Homework 2

1.2

By definition $C_{class} L_{0-1}(c(h), y) \leq L_{class}(h, y) \quad \forall y \in Y_{\pm} \ \forall h \in \mathbb{R}$

Summing over all data points then we get:

$$\sum_{i=1}^{M} C_{class} L_{0-1}(c(h), y) \leq \sum_{i=1}^{M} L_{class}(h, y) \quad \forall h \in \mathbb{R}$$

$$C_{class} \frac{1}{M} \sum_{i=1}^{M} L_{0-1}(c(h), y) \leq \frac{1}{M} \sum_{i=1}^{M} L_{class}(h, y) \quad \forall h \in \mathbb{R}$$

$$C_{class} \hat{L}_{0-1}(c(h)) \leq \hat{L}_{class}(h) \quad \forall h \in \mathbb{R}$$

$$\hat{L}_{0-1}(c(h)) \leq \frac{1}{C_{class}} \hat{L}_{class}(h) \quad \forall h \in \mathbb{R}$$

$\square$

1.3

(i) Yes: - If $sgn(h) = y$ then $L_{class}(h, y) = (h-y)^2 \geq 0 = L_{0-1}(c(h), y)$

- If $sgn(h) \neq y$ then $|h-y| \geq 1$ and thus:

$$L_{class}(h, y) = (h-y)^2 \geq 1 = L_{0-1}(c(h), y)$$

$\square$

The best constant is $C_{class} = 1$ since for any other constant $C' = 1+\varepsilon$ for $0 < \varepsilon < 1$ we can have $h = \varepsilon/4$, $y = -1$ and then

$$L_{class}(h, y) = (1+\varepsilon/4)^2 = 1 + \varepsilon/2 + \varepsilon^2/16 < 1+\varepsilon = (1+\varepsilon) L_{0-1}(c(h), y)$$

☆ if $C' = 1+\varepsilon$ for $\varepsilon \geq 1$ then pick $h = 1/10$, $y = -1$ to get

$$L_{class}(h, y) = (1.1)^2 < 2 \leq C' = C' L_{0-1}(c(h), y)$$

(ii) For any constant $C_{class} > 0$ pick $h = 1 + \frac{C_{class}}{2}$, $y = -1$

Then $L_{class}(h, y) = \left|1 + \frac{C_{class}}{2} - 1\right| = \frac{C_{class}}{2} < C_{class} = C_{class} L_{0-1}(c(h), y)$

so $|h+y|$ is not an upper bound for the zero-one loss $\square$

(iii) let $h^* < 0$ and $L_{class}(h^*, 1) = 0$

Then $L_{class}(h^*, 1) = 0 < C_{class} = C_{class} L_{0-1}(h^*, 1)$

for any $C_{class} > 0$. So this cannot be an upper bound for the zero-one loss

(iv) Let $L_{class}(h, y) = ||h| + y|$ then taking $h = 1 + \frac{C_{class}}{2}$, $y = -1$

we get that for every $C_{class} > 0$:

$$L_{class}(h, y) = \frac{C_{class}}{2} < C_{class} = C_{class} L_{0-1}(c(h), y)$$

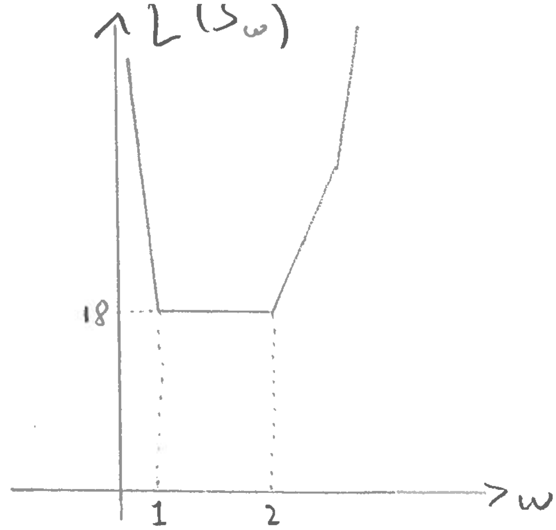
So $L_{class} = ||h| + y|$ is not an upper bound for the zero-one loss but there exists no $h < 0$ with $L_{class}(h, 1) = 0$. (since $L_{class}(h, 1) > 1 \ \forall h$), so the converse is not true.

