COMS 4771 Fall 2016 Homework 4

Name - Pulkit Jain UNI - pj2313

Solution 1. (a) The objective function is,

$f(w) = \frac{\lambda}{2}||w||_2^2 + \frac{1}{|S|}\Sigma_{(x,y)\in S}(< w, x > -y)^2$
$f'(w) = \lambda w + \frac{2}{|S|}\Sigma_{(x,y)\in S}x(< w, x > -y)$

$$\frac{d}{dw}w = \begin{bmatrix} \frac{\partial w_1}{\partial w_1} & \frac{\partial w_1}{\partial w_2} & \dots & \frac{\partial w_1}{\partial w_d} \\ \frac{\partial w_2}{\partial w_1} & \frac{\partial w_2}{\partial w_2} & \dots & \frac{\partial w_2}{\partial w_d} \\ \dots & \dots & \dots & \dots \\ \frac{\partial w_d}{\partial w_1} & \frac{\partial w_d}{\partial w_2} & \dots & \frac{\partial w_d}{\partial w_d} \end{bmatrix} = I$$

$$\frac{d}{dw}\Sigma_{(x,y)\in S}x(< w, x > -y) = \begin{bmatrix} \frac{\partial x_1(<w,x>)}{\partial w_1} & \frac{\partial x_1(<w,x>)}{\partial w_2} & \dots & \frac{\partial x_1(<w,x>)}{\partial w_d} \\ \frac{\partial x_2(<w,x>)}{\partial w_1} & \frac{\partial x_2(<w,x>)}{\partial w_2} & \dots & \frac{\partial x_2(<w,x>)}{\partial w_d} \\ \dots & \dots & \dots & \dots \\ \frac{\partial x_d(<w,x>)}{\partial w_1} & \frac{\partial x_d(<w,x>)}{\partial w_2} & \dots & \frac{\partial x_d(<w,x>)}{\partial w_d} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 & x_1x_2 & \dots & x_1x_d \\ x_2x_1 & x_2^2 & \dots & x_2x_d \\ \dots & \dots & \dots & \dots \\ x_dx_1 & x_dx_2 & \dots & x_d^2 \end{bmatrix} = XX^T$$

$=> f"(w) = \lambda I + \frac{2}{|S|}XX^T$
for some non zero vector Z, $I = Z^T I Z > 0$
$=> I = Z^T(I\lambda)Z$
$\mathrm{XX}^T issymmetric, So, \mathrm{XX}^T = Z^T(M)Z = Z^T(XX^T)Z = (X^TZ)^2 >= 0$
$=> M >= 0$
$=> f"(w) = Z^T(\lambda I + M)Z$
Since, $Z^T(\lambda I)Z > 0 and Z^T(M)Z >= 0$
$=> Z^T(\lambda I + M)Z > 0$ So it is positive definite, $f"(w) > 0$ So, the optimization problem is convex

Algorithm 1 Gradient Descent

(b)
eta ← step size *string*
noi ← *no of iterations*
w ← *initial solution*
x ← *input data*
y ← *input labels*
n ← *size of input labels*
gradient ← 0
for $j = 1$ to *noi* **do**
  for $i = 1$ to n **do**
    $gradient \leftarrow gradient + x(i)(dot(w, x(i)) - y(i))$
  $w \leftarrow w - (eta * \lambda)$

Final weight vector for this gradient descent will be w.

(c) yes

(d) no

(e) yes

Solution 2. (a) Gradient Descent Algorithm

Algorithm 2 Gradient Descent

$eta \leftarrow$ step size *string*

$noi \leftarrow no\ of\ iterations$

$\beta_0, \beta \leftarrow initial\ solution$

$x \leftarrow input\ data$

$x \leftarrow [x, 1]$ ($x \in R^d$ to $x \in R^{d+1}$)

$w \leftarrow [\beta, \beta_0]$ ($sizeof\beta_i = d$, $sizeofw_i = d + 1$)

$y \leftarrow input\ labels$

$n \leftarrow size\ of\ input\ labels$

for $j = 1$ to noi **do**

  $\lambda \leftarrow dot((1/(1 + exp(-dot(w, x)))) - y, x)/n$

  $w \leftarrow w - (eta * \lambda)$

Final weight vector for this gradient descent will be w.

(b) The number of iterations needed = 22385

(c) When all three features for all the data given are plotted we see that the second feature is in range [0,1] while the first and third features have range [0,20]. So the first and third features are normalized so that they too are in range [0,1]. This will lead to faster gradient decent as, step size remaining the same, the gradient will be steeper and hence gradient decent will reach more closer to the solution in each step than without normalization.

The linear transformation is A= $\begin{bmatrix} 1/20 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/20 \end{bmatrix}$

377 iterations are required to achieve the objective value.

(d) Before Transformation

the number of iterations executed = 32

the final objective value = 0.6914

the final hold-out error rate = 0.5244

After Transformation

the number of iterations executed = 32

the final objective value = 0.6648

the final hold-out error rate = 0.3793