Problem 1.

1.

True.

2.

True.

3.

True.

4.

False.

Problem 2.

(a)

Gradient of the ith parameter θ^i at iteration t:

$g_t^i = \nabla_{\theta^i} \ell(\theta_t)$

The adaptive learning rate used by Adagrad:

$\eta_t^i = \frac{\eta_0}{\sqrt{\sum_{s=0}^{t} (g_s^i)^2}}$

The Adagrad with minibatch:

$\hat{g}_t \leftarrow \nabla_{\hat{\theta}} \frac{1}{m} \sum_{i=1}^{m} L(f(x_i; \tilde{\theta}_t), y_i)$

$r_t \leftarrow r_{t-1} + \hat{g}_t \odot \hat{g}_t$

$\triangle\theta \leftarrow -\frac{\eta_0}{\delta + \sqrt{r_t}} \odot \hat{g}_t$

$\theta_{t+1} \leftarrow \theta_t + \triangle\theta$

Adagrad algorithm updates different learning rate for each parameter. It can be seen from the AdaGrad algorithm that as the algorithm continues to iterate, r_t will become larger and larger, and the overall learning rate will become smaller and smaller. Therefore, in general, the AdaGrad algorithm is incentive convergence at the beginning, and then gradually becomes penalty convergence,

and the speed is getting slower and slower.

In SGD, as the gradient increase,our learning step length should be increased. But in AdaGrad, as the gradient $\hat{g}_t$ increases, our r_t is gradually increasing, and r_t is on the denominator when the gradient is updated, that is, the entire learning rate is reduced. But if the initial $\hat{g}_t$ is large, the training speed would be slow.

(b)

For an L layer neural network:

$f = f_L \circ f_{L-1} ... \circ f_1$

$f(x) = f_L(f_{L-1}(...f(x)...))$

So, the Gradient of f with respect input x is the product:

$f' = f'_L ... f'_1$

when the number of layers increases, if the derivative part is larger than 1, the final gradient update will increase exponentially and gradient explosion occurs; if the derivative part is smaller than 1, the calculated gradient update information will decay exponentially and vanishing gradient occurs.

Problem 3.

(a)

When VC dimension of $\mathcal{H}$ equals to 1, we have 2 choices of y labels for 1 point: 0 and 1. For the chosen point $i \in R$ and the threshold point a, we can always set $i_1 = a + j$, where $j \in R$ and $i_2 = a - j$, where $j \in R$. Then we could always obtain $y_{i_1} = 1$ for i_1 and $y_{i_2} = 0$ for i_2 using function class $\mathcal{H}$. Thus, $VC(\mathcal{H}) \geq 1$.

When VC dimension of $\mathcal{H}$ equals to 2, we have four different combinations: 00,11,01, and 10. The 10 combination could not be implemented by the function class $\mathcal{H}$. First, we set $i_1 \leq i_2$ and assume there exist an a = j($j \in R$) let $y_{i_1} = 1$ and $y_{i_2} = 0$.

When $y_{i_1} = 1$, $i_1 > a$. When $y_{i_2} = 0$, $i_2 \leq a$. Thus $i_2 < i_1$, which contradicts our assumption. So there is no a satisfies the 10 combination using function class $\mathcal{H}$ and the VC dimension of $\mathcal{H}$ equals to 1.

(b)

$\forall$ lines $\in R^2$, the lines could intersect with a circle at most 2 points.

$\forall$ triangles $\in R^2$, the triangles could intersect with a circle at most 6 points and partition the circle into 6 different parts.

For $VC(\mathcal{H}) \leq 6$, the triangle could shatter $(x_1, x_2, x_3, x_4, x_5, x_6)$. Because the most complicate situation would be: $(x_1(+), x_2(-), x_3(+), x_4(-), x_5(+), x_6(-))$ or $(x_1(-), x_2(+), x_3(-), x_4(+), x_5(-), x_6(+))$laying on the circle and our 6 different parts on the circle partitioned by the triangle could correspond each point to a part, which perfectly shatters 6 points.