4.1

We Minimize the total loss $\hat{L}(S_\omega)$: (we assume green lines represent $y = +1$ while red crosses represent $y = -1$ since the score should increase with the chance of $y = +1$)

$$\hat{L}(S_\omega) = 12\max(1-\omega, 0) + 8\max(\omega-1, 0) + 2\max(2-\omega, 0) + 4\max(\omega-2, 0)$$
$$+ 2\max(3-\omega, 0) + 4\max(\omega-3, 0) + 4\max(4-\omega, 0) + 12\max(\omega-4, 0).$$

The following graph illustrates $\hat{L}(S_\omega)$:



So any $w \in (1,2)$ is a minimizer.

In this case this results in the same as the majority classifier, which predicts $y=-1$ for bin 1 and $y=+1$ for bins 2, 3, and 4.

If the bins are relabelled to any non-decreasing values then the classifier will remain unchanged since we have a distance score $x-w$ for which the condition of a minimizer is $FP=FN$ (false positives = false negatives). And this will always lie between bins 1 and 2.

4.2

- If incorrect then $y=1$ and $s<0$ in which case $L_{margin}(s,y)=\max(0,1-s) = 1-s \in [1,\infty)$
or $y=-1$ and $s\geq 0$ in which case $L_{margin}(s,y)=\max(0,1+s)=1+s \in [1,\infty)$

If marginal then $y=1$ and $0\leq s\leq 1$ so $L_{margin}(s,y)=\max(0,1-s)=1-s \in [0,1]$
or $y=-1$ and $-1\leq s\leq 0$ so $L_{margin}(s,y)=\max(0,1+s)=1+s \in [0,1]$

- If confident then $y=1$ and $s\geq 1$ so $L_{margin}(s,y)=\max(0,1-s)=0$
or $y=-1$ and $s\leq -1$ so $L_{margin}(s,y)=\max(0,1+s)=0$

So $L_{margin} \in \begin{cases} [1,\infty), & \text{incorrect} \\ [0,1], & \text{marginal} \\ =0, & \text{confident} \end{cases}$ as expected.

Note that this still holds for the t-margin loss, given the generalizations in excercise 4.4

4.3

Note that if $L_{margin,t}$ is marginal or confident then $c(s)=\gamma$, so $L_{0-1}(c(s),\gamma)=0$ and $L_{margin,t} \geq L_{0-1}$ (since $L_{margin-t} \geq 0$)

And if $L_{margin,t}$ is incorrect then $c(s) \neq \gamma$, so $L_{0-1}(c(s),\gamma)=1$ but $L_{margin,t} \in [1,\infty)$ from the previous excercise.

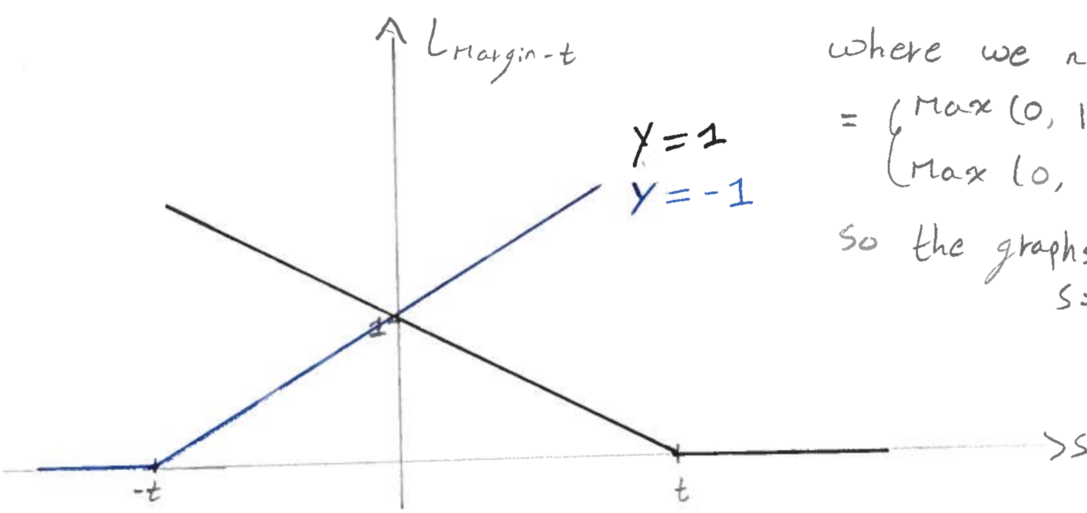
So $L_{margin,t}(s,\gamma) \geq L_{0-1}(c(s),\gamma)$ $\forall s, \forall \gamma$ and thus $[L_{margin,t}, C=sgn, C_{class}=1]$ is an upper bound for the classification error. Summing over all $\gamma \in Y_t$ we get (4).

