Machine Learning Hw3

r03922145 Yi Huang

Question 1

```
sigma <- 0.1
d <- 8
N <- c(10,25,100,500,1000)
Ein <- sigma^2*(1-(d+1)/N)
print(data.table(N,Ein))
```

```
##        N     Ein
## 1:    10 0.00100
## 2:    25 0.00640
## 3:   100 0.00910
## 4:   500 0.00982
## 5:  1000 0.00991
```

Question 2

[**a**] H is positive semi-definite if all eigenvalues of H are non-negative. Suppose λ is the eigenvalue and b is the eigenvector.

$$\begin{aligned} Hb &= \lambda b \\ H^2 b &= \lambda H b \\ &= \lambda(\lambda b) \\ &= \lambda^2 b \end{aligned}$$

Because of "idempotent", $H^2 b = Hb$, we have

$$\lambda = \lambda^2 b$$

The solutions of the equation is either 0 or 1 which means H is positive semi-definite.

[**d**] We know that H and $(I - H)$ are symmetric matrix and $trace(I - H) = N - (d + 1)$. The trace of a symmetric matrix equals to the sum of its diagonal elements, thus we have $trace(H) = d + 1$. Because the trace of matrix is also the sum of its eigenvalues, the sum of eigenvalues of the hat matrix is $d + 1$. As mentioned above, the hat matrix is positive semi-definite, the eigenvalue of hat matrix is eigher 0 or 1, so $d + 1$ eigenvalues of H are 1.

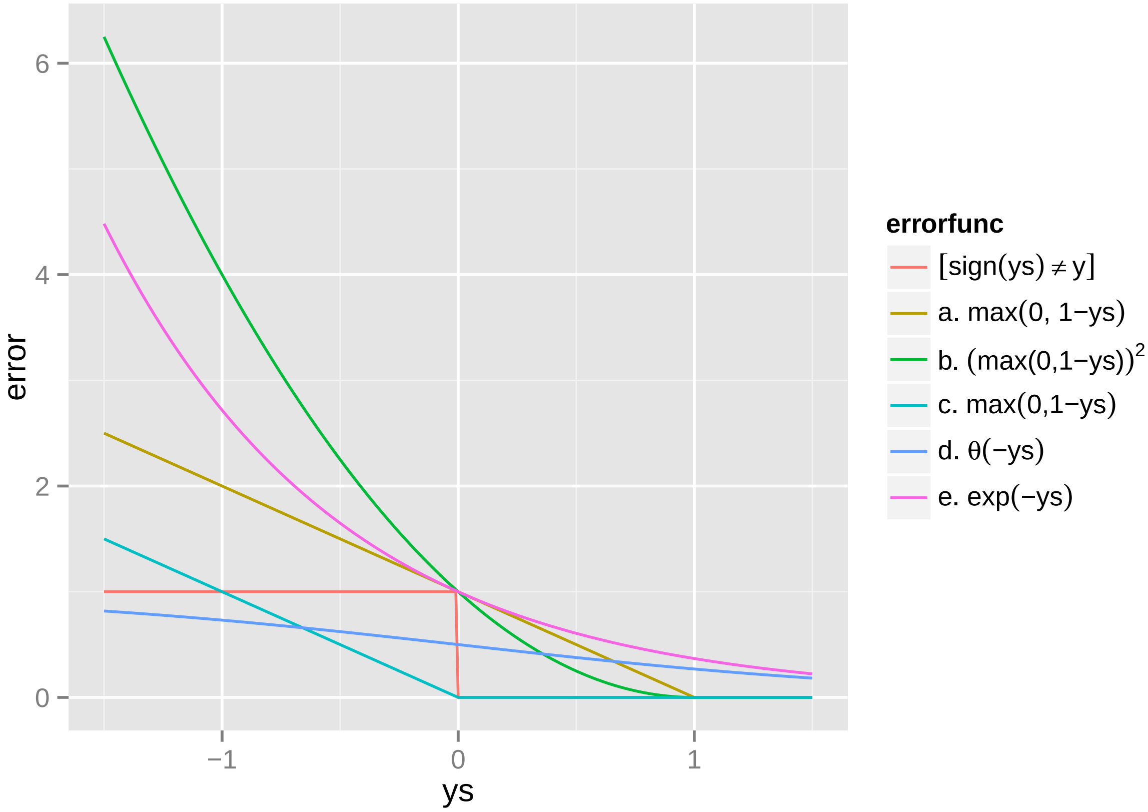
[**e**]

$$\begin{aligned} H^2 &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) \\ &= X(\ (X^TX)^{-1}(X^TX)\ )(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

H is idempotent, therefore $H^{1126} = H$.