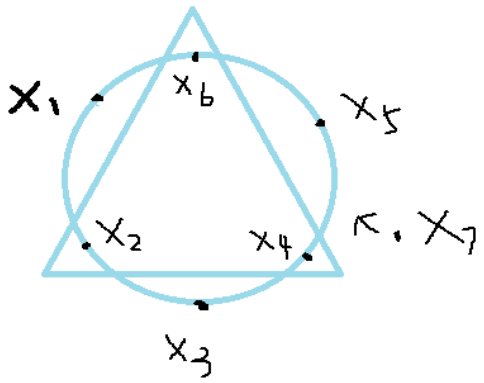

Figure 1: Adding the seventh point

For $VC(\mathcal{H}) = 7$, a new point, either positive or negative will be added on the circle. Then the problem converts to whether we could use 2 edges to partition 3 combinations$(+--+, ++-+, +--++)$ or $(-++-, --+-, -+--)$ while the third side of the triangle stays the same. The answer is yes. So, the VC$\mathcal{H} \geq 7$.

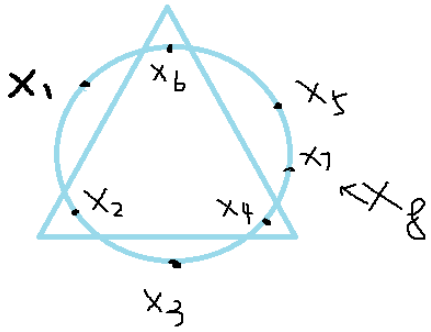

Figure 2: Adding the eighth point

For $VC(\mathcal{H}) = 8$, one more new point, either positive or negative will be added on the circle. If the x_8 locates in the circle, we cannot separate $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ from x_8. If the x_8 locates on the circle or outside the circle, we cannot separate (x_1, x_3, x_5, x_7) from (x_2, x_4, x_6, x_8). In Figure 2, x_3,x_5 and x_7 would be the same sign and x_4,x_8 would be the same sign, which means we need to use 2 edges to partition 5 points with different sign $(+-+-+)$ or $-+-+-$ .This partition is impossible. So the VC($\mathcal{H}$) < 8, which means it equals to 7.

Problem 4.

(a)

$E_{x,y,D}[(f_D(x) - \bar{f}(x))(\bar{f}(x) - y)]$

$= E_{x,y}[E_D(f_D(x) - \bar{f}(x))(\bar{f}(x) - y)]$ (D is independent of (x,y))

$= E_{x,y}[(E_D(f_D(x)) - \bar{f}(x))(\bar{f}(x) - y)]$

$= E_{x,y}[(\bar{f}(x) - \bar{f}(x))(\bar{f}(x) - y)]$

$= 0$

(b)

$E_{x,y}[(\bar{f}(x) - \bar{y}(x))(\bar{y}(x) - y)]$

$= E_x[E_{y|x}[(\bar{f}(x) - \bar{y}(x))(\bar{y}(x) - y)]]$

$= E_x[(\bar{f}(x) - \bar{y}(x))(\bar{y}(x) - E_{y|x}(y))]$

$= E_x[(\bar{f}(x) - \bar{y}(x))(\bar{y}(x) - \bar{y}(x)]$

$= 0$

Problem 5.

(a)

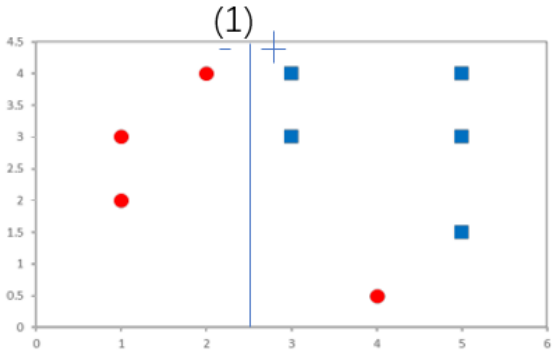

Figure 3: Adaboost (a)

$\theta \in (2, 3]$, where θ is the boundary. Right side indicates positive labels and left side indicates negative labels.

(b)

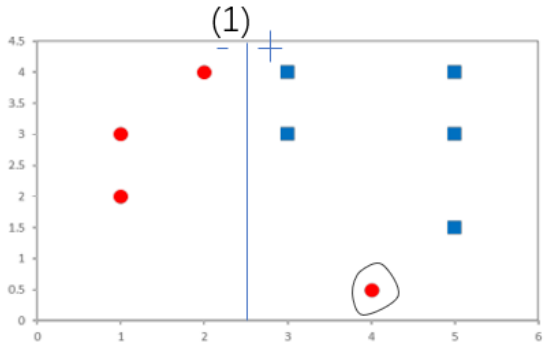

Figure 4: Adaboost (b)

(c)