4.4

(i) Letting $t=1$ we get $L_{margin,1} = \begin{cases} \max(0,1-s), & \gamma=1 \\ \max(0,1+s), & \gamma=-1 \end{cases} = L_{margin}$

(ii)
- incorrect: $c(s) \neq \gamma$
- false positive: $\gamma=-1$, $s>0$
- marginal positive: $\gamma=1$, $0 \leq s \leq t$
- false negative: $\gamma=1$, $s<0$
- marginal negative: $\gamma=-1$, $-t \leq s \leq 0$
- marginal: $\gamma=c(s)$, $|s| \leq t$
- confident: $c(s)=\gamma$ and $|s| \geq t$

4.5



where we note that $L_{margin-t}(-s,-\gamma) = \begin{cases} \max(0, 1+s/t), & \gamma=-1 \\ \max(0, 1-s/t), & \gamma=1 \end{cases} = L_{margin-t}(s,\gamma)$

So the graphs are symmetric about $s=0$.

4.6

$$0 = \hat{L}'_{margin\text{-}t}(S_\omega) = \frac{1}{m}\sum_{i=1}^{m} l'_{margin\text{-}t}(x-\omega, y)$$

$$\Rightarrow 0 = \sum_{i=1}^{m} l'_{margin\text{-}t}(x-\omega, y) = \sum_{i\in J^+} l'_{margin\text{-}t}(x-\omega, y) + \sum_{i\in J^-} l'_{margin}(x-\omega, y)$$

where J^+ are those data points with $y=1$ and J^- those with $y=-1$.

$$\Rightarrow 0 = \sum_{i\in\{y=1,\text{ confident}\}} l'_{margin\text{-}t}(x-\omega, y) + \sum_{i\in\{MP, FN\}} l'_{margin\text{-}t}(x-\omega, y) + \sum_{i\in\{y=-1,\text{ confident}\}} l'_{margin\text{-}t}(x-\omega, y) + \sum_{i\in\{MN, FP\}} l'_{margin\text{-}t}(x-\omega,$$

From excercise 4.5 (the graph of $l_{margin\text{-}t}$) we have that

$$l'_{margin\text{-}t} = \begin{cases} 0, & \text{for confident} \\ 1/t, & \text{for } \{MN, FP\} \\ -1/t, & \text{for } \{MP, FN\} \end{cases}$$

MN = Marginal negative
MP = Marginal positive
FN = false negative.
FP = false positive

So we get $0 = \sum_{i\in\{MP, FN\}} -1/t + \sum_{i\in\{MN, FP\}} 1/t \Rightarrow \#\{MP, FN\} = \#\{MN, FP\}$

$\square$

5.1 (i)

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x(1+e^{-x})} = \frac{e^x}{1+e^x} = p(e^x)$$

$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(r(p))$ by definition

(ii)

$$\text{logit}(\sigma(x)) \underbrace{=}_{\text{from (i)}} \text{logit}(p(e^x)) \underbrace{=}_{\text{from (i)}} \log(r(p(e^x))) \underbrace{=}_{\text{since } r=p^{-1}} \log(e^x) = x$$

$$\sigma(\text{logit}(p)) \underbrace{=}_{\text{from (i)}} p(e^{\text{logit}(p)}) \underbrace{=}_{\text{from (i)}} p(e^{\log(r(p))}) = p(r(p)) \underbrace{=}_{\text{since } p=r^{-1}} p$$

So σ and logit are inverses.

5.2

$\rightarrow 2\sigma(x) = \frac{2}{1+e^{-x}} = \frac{2e^x}{e^x+1} = 1 + \frac{e^x-1}{e^x+1} = 1+\tanh(x/2)$

$\rightarrow 1-\sigma(x) = 1 - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{x}} = \sigma(-x)$

$\rightarrow \sigma'(x) = \left(\frac{1}{1+e^{-x}}\right)' = -\frac{1}{(1+e^{-x})^2}\cdot(-e^{-x}) = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\cdot(1-\sigma(x))$

5.3

equation (8) states that the probability classifier is the same as the score-based classifier for the score that gives the desired probability.