Question 3

Let $s = w^T x$, $ys = yw^T x$. Plot 5 error functions:



So the correct answer is [a],[b],[e].

Question 4

Let $s = w^T x$, $ys = yw^T x$.

[**a**] wrong
$err(ys) = max(0, 1 - ys)$ is not differentiable at $ys = 1$

[**b**] correct

$$err^{'}(ys) = \begin{cases} 0 & ys > 1 \\ -2(1-ys) & ys < 1 \end{cases}$$

and when $ys = 1$, $0 = -2(1-1) = 0$

[**c**] wrong
$err(ys) = max(0, -ys)$ is not differentiable at $ys = 0$

[**d**] correct

$$err^{'}(ys) = \frac{e^{ys}}{(1+e^{ys})^2}$$

which is continous of ys everywhere

[**e**] correct

$$err^{'}(ys) = -exp(-ys)$$

which is continous of ys everywhere

Question 5

The prerequisity of the halting of PLA is the $\mathcal{D}$ is linear separable, that means the final error should result in zero. In other words, if $yw^Tx > 0$, $err(w) = 0$.Then only $err(w) = max(0, -yw^Tx)$ can satisfy such property.

Question 6

$$\frac{\partial E(u,v)}{\partial u} = e^u + ve^{uv} + 2u - 2v - 3$$
$$\frac{\partial E(u,v)}{\partial v} = 2e^{2v} + ue^{uv} - 2u + 4v - 2$$
$$\bigtriangledown E(u,v) = (\frac{\partial E(u,v)}{\partial u}, \frac{\partial E(u,v)}{\partial v})$$

The gradient $\bigtriangledown E(u,v)$ around $(0,0)$ is $(-2,0)$.

Question 7

```
## u1 = 0.020 , v1 = 0.000
## u2 = 0.039 , v2 = 0.000
## u3 = 0.058 , v3 = 0.001
## u4 = 0.076 , v4 = 0.001
## u5 = 0.094 , v5 = 0.002
```

After five updates, $E(u_5, v_5) = E$ (0.094, 0.002) = 2.825.

Question 8

$\hat{E}_2(\Delta u, \Delta v)$ is the second-order Taylor's expansion of E around (u,v), then we have:

$$\hat{E}_2(\Delta u, \Delta v) = E(u,v) + \Delta E(u,v) \cdot (\Delta u, \Delta v) + \frac{1}{2}(\Delta u, \Delta v)\Delta^2 E(u,v) \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}$$

where $\Delta^2 E(u,v)$ is the Hessan matrix of E:

$$\Delta^2 E = \begin{pmatrix} \frac{\delta^2 E}{\delta u^2} & \frac{\delta^2 E}{\delta u \delta v} \\ \frac{\delta^2 E}{\delta v \delta u} & \frac{\delta^2 E}{\delta v^2} \end{pmatrix} = \begin{pmatrix} e^u + v^2 e^{uv} + 2 & e^{uv} + uve^{uv} - 2 \\ e^{uv} + uve^{uv} - 2 & 4e^{2v} + u^2 e^{uv} + 4 \end{pmatrix}$$

Then $\hat{E}_2$ around $(0,0)$ is:

$$3 + (-2,0) \cdot (\Delta u, \Delta v) + \frac{1}{2}(\Delta u, \Delta v) \begin{pmatrix} 3 & -1 \\ -1 & 8 \end{pmatrix} \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}$$
$$= 1.5 \cdot (\Delta u)^2 + 4 \cdot (\Delta v)^2 - 1 \cdot (\Delta u)(\Delta v) - 2 \cdot \Delta u + 0 \cdot \Delta v + 3$$

