Name: Yuhan Shao
Student Number: 1002281327
UTORid: shaoyuha

CSC411 Assignment 2

1.(a) WTP: $H(X) = \sum_x p(x) \log_2(\frac{1}{p(x)}) \geq 0$

- Since X is a discrete random variable, probability mass function $p(x) \in [0,1]$
- $\boxed{\text{Case 1}}$: when $p(x)=0$, by the condition in the question:
  $p(x)\log_2(\frac{1}{p(x)}) = 0$
- $\boxed{\text{Case 2}}$: when $p(x) \in (0,1] \Leftrightarrow \frac{1}{p(x)} \in [1,+\infty) \Leftrightarrow \log_2(\frac{1}{p(x)}) > 0$
  $\Leftrightarrow p(x)\log_2(\frac{1}{p(x)}) > 0$
- Therefore, for each possible value of $x \in X$, we have:
  $p(x)\log_2(\frac{1}{p(x)}) \geq 0$
  $\Leftrightarrow \sum_x p(x)\log_2(\frac{1}{p(x)}) \geq 0$
  $\Leftrightarrow H(x) \geq 0$, i.e.: entropy $H(X)$ is non-negative.

(b) WTP: $KL(p\|q) = \sum_x p(x)\log_2\frac{p(x)}{q(x)} \geq 0$.

- By appendix: $\log(x)$ is concave on the set of positive real number
  → In order to use Jensen's inequality, which states:
  if $\phi(t)$ is a convex function of t, then:
  $\phi(E[t]) \leq E[\phi(t)]$
  → We need to transfer concave function $\log(x)$ to convex one.
  By the def$^{\underline{n}}$ of concave function, we have:
  f is concave if $-f$ is convex
  → Therefore $-\log(x)$ is convex function on the set of positive real number.
- $KL(p\|q) = \sum_x p(x)\log_2\frac{p(x)}{q(x)}$
  $= \sum_x p(x)(-\log_2\frac{q(x)}{p(x)})$ # transfer to convex function.
  $= E[-\log_2\frac{q(x)}{p(x)}]$ # $\phi(t) = -\log_2 t$; $t = \frac{q(x)}{p(x)}$
  $\geq -\log_2(\sum_x p(x)\frac{q(x)}{p(x)})$ # By Jensen's inequality.
  $= -\log_2(\sum_x q(x)) = -\log_2 1 = 0$
- Therefore, $KL(p\|q) \geq 0$, i.e: $KL(p\|q)$ is non-negative.

(C) WTS: $I(Y;X) = KL(p(x,y)\|p(x)p(y))$
where $p(x) = \sum_y p(x,y)$ is the marginal distribution of X.

- $I(Y;X) = H(Y) - H(Y|X)$ # By def$^{\underline{n}}$ of $I(Y;X)$.
  $= (-\sum_y p(y)\log p(y)) - (-\sum_x\sum_y p(x,y)\log p(y|x))$ # By def$^{\underline{n}}$ of $H(X)$
  $= (-\sum_y(\sum_x p(x,y))\log p(y)) + \sum_x\sum_y p(x,y)\log p(y|x)$ # By marginal distribution of Y
  $= \sum_x\sum_y p(x,y)\log\frac{1}{p(y)} + \sum_x\sum_y p(x,y)\log p(y|x)$
  $= \sum_x\sum_y p(x,y)(\log\frac{1}{p(y)} + \log p(y|x))$
  $= \sum_x\sum_y p(x,y)\cdot\log\frac{p(x,y)}{p(x)p(y)}$ (*)
- $KL(\underbrace{p(x,y)}_{p}\|\underbrace{p(x)p(y)}_{q}) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$ # By def$^{\underline{n}}$ of KL
  $= \sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$ (**)
- By (*) and (**), we conclude that:
  $I(Y;X) = KL(p(x,y)\|p(x)p(y))$

2. WTP: $\forall x, t \quad L(\bar{h}(x), t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t)$

where $\bar{h}(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(x)$

- $\bar{h}(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(x) = E[h_i(x)]$ # expectation of $h_i(x)$ $(*)$
- $L(\bar{h}(x), t) = \frac{1}{2}(\bar{h}(x) - t)^2$ # By def^{n} of squared error loss function

$= \frac{1}{2}(E[h_i(x)] - t)^2$ # By $(*)$, Let $\phi(s) = \frac{1}{2}(s-t)^2$, $s = h_i(x)$

$\leq E[\frac{1}{2}(h_i(x) - t)^2]$ # By Jensen's inequality.