If $c(h)=y$ then $L_{0-1}(c(h),y)=0$ but $l_{score,log}(h,y)=l_{log}(\sigma(h),y)\geq 0$

so $l_{score,log}(h,y)\geq L_{0-1}(c(h),y)$ (since l_{log} is non-negative).

If $c(h)\neq y$ then $L_{0-1}(c(h),y)=1$ and $\begin{cases}\sigma(h)<0.5, & \text{for } y=1\\ \sigma(h)\geq 0.5, & \text{for } y=-1\end{cases}$

in which case $l_{score,log}(h,y) = l_{log}(\sigma(h),y) = -\log(p')$ for $p'\in[0,0.5)$

thus $l_{score-log}(h,y) = -\log(p') \geq -\log(1/2) = -\log(1/2)\, L_{0-1}(c(h),y)$

so $l_{score,log}(h,y)\geq -\log(1/2)\, L_{0-1}(c(h),y)$ $\forall h, \forall y$ so $l_{score-log}$ is an upper bound for the error (with $c(h)=\text{round}(\sigma(h))$ and any $C_{class}\in(0,-\log(1/2)$

The best constant is $C_{class} = -\log(1/2) = \log(2)\approx 0.69$

For this special case, theorem 1.5 states: $\hat{L}_{0-1}(c(h)) \leq \frac{1}{\log(2)}\hat{L}_{score,log}(h)$

5.4

$\hat{L}_{log}(P_w) = \frac{1}{M}\sum_{j\in J^+} -\log(\sigma(x_j-w)) + \frac{1}{M}\sum_{j\in J^-} -\log(1-\sigma(x_j-w))$

$\hat{L}'_{log}(P_w) = \frac{1}{M}\sum_{j\in J^+}\frac{1}{\sigma(x_j-w)}\cdot\sigma'(x_j-w) + 1/M\sum_{j\in J^-}\frac{-1}{1-\sigma(x_j-w)}\cdot\sigma'(x_j-w)$

$$\hat{L}'_{log}(P_\omega) = \frac{1}{M}\sum_{j\in J^+} \frac{\cancel{\sigma(x_j-\omega)}\cdot(1-\sigma(x_j-\omega))}{\cancel{\sigma(x_j-\omega)}} - \frac{1}{M}\sum_{j\in J^-} \frac{\sigma(x_j-\omega)\cdot\cancel{(1-\sigma(x_j-\omega))}}{\cancel{1-\sigma(x_j-\omega)}}$$

$$= \frac{1}{M}\sum_{j\in J^+}(1-\sigma(x_j-\omega)) - \frac{1}{M}\sum_{j\in J^-}\sigma(x_j-\omega)$$

Setting $\hat{L}'_{log}(P_\omega)=0$ we get $\sum_{j\in J^+}(1-\sigma(x_j-\omega)) = \sum_{j\in J^-}\sigma(x_j-\omega)$ as expected. $\square$

6.1

(i) $\frac{\partial l_2(p,y)}{\partial p} = p - y^+ \implies \frac{\partial^2 l_2(p,y)}{\partial p^2} = 1 \quad \forall p$ so l_2 is convex

$$\frac{\partial l_{log}(p,y)}{\partial p} = \begin{cases} \frac{-1}{p}, & y=1 \\ \frac{1}{1-p}, & y=-1 \end{cases} \implies \frac{\partial^2 l_{log}(p,y)}{\partial p^2} = \begin{cases} \frac{1}{p^2}, & y=1 \\ \frac{1}{(1-p)^2}, & y=-1 \end{cases}$$

which is positive $\forall p$ so l_{log} is convex.

(ii) $\hat{L}_2(p) = \frac{1}{M}\sum_{i=1}^{M} l_2(p, y_i) = \frac{1}{M}\sum_{i=1}^{M}(p-y^+)^2/2$

$$\hat{L}'_2(p) = \frac{1}{M}\sum_{y_i=1}(p-1) + \frac{1}{M}\sum_{y_i=-1} p = \frac{q}{M}(p-1) + \frac{(1-q)}{M}p$$

