STAT 598Z Bonus Project
Instructor: Prof. S V N Vishwanathan **Jiajie Huang**
Due April 23th, 2013 **huang147@purdue.edu**

Problem 1

- Convexity of $J(w)$ in w. Let's first prove it in 2 dimensions. Set $w = (w_1, w_2)$, $x_i = (x_{i1}, x_{i2})$, and we know that y_i is a scalar, then

$$\begin{aligned} J(w) &= J(w_1, w_2) \\ &= \frac{\lambda}{2}\|w\|^2 + \sum_i max(0, 1 - y_i w^T x_i)^2 \\ &= \frac{\lambda}{2}(w_1^2 + w_2^2) + \sum_i max(0, 1 - y_i(w_1 x_{i1} + w_2 x_{i2}))^2 \end{aligned}$$

  Obviously, $f(w_1, w_2) = \frac{\lambda}{2}(w_1^2 + w_2^2)$ is a convex function (quadratic function). Moreover, both $g(w_1, w_2) = 1 - y_i(w_1 x_{i1} + w_2 x_{i2})$(linear function) and $h(w_1, w_2) = 0$ (linear function) are convex functions as well; therefore $max(g(w_1, w_2), h(w_1, w_2)$ is convex; then we know $max(g(w_1, w_2), h(w_1, w_2))^2$ is also convex as quadratic function of a convex function is also convex. Then the linear combination $\sum_i max(g(w_1, w_2), h(w_1, w_2))^2$ is convex, with no doubt. Combining $f(w_1, w_2)$ and $\sum_i max(g(w_1, w_2), h(w_1, w_2))^2$, we get a convex function $J(w_1, w_2)$, or $J(w)$. The same applies to high dimensions of w.

- I used the same gradient descent solver in the main project to find out the optimal w that minimizes $J(w)$, with $\lambda = 0.001$, on the data set $a1a$. The returned w is a multidimensional vector (123×1) that has the same number of variables $(d = 123)$ as in the data set.

- If the test data set $a1at$ is a $n \times d(d = 123)$ matrix X, and the label of $a1at$ is a $n \times 1$ vector Y, then The accuracy is calculated by

$$Accuracy = \frac{n_{sign(w^T X)=Y}}{n}$$

  I use $\epsilon = 0.01$, $\eta = 1e^{-5}$ and vary the value of λ. I also set the maximum number of iterations to be 1000 in my gradient descent solver. The results are summarized as below. From linear algebra we can rewrite the vectorized gradient as

$$\nabla J(w) = \lambda w + \nabla(\sum_i max(0, 1 - y_i w^T x_i)^2)$$

  When λ is small, the first part of $\nabla J(w)$ is dominated by the second part and causes ill conditions. When λ increases, the accuracy increases as well, which means the first term of $\nabla J(w)$ is dominating the gradient descent calculation.

```
Converged in  42 iterations
Lambda: 0.001  Accuracy: 0.687911874919

Converged in  42 iterations
Lambda: 0.0001  Accuracy: 0.687911874919

Converged in  42 iterations
Lambda: 0.0  Accuracy: 0.687911874919

Converged in  42 iterations
Lambda: 0.01  Accuracy: 0.687911874919

Can not converge in 1000 iterations
Lambda: 0.1  Accuracy: 0.748029461171

Can not converge in 1000 iterations
Lambda: 1.0  Accuracy: 0.747997157255

Can not converge in 1000 iterations
Lambda: 100.0  Accuracy: 0.77955808244

Converged in  826 iterations
Lambda: 1000.0  Accuracy: 0.759465047164

Converged in  114 iterations
Lambda: 10000.0  Accuracy: 0.759465047164

Converged in  9 iterations
Lambda: 100000.0  Accuracy: 0.759465047164
```