So $(b_{uu}, b_{vv}, b_{uv}, b_u, b_v, b) = (1.5, 4, -1, -2, 0, 3)$

Question 9

$\hat{E}_2$ attains its minimum when its derivative with respect to $(\Delta u, \Delta v)$ is queal to zero:

$$\frac{\hat{E}_2(\Delta u, \Delta v) - E(u,v)}{(\Delta u, \Delta v)} = 0 \Leftrightarrow \Delta E(u,v) + \Delta^2 E(u,v) \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = 0$$

Then the optimal $(\Delta u, \Delta v)$ is:

$$\begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = -(\Delta^2 E(u,v))^{-1} \Delta E(u,v)$$

Question 10

```
## [1] 2.361
```

Question 11

The union set of quadratic, linear, or constant hypotheses in $\mathbb{R}^2$ is just as linear hypotheses in $\mathcal{Z}$ - a space after $\phi_2(x)$ transformation.

$$\begin{aligned} &for\ vector\ x = (x1, x2) \in \mathbb{R}^2 \\ &\phi_2(x) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2) \end{aligned}$$

Then we have

$$\phi_2(X) = \begin{bmatrix} (\phi_2(x_1))^T \\ (\phi_2(x_2))^T \\ (\phi_2(x_3))^T \\ (\phi_2(x_4))^T \\ (\phi_2(x_5))^T \\ (\phi_2(x_6))^T \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The determinant of this matrix $det(\phi_2(X)) = 16 \neq 0$, which means that all six points can be shattered by the union of quadratic, linear of constant hypotheses of x.

Question 12

Because $\mathcal{Z} - space$ can "memorise" all points from $\mathcal{X} - space$ and store it in its $n - th$ dimension. All the points in $\mathcal{Z} - space$ are non-colinear, then can be shattered by linear classifier. So the "maximum"" number of points that can be shattered by the process is ∞.

Question 16

The likelihood that h generate $\mathcal{D}$ is $\prod_{n=1}^{N} P(x_n) h_{y_n}(x_n)$, then we have:

$$
\begin{aligned}
likelihood(h) &= \prod_{n=1}^{N} P(x_n) h_{y_n}(x_n) \\
&\propto \prod_{n=1}^{N} h_{y_n}(x_n) \\
&= \prod_{n=1}^{N} \frac{exp(w_{y_n}^T x_n)}{\sum_{i=1}^{K} exp(w_i^T x_n)} \\
&\propto ln \prod_{n=1}^{N} \frac{exp(w_{y_n}^T x_n)}{\sum_{i=1}^{K} exp(w_i^T x_n)} \\
&\propto \frac{1}{N} \sum_{n=1}^{N} (ln(exp(w_{y_n}^T x_n)) - ln(\sum_{i=1}^{K} exp(w_i^T x_n))) \\
&= \frac{1}{N} \sum_{n=1}^{N} (w_{y_n}^T x_n - ln(\sum_{i=1}^{K} exp(w_i^T x_n)))
\end{aligned}
$$

So a sound $E_{in}(w_1, ..., w_K)$ that minimizes the negative log likelihood is:

$$
\frac{1}{N} \sum_{n=1}^{N} (ln(\sum_{i=1}^{K} exp(w_i^T x_n)) - w_{y_n}^T x_n)
$$

Question 17

For

$$
E_{in} = \frac{1}{N} \sum_{n=1}^{N} (lnA - w_{y_n}^T x_n)
$$

where

$$
A = \sum_{i=1}^{K} exp(w_i^T x_n)
$$

Then

$$
\begin{aligned}
\frac{\partial E_{in}}{\partial w_i} &= \frac{1}{N} \sum_{n=1}^{N} (\frac{\partial (lnA)}{\partial w_i} - \frac{\partial (w_{y_n}^T x_n)}{\partial w_i}) \\
&= \frac{1}{N} \sum_{n=1}^{N} (\frac{1}{A} \cdot \frac{\partial A}{\partial w_i} - [\![y_n = i]\!] x_n) \\
&= \frac{1}{N} \sum_{n=1}^{N} (\frac{x_n exp(w_i^T x_n)}{A} - [\![y_n = i]\!] x_n) \\
&= \frac{1}{N} \sum_{n=1}^{N} ((\frac{exp(w_i^T x_n)}{\sum_{i=1}^{K} exp(w_i^T x_n)} - [\![y_n = i]\!]) x_n) \\
&= \frac{1}{N} \sum_{n=1}^{N} ((h_i(x) - [\![y_n = i]\!]) x_n)
\end{aligned}
$$

Question 21

The least numnber of queries is $N+1$.

$$RMSE(h) = 0 \rightarrow \sum_{n=1}^{N}(y_n - h(x_n))^2 = 0$$

Let $k \in \mathbb{R}^N$ and $query(k) = \sum_{n=1}^{N}(y_n - k_n)^2 = d$.

