CSC 411 Fall 2018
Machine Learning and Data Mining Homework 2

Family name: Bhatia Given name: Pooja

Homework 2

Q1 solution

Ans 1 (a) To Prove that entropy $H(X)$ is non-negative.

X is a discrete random variable and with probability mass function p

Entropy of variable X is

$$H(X) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)} \right) \text{ where}$$

$x \in X$.

$$H(X) = -\sum_x p(x) \log_2 p(x) \quad \rightarrow (1)$$

since, $0 \leq p(x) \leq 1$ (range of $p(x)$) then we can say that $-\log_2 p(x) \geq 0$ $\rightarrow (2)$

from (1) and (2)

we can say that

$$H(X) \geq 0$$

Hence proved Entropy $H(X)$ is non negative.

Q1
Ans
(b)

$$KL(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

If $\phi(x) = -\log_2(x)$

$$\phi'(x) = -\frac{1}{x}$$

$$\phi''(x) = \frac{1}{x^2}$$

for $x \in X$ $\phi''(x)$ is always positive

Hence $\phi(x)$ is Convex.

Also, this can be followed through the Appendix of the assignment where it mentions that $\log_2(x)$ on a set of positive real numbers is concave through which it can be inferred that $-\log_2(x)$ is Convex

NOW, The Jensen's Inequality where X is a random variable

and ϕ be a convex function is

$$E[\phi(X)] \geq \phi(E[X])$$

$$KL(p||q) = \sum_x p(x) \log_2 \left[\frac{p(x)}{q(x)}\right]$$

$$= \sum_x p(x) \{\log_2 (p(x)) - \log_2 (q(x))\}$$

// by applying Log Quotient rule

$$= -\sum_x p(x) \{\log_2 (q(x)) - \log_2 (p(x))\}$$

$$= -\sum_x p(x) \log_2 \left[\frac{q(x)}{p(x)}\right] \rightarrow (1)$$

Now Applying Jensen's Inequality (as stated above) for convex functions in Equation 1

$$KL(p||q) \geq -\log_2 \sum_x p(x) \frac{q(x)}{p(x)}$$

$$\geq -\log_2 \sum_x q(x)$$

$$KL(p\|q) \geq -\log_2 \sum_x q(x)$$

$$\geq -\log_2(1)$$

$$\geq 0$$

$$\boxed{KL(p\|q) \geq 0}$$

Hence proved $KL(p\|q)$ is non-negative.

Ans Q1 (c)

$$I(Y;X) = H(Y) - H(Y|X) \quad \text{——} (1)$$

$$KL(p\|q) = \sum p(x) \log_2 \frac{p(x)}{q(x)} \quad \text{——} (2)$$

To show

$$I(Y;X) = KL(p(x,y) \| p(x)p(y))$$

By applying equation (2) to RHS

$$KL(p(x,y) \| p(x)p(y))$$

$$= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Now according to chain rule which gives relation between joint distribution and conditional distribution

$$p(x,y) = p(y)\,p(x|y) = p(x)\,p(y|x)$$

hence,

$$KL(p(x,y)\,||\,p(x)p(y))$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)\,p(y)}$$

$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(y|x)}{p(y)}$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(y)}{p(y|x)}$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x)$$

$$KL(p(x,y) \,||\, p(x)p(y)) =$$

$$-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x) \quad \rightarrow (3)$$

using formula

(A) $$H(Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y)$$

(B) $$H(Y|X) = \sum_{x\in X} p(x)\, H(Y|X=x)$$

$$= \sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log_2 p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x)$$

Applying formula A and B to equation (3) we get

$$KL(p(x,y)||p(x)p(y)) = H(Y) - H(Y|X)$$

Applying Equation ① stated at the beginning of the solution

$$KL(p(x,y)||p(x)p(y)) = I(Y;X)$$

{since $I(Y;X) = H(Y) - H(Y|X)$}

Hence shown

$$I(Y;X) = KL(p(x,y)||p(x)p(y))$$

Solution Q2

Squared error Loss function is

$$L(y,t) = \frac{1}{2}(y-t)^2 \quad \text{— (1)}$$

Average estimator for $h_1 \dots h_m$ is

$$\bar{h}(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(x) \quad \text{— (2)}$$

To prove,

$$L(\bar{h}(x),t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x),t)$$

That is loss of average estimator is smallar than average loss of the estimators, for any x and t.

from (1) and (2)

$$L(\bar{h}(x),t) = \frac{1}{2}(\bar{h}(x)-t)^2$$

$$= \frac{1}{2}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_i(x)\right) - t\right]^2 \quad \text{— (3)}$$

Now, as mentioned in the Appendix x^p is convex on the set of positive values (real) when $P \geq 1$ or $P \leq 0$.

