^*-y)(Xw-y)^\top$.

(c)

$$\nabla J(w) = \frac{1}{n}\sum_{i=1}^{n} L'(w^\top X_i, y_i)X_i + 2\lambda w$$

If L is convex in its first argument, because $\lambda|w|^2$ is also convex in the domain of w, then $\nabla J(w^*) = 0$. Therefore,

$$w^* = \sum_{i=1}^{n} -\frac{1}{2\lambda}\frac{1}{n}L'(w^\top X_i, y_i)X_i = \sum_{i=1}^{n} a_i X_i$$

Even though the loss function is not convex, $\nabla J(w^*) = 0$ is still true, so w^* also has the same form.

4. Franzia Classification + Logistic Regression = Party!

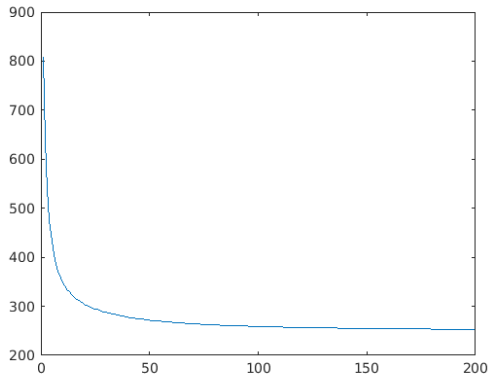
(a) The gradient of logistic regression with l_2 regularization is:

$$\nabla_w J = -X^\top (y - s(Xw)) + 2\lambda w$$

So the batch gradient descent update equation is:

$$w^{t+1} = w^t - \epsilon \nabla_w J = w^t - \epsilon[-X^\top (y - s(Xw)) + 2\lambda w]$$

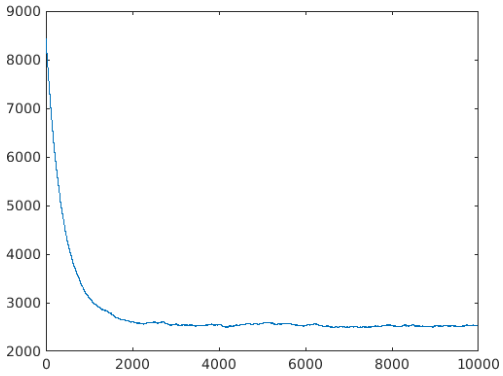
The learning rate of 1e-3 was chosen and $\lambda = 1$. The cost function versus the number of iterations is as follows:



(b) Similar to the batch gradient descent, the stochastic gradient descent update equation is:

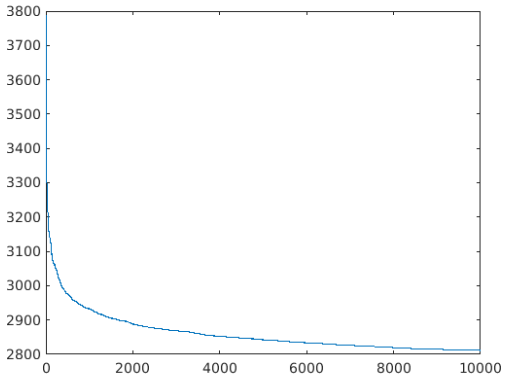
$$w^{t+1} = w^t - \epsilon \nabla_{w,i} J = w^t - \epsilon[-X_i^\top (y_i - s(X_i w)) + 2\lambda w]$$

Where X_i is the ith row of X, y_i is the ith label in y.
The learning rate and λ are the same as in (a). The cost function versus the number of iterations is as follows:



Notice that in Batch GD only 200 iterations are conducted but in Stochastic GD, 10000 iterations are done. Although the number of iterations required by SGD to converge seems to be much higher than that of BGD, SGD requires much less computations because in each iteration, SGD calculates the gradient using only one sample instead of all samples in BGD.

(c) In this case, the learning rate is 0.1 and $\lambda = 1$. The cost is as follows:



This changing learning rate scheme seems to benefit the SGD by making the later iterations more stable.

(d) Because the training set size is only 6000*12, which can be easily handled by batch gradient descent, so BGD is adopted to train the classifier. The learning rate was chosen to be 0.01. To determine the regularization parameter λ, 1/6 of the dataset is used as validation set. And the validation accuracy was calculated with different λ. Then the *lambda* that gave the highest accuracy was adopted, which is $\lambda = 1$.
Kaggle display name: Yicheng Chen
Kaggle score: 0.99597(03/09/2017)

5. First Question

The reason why the linear SVM cannot utilize the new feature well is that the feature is not linearly seperable. Since the time stamp is the number of miliseconds since the previous midnight, the spams messages will spike at around 0 and around MAX, which is not linearly seperable in this case.
To improve the results, he can consider counting the number of miliseconds to the noon of that day. Or he can normalize the original time stamp and use a quadratic kernel in SVM.