Start with $k_0 = \{0\}^N$, then $query(k_0) = \sum_{n=1}^{N} y_n^2 = d_0$

Counstruct a query $query(k_i)$ then substract from $query(k_0)$, we get :

$$\begin{aligned} query(k_0) - query(k_i) &= \sum_{n=1}^{N} y_n^2 - \sum_{n=1}^{N}(y_n - k_{in})^2 \\ &= \sum_{n=1}^{N}(2k_{in}y_n - k_{in}^2) \\ &= d_i \end{aligned}$$

which is a linear equation about y. If we want to solve y, we need at least N such linear equations about y.

e.g. Consider a resonable K given below:

$$K = \begin{bmatrix} k_1 \\ \vdots \\ k_i \\ \vdots \\ k_N \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

$$query(K) = \begin{bmatrix} query(k_1) \\ \vdots \\ query(k_i) \\ \vdots \\ query(k_N) \end{bmatrix}$$

$$query(k_0) - query(K) = \begin{bmatrix} query(k_0) - query(k_1) \\ \vdots \\ query(k_0) - query(k_i) \\ \vdots \\ query(k_0) - query(k_N) \end{bmatrix} = \begin{bmatrix} 2y_1 - 1 \\ \vdots \\ 2y_i - 1 \\ \vdots \\ 2y_N - 1 \end{bmatrix} = \begin{bmatrix} d_1 \\ \vdots \\ d_i \\ \vdots \\ d_N \end{bmatrix}$$

Then we can solve it easily. So the total number of queries we need is at least $N+1$.

Question 23

$$\min_{w_1, w_2, \ldots, w_K} RMSE(H) \rightarrow \min_{w_1, w_2, \ldots, w_K} \sum_{n=1}^{N}(y_n - H(x_n))^2$$

Let

$$
\begin{aligned}
f(w) &= \sum_{n=1}^{N}(y_n - H(x_n))^2 \\
&= \sum_{n=1}^{N} y_n^2 - 2\sum_{n=1}^{N} y_n H(x_n) + \sum_{n=1}^{N} H^2(x_n)
\end{aligned}
$$

Where $H(x_n) = \sum_{k=1}^{K} w_k h_k(x_n)$ If f is minimized, the partial derivative of all w_i should be 0.

$$
\begin{aligned}
\frac{\partial f}{\partial w_i} &= -2\sum_{n=1}^{N} y_n h_i(x_n) + 2\sum_{n=1}^{N} H(x_n)h_i(x_n) \\
&= -2h_i^T y + 2\sum_{n=1}^{N}(h_i(x_n)\cdot\sum_{k=1}^{K} w_k h_k(x_n)) \\
&= -2h_i^T y + 2\sum_{k=1}^{K}(\sum_{n=1}^{N} h_i(x_n)h_k(x_n))\cdot w_k \\
&= -2h_i^T y + 2\sum_{k=1}^{K} h_i^T h_k w_k \\
&= 0
\end{aligned}
$$

That means we should solve such linear equations:

$$
\begin{aligned}
h_1^T h_1 w_1 + h_1^T h_2 w_2 + \cdots + h_1^T h_K w_K &= h_1^T y \\
h_2^T h_1 w_1 + h_2^T h_2 w_2 + \cdots + h_2^T h_K w_K &= h_2^T y \\
&\vdots \\
h_K^T h_1 w_1 + h_K^T h_2 w_2 + \cdots + h_K^T h_K w_K &= h_K^T y
\end{aligned}
$$

Followed by question 22, as we have already known $h_i^T h_j$, the least number of queries to get $h_1^T y \sim h_K^T y$ is $K+1$, so the least number of queries would be $K+1$ to solve $\underset{w_1,w_2,\ldots,w_K}{min} RMSE(H)$.