Setting to zero:

$$0 = q(p-1) + (1-q)p$$
$$0 = \cancel{qp} - q + p - \cancel{pq}$$
$$p = q$$

so l_2 is proper

$$\hat{L}_{log}(p) = \frac{1}{M}\sum_{i=1}^{M} l_{log}(p, y_i) = \frac{1}{M}\sum_{y_i=1} -\log(p) + \frac{1}{M}\sum_{y_i=-1} -\log(1-p)$$

$$\hat{L}'_{log}(p) = \frac{1}{M}\sum_{y_i=1}\frac{-1}{p} + \frac{1}{M}\sum_{y_i=-1}\frac{1}{1-p} = \frac{-q}{M}\cdot\frac{1}{p} + \frac{(1-q)}{(1-p)}\cdot\frac{1}{M}$$

setting to zero: $0 = \frac{-q}{p} + \frac{1-q}{1-p} \implies \frac{q}{p} = \frac{1-q}{1-p} \implies q - \cancel{qp} = p - \cancel{qp}$

$$q = p$$

so l_{log} is proper $\square$

6.2

(i) $$\frac{\partial l_s(p,1)}{\partial p} = \frac{-(p^2+(1-p)^2)^{1/2} + \frac{p}{2}(p^2+(1-p)^2)^{-1/2}\cdot(2p-2(1-p))}{p^2+(1-p)^2}$$

$$= \frac{-(p^2+(1-p)^2) + 2p^2 - p}{(p^2+(1-p)^2)^{3/2}} = \frac{p-1}{(p^2+(1-p)^2)^{3/2}}$$

$$\frac{\partial^2 l_s(p,1)}{\partial p^2} = \frac{(p^2+(1-p)^2)^{3/2} - \frac{3}{2}(p-1)\cdot(p^2+(1-p)^2)^{1/2}\cdot(4p-2)}{(p^2+(1-p)^2)^3}$$

$$= \frac{p^2+(1-p)^2 - 6p^2+9p-3}{(p^2+(1-p)^2)^{5/2}} = \frac{-4p^2+7p-2}{(p^2+(1-p)^2)^{5/2}}$$

But plugging in $p=0$ we get $\frac{\partial^2 L_s(0,1)}{\partial p^2} = -2 < 0$

so $L_{spherical}$ is not convex in $[0,1]$

(ii) $$\frac{\partial L_s(p,-1)}{\partial p} = \frac{(p^2+(1-p)^2)^{1/2} + (1-p)/2\cdot(p^2+(1-p)^2)^{-1/2}\cdot(2p-2(1-p))}{p^2+(1-p)^2}$$

$$= \frac{p^2+(1-p)^2+\left(\frac{1-p}{2}\right)\cdot(4p-2)}{(p^2+(1-p)^2)^{3/2}} = \frac{p}{(p^2+(1-p)^2)^{3/2}}$$

Now $\hat{L}_s(p) = \frac{1}{M}\sum_{i=1}^{M} L_s(p) = \frac{q}{M} L_s(p,1) + \frac{(1-q)}{M} l_s(p,-1)$

so $\hat{L}'_s(p) = \frac{q}{M}\left(\frac{p-1}{(p^2+(1-p)^2)^{3/2}}\right) + \frac{(1-q)}{M}\left(\frac{p}{(p^2+(1-p)^2)^{3/2}}\right)$

setting to zero: $0 = q(p-1) + (1-q)p$ (cancelling the equal denominators)

$0 = \cancel{qp} - q + p - \cancel{qp}$

$p = q$ so L_s is proper

$\square$

6.3

(i) $$\frac{\partial l_1(p,y)}{\partial p} = \begin{cases} 1, & y=-1 \\ -1, & y=1 \end{cases} \implies \frac{\partial^2 l_1(p,y)}{\partial p^2} = 0 \quad \forall p, \forall y$$

So l_1 is convex (it is linear so it is convex).

(ii) $$\hat{L}_1(p,y) = \frac{1}{M}\sum_{i=1}^{M} l_1(p,y) = \frac{q}{M}\sum_{y_i=1} l_1(p,1) + \frac{(1-q)}{M}\sum_{y_i=-1} l_1(p,-1)$$

$$= \frac{q(1-p)}{M} + \frac{(1-q)p}{M}$$

$$= \frac{q+p-2qp}{M}$$

$$= \left(\frac{1-2q}{M}\right)p + \frac{q}{M}$$

is linear in p, which means that the minimizer will occur on the boundary. So the minimizer will be $p=0$ or $p=1$ regardless of the value of q (except when $q=1/2$, when any $p \in [0,1]$ is a minimizer). So l_1 is not proper

$\square$