$= \frac{1}{2}(E[(h_i(x) - t)^2])$ # By the property of expected value.

$= \frac{1}{2} \cdot \frac{1}{m}\sum_{i=1}^{m}(h_i(x) - t)^2$

$= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(h_i(x) - t)^2$

$= \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t)$ # By def^{n} of squared error Loss function

3. WTS: $err_t' = \frac{\sum_{i=1}^{N} w_i' I\{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'} = \frac{1}{2}$

- By the condition in the question we have:

$t^{(i)} \in \{-1, 1\}$ and $h_t(x^{(i)}) \in \{-1, 1\}$

- Therefore, we can define:

$$\begin{cases} t^{(i)} h_t(x^{(i)}) = 1, & i \in E^c \quad \text{\# correct classify} \\ t^{(i)} h_t(x^{(i)}) = -1, & i \in E \quad \text{\# misclassify.} \end{cases}$$

- $\boxed{\text{Case 1}}$. When $i \in E^c$, by def^{n} of w_i', we have:

$w_i' \leftarrow w_i \exp(-\alpha_t)$

$\leftarrow w_i \exp(-\frac{1}{2}\log\frac{1-err_t}{err_t})$ # By def^{n} of α_t

$\leftarrow w_i \exp(\log\sqrt{\frac{err_t}{1-err_t}})$

$\leftarrow w_i \sqrt{\frac{err_t}{1-err_t}}$

- $\boxed{\text{Case 2}}$: When $i \in E$, by def^{n} of w_i', we have:

$w_i' \leftarrow w_i \exp(\alpha_t)$

$\leftarrow w_i \exp(\frac{1}{2}\log\frac{1-err_t}{err_t})$

$\leftarrow w_i \exp(\log\sqrt{\frac{1-err_t}{err_t}})$

$\leftarrow w_i \sqrt{\frac{1-err_t}{err_t}}$

- Therefore, $err_t' = \frac{\sum_{i=1}^{N} w_i' I\{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'}$

$$= \frac{\sum_{i \in E} w_i \sqrt{\frac{1-err_t}{err_t}}}{\sum_{i \in E} w_i \sqrt{\frac{1-err_t}{err_t}} + \sum_{i \in E^c} w_i \sqrt{\frac{err_t}{1-err_t}}}$$

$$= \frac{(\frac{1-err_t}{err_t})^{1/2}\sum_{i \in E} w_i}{(\frac{1-err_t}{err_t})^{1/2}\sum_{i \in E} w_i + (\frac{1-err_t}{err_t})^{-1/2}\sum_{i \in E^c} w_i}$$

$$= \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + (\frac{err_t}{1-err_t})\sum_{i \in E^c} w_i} \quad (*)$$

- We also know that:

$$err_t = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E^c} w_i}$$

$\Rightarrow \sum_{i \in E} w_i = err_t(\sum_{i \in E} w_i + \sum_{i \in E^c} w_i)$

$\Rightarrow \sum_{i \in E^c} w_i = (\frac{1-err_t}{err_t})\sum_{i \in E} w_i \quad (**)$

- Plug $(**)$ into $(*)$, we have:

$$err_t' = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + (\frac{err_t}{1-err_t})(\frac{1-err_t}{err_t})\sum_{i \in E} w_i} = \frac{\sum_{i \in E} w_i}{\sum_{i \in E} w_i + \sum_{i \in E} w_i} = \frac{1}{2}$$

Interpretation:
In the class, we know that a single weak classifier is not capable of making the training error very small. It only performs slightly better than chance, i.e.: the error of classifier h according to the given weights $\{w^{(1)}, \dots, w^{(N)}\}$ is at most $\frac{1}{2} - \gamma$ for some $\gamma > 0$.
After the t^{th} iteration, reweight all the weights and apply the old weak learn, now we get the error rate proven above: $err_t' = \frac{1}{2}$, which implies that the error rate is now at its maximum and therefore, we cannot learn anything new with the old weak learner. This result forces us to use a new weak learner in the next iteration in order to decrease the error rate.