$\epsilon_t = \sum_i w_t(i) 1[G_t(x_i) \neq y_i]$

$= \frac{1}{9}$

$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$

$= \frac{1}{2} ln(8)$

$w_2(i) = \frac{w_1(i) exp(-\alpha_t y_i G_t(x_i))}{Z_t}$

w_2 (wrong classified point) $= \frac{\frac{1}{9} exp(\frac{1}{2} ln(8))}{Z_t}$

$= \frac{8^{0.5}}{9Z_t}$

w_2 (correct classified point) $= \frac{\frac{1}{9} exp(-\frac{1}{2} ln(8))}{Z_t}$

$= \frac{8^{-0.5}}{9Z_t}$

$\because Z_t$ is normalization factor

$\therefore Z_t = \frac{8^{0.5}}{9} + 8 * \frac{8^{-0.5}}{9}$

w_2 (wrong classified point) $= \frac{8^{0.5}}{9Z_t}$

$= \frac{8^{0.5}}{9} * \frac{9}{8*8^{-0.5}+8^{0.5}}$

$= \frac{1}{2}$

w_2 (correct classified point) $= \frac{8^{-0.5}}{9Z_t}$

$= \frac{8^{-0.5}}{9} * \frac{9}{8*8^{-0.5}+8^{0.5}}$

$= \frac{1}{16}$

The weighted error of the first decision stump after the first boosting iteration would be: $\frac{1}{2}$

(d)

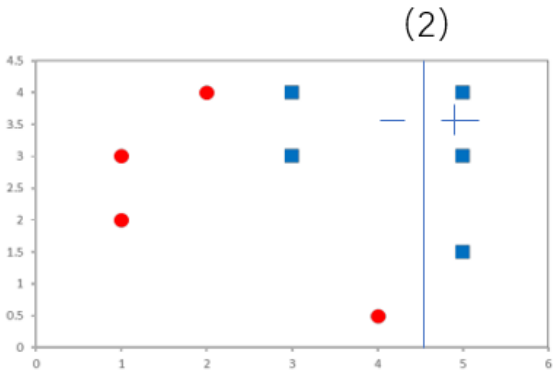

Figure 5: Adaboost (c)

$\theta \in (4, 5]$, where θ is the boundary. Right side indicates positive labels and left side indicates negative labels.

Extra Credit

a

First, set $\frac{t^2/2}{\sum_{i=1}^{n} E[X_i^2]+Mt/3} = i$

Because the mean of random variable is 0, $E[X^2] = D[X] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2]$

Then, $t^2 = 2nE[X^2]i + 2Mti/3$

$t^2 - 2nE[X^2]i - 2Mti/3 = 0$

$t = \frac{2Mi/3 \pm \sqrt{\frac{4Mti}{9}+8nE[X^2]i}}{2} = \frac{Mi}{3} \pm \sqrt{\frac{M^2i^2}{9} + 2nE[X^2]i}$

Equation 2:

$=P(\sum_{i=1}^{n} X_i >= \frac{Mi}{3} + \sqrt{\frac{M^2i^2}{9} + 2nE[X^2]i}) <= 1 - e^{-i}$ (t needs to be greater than 0)

$=P(n\bar{X} >= \frac{Mi}{3} + \sqrt{\frac{M^2i^2}{9} + 2nE[X^2]i}) <= 1 - e^{-i}$

$=P(n\bar{X} <= \frac{Mi}{3} + \sqrt{\frac{M^2i^2}{9} + 2nE[X^2]i}) >= 1 - e^{-i}$

$=P(n\bar{X} <= \frac{Mi}{3} + \sqrt{\frac{M^2i^2}{9}} + \sqrt{2nE[X^2]i}) >= 1 - e^{-i}$(Cauchy Inequality)

$=P(n\bar{X} <= \frac{2Mi}{3} + \sqrt{2nE[X^2]i}) >= 1 - e^{-i}$

$=P(\bar{X} <= \frac{2Mi}{3n} + \sqrt{2E[X^2]i/n}) >= 1 - e^{-i}$

Set $i = log(2/\delta), e^{-(log(2/\delta))} = e^{(log(2/\delta)^{-1})} = \delta/2$

$= P(\bar{X} <= \frac{2Mlog(2/\delta)}{3n} + \sqrt{2E[X^2]log(2/\delta)/n}) >= 1 - \delta/2$

$= P(|\bar{X}| <= \frac{2Mlog(2/\delta)}{3n} + \sqrt{2E[X^2]log(2/\delta)/n}) >= 1 - \delta$ ($X_1, ...X_n$ are i.i.d. real-valued random variables with mean zero)

b

Dont know.