CS 189 : Homework 3

Problem 1

1. (Code included).

2. My RSS value (computed in code again!) is 8.1761×10^{12}.

3. (In code): Look for figure 1.

4. (In code): Look for figure 2. This histogram represents a gaussean distribution with a mean centered around zero.

Problem 2 :

1. $R[w] = \sum_{i=1}^{n} \log(1 + e^{-z(i)})$

where $z^{(i)} = y^{(i)} f(x^{(i)})$ and $f(x) = w^T x$

Now $f(x) = \sum_i w_i x_i$

$z = yf(x) = \sum_i w_i y x_i$

$\frac{\partial z}{\partial w_i} = y x_i$

$\Rightarrow \frac{\partial z}{\partial w} = Y^T X$ (1×feature) vector.

$\text{Loss}_{\log} = \log(1+e^{-z})$ [Putting risk fn together]

$\frac{\partial \text{Loss}}{\partial z} = \frac{-e^{-z}}{1+e^{-z}}$ (scalar) $= A$

$\Rightarrow \Delta w_i = -\eta \left(\frac{\partial \text{Loss}}{\partial z}\right)\left(\frac{\partial z}{\partial w}\right) = +\eta A Y^T X$

2.

continued to next page....

$$R[\omega] = \sum_{i=1}^{n} \log\left(1+e^{-y^{(i)}(\omega^T x^{(i)})}\right)$$

$$\frac{\partial R[\omega]}{\partial x_i} = \sum_{i=1}^{n} \frac{1 \times \left(e^{-y^{(i)}(\omega^T x^{(i)})}\right)\left(-y^{(i)}\omega^{(i)}\right)}{\left(1+e^{-y^{(i)}(\omega^T x^{(i)})}\right) \times 1}$$

$$\frac{\partial R[\omega]}{\partial x_i \partial y_i} = \frac{\partial}{\partial y^{(i)}}\left[\frac{f\cdot(-y^{(i)}\omega^{(i)})}{1+f}\right] \quad \left[\text{NOTE: } f = e^{-y^{(i)}(\omega^T x^{(i)})}\right]$$

$$= \cancel{-\omega^{(i)}} - \omega(i)\frac{\partial}{\partial y^{(i)}}\left(\frac{f\cdot y^{(i)}}{1+f}\right)$$

$$= -\omega(i)\frac{\partial}{\partial y^{(i)}}\left[y^{(i)}\left(1/f+1\right)^{-1}\right]$$

$$= -\omega(i)\frac{\partial}{\partial y^{(i)}}\left[\left(1/f+1\right)^{-1} + y^{(i)}\frac{\partial}{\partial y^{(i)}}\left(1/f+1\right)^{-1}\right]$$

$$= -\omega(i)\left[\frac{f}{f+1} + y^{(i)}\left(-1\left(1/f+1\right)^{-2}\frac{\partial}{\partial y^{(i)}}\left(1/f\right)\right)\right]$$

$$= -\omega(i)\left(\frac{e^{-y(i)(\omega^T x^{(i)})}}{e^{-y^{(i)}\omega^T x^{(i)}}+1} + y^{(i)}\left(\frac{f}{f+1}\right)^2 \omega^T x^{(i)} e^{y(i)\,\omega^T x^{(i)}}\right)$$

$$= \omega(i)\left(\frac{f}{f+1}\right)\left[y^{(i)}\left(\frac{\cancel{f}}{f+1}\right)\frac{\omega^T x^{(i)}}{\cancel{f}} + 1\right]$$

which is ≥ 0 for all conditions.

$\Rightarrow$ The matrix is positive semidefinitive!

3\. (a) $\mu^{(0)} = \begin{bmatrix} 0.9526 \\ 0.7311 \\ 0.2889 \\ 0.7311 \end{bmatrix}$

(b) $w^{(1)} = \begin{bmatrix} -2 \\ 6.2655 \\ 0 \end{bmatrix}$

(c) $\mu^{(1)} = \begin{bmatrix} 1.0000 \\ 1.0000 \\ 0.0019 \\ 0.0139 \end{bmatrix}$

(d) $w^{(2)} = \begin{bmatrix} -2 \\ 14.2025 \\ 0 \end{bmatrix}$

Problem 3

1. (code attached) : Plots (figures (1), (2) & (3)).

2. (code attached): Plots (figures (4), (5), (6))
   We notice that while I could converge very quickly in batch gradient descent, it took me much longer for stochastic graddient descent.

3. The learning rate $\eta \propto 1/t$ did help the convergence a lot. (Plots included) (figures (7), (8), (9))
   * This is visible more for $\eta \propto 1/10t$ more.

4. a) $f(x) = \sum_{i=1}^{n} \alpha_i K(x^{(i)}, x)$

$$K(x^{(i)}, x) = (x^T x^{(i)} + 1)^2$$

$$f(x) = \sum_i \overset{w_i}{\cancel{\alpha_i}} \Phi_i(x)$$

$$= \sum_n \alpha_n K(x^n, x).$$

similarly proceeding by Question 2,

$$\Rightarrow \Delta w = \eta S(-Z) y^k \phi(x^k)$$

but by kernel trick, $\Delta w = \Delta \alpha_k \phi(x^k)$

$$\Rightarrow \boxed{\Delta \alpha_k = \eta S(-Z) y^k}$$

Scanned by CamScanner

$$\Rightarrow \quad \alpha_i \leftarrow \alpha_i + \eta S(-z^{(i)}) y^{(i)}$$

where $z^{(i)} = y^{(i)} f(x^{(i)})$.

$$\frac{\partial L}{\partial \alpha_i} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha_i} \left(\sum_{i=1}^{n} \log(1+e^{-z^{(i)}}) \right)$$

$$= \sum_{i=1}^{n} - \frac{e^{-z^{(i)}} \cdot \partial/\partial\alpha_i (z^{(i)})}{1+e^{-z^{(i)}}}$$

$$= \sum_{i=1}^{n} - \frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}} y^{(i)} \, \partial/\partial\alpha_i (f(x^{(i)}))$$

$$= - \frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}} y^{(i)} K(x^{(i)}, x)$$

$$= \eta S(-z) y^{(i)} K(x^{(i)}, x).$$

which is different from our answer.

(b) Code is present. Generates 3 graphs.
fig 10 : training_risk
fig 11 : ~~test~~ validation_risk
fig 12 : difference (hopefully close to 0).

(c) The quadratic kernel does not ~~overfits~~ overfit compared to linear kernel. I found that as I change η, so to a certain point, there was bad fit but when I increase a bit, I get a better fit.

Problem 4. :-

I feel that this may have happened because of 2 reasons.

1. Daniel overcomplicated his model thus make it overfit data during A/B testing.
2. Daniel forgot to account for noise.
3. used milliseconds thus overfitting

We ~~us~~ could use a quadratic/linear kernel to count the number of 30 ~~s~~ minute intervals between midnight. Also use a gaussean noise filter to remove the good ~~emails~~ emails from the previous day (in a different timezone etc).

~~I would also check to find the emails~~

~~co~~