$\phi(x) = x^2$; $P = 2$ which is $P \geq 1$ hence it is convex.

Jensen's inequality for $\phi(x)$ convex function

$$E[\phi(X)] \geq \phi(E[X])$$

Now, since x^2 is a convex function we apply Jensen's Inequality (convex function)

$$= \frac{1}{2}\left\{\frac{\frac{1}{2}(y_1-t)^2 + \frac{1}{2}(y_2-t)^2 \cdots + \frac{1}{2}(y_m-t)^2}{m}\right\} \geq \phi\left(\frac{\frac{1}{2}(y_1-t) + \frac{1}{2}(y_2-t) \cdots \frac{1}{2}(y_m-t)}{m}\right)$$

Applying ① (squared loss function)

$$= \frac{1}{2}\left\{\frac{L(y_1,t) + L(y_2,t) \cdots L(y_m,t)}{m}\right\} \geq \phi\left(\frac{1}{2m}\left\{(y_1 + \cdots y_m) - mt\right\}\right)$$

$$= \frac{\frac{1}{2}\left\{\frac{1}{2}(y_1-t)^2+\frac{1}{2}(y_2-t)^2 \cdots \frac{1}{2}(y_m-t)^2\right\}}{m} \geq \phi\left(\frac{\frac{1}{2}(y_1-t) \cdots \frac{1}{2}(y_m-t)}{m}\right) \quad (1)$$

Applying ① (Squared Loss function)

$$= \frac{1}{2}\frac{\{L(y_1,t) \cdots L(y_m,t)\}}{m} \geq \phi\left(\frac{1}{2m}\{(y_1+\cdots y_m)-mt\}\right)$$

$$\frac{1}{2m}\sum_{i=1}^{m} L(y_i,t) \geq \left(\frac{1}{2m}\left\{\sum_{i=1}^{m} y_i - mt\right\}\right)^2$$

Since, $y_i = h_i(x)$

$$\frac{1}{2m}\sum_{i=1}^{m} L(h_i(x),t) \geq \left(\frac{1}{2}\left\{\frac{1}{m}\sum_{i=1}^{m} y_i - t\right\}\right)^2$$

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

$$\geq \frac{1}{2}\left(\frac{1}{2}\left[\left(\frac{1}{m}\sum_{i=1}^{m} h_i(x)\right) - t\right]^2\right.$$

from equation ③

$$\frac{1}{2m}\sum_{i=1}^{m} L(h_i(x),t) \geq \frac{1}{2}(\bar{h}(x),t)$$

$$\therefore L(\bar{h}(x),t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x),t)$$

Hence proved

Q3 Ans AdaBoost

target labels are set from $\{-1, +1\}$ and weak Learner returns a classifier whose output also belong to $\{-1, +1\}$.

$$h_t \leftarrow \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \, \mathbb{I}\{h(x^{(i)}) \neq t^{(i)}\}$$ // Incase of weak Learner

$$err_t = \frac{\sum_{i=1}^{N} w_i \, \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i}$$

$$E = \{i : \mathbb{I}\{h_t(x^{(i)}) \neq t^{(i)}\}\}$$

$$E_c = \{i : \mathbb{I}\{h_t(x^{(i)}) = t^{(i)}\}\}$$

so basically,

$$err_t = \frac{\sum_{i \in E} w_i}{\sum_{i=1}^{N} w_i}$$

Now,

$$w_i' = w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)}))$$

where w_i' updated weights.

$$w_i' = w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)}))$$

$$w_i' = \sum_{i: t^{(i)} = h_t(x^{(i)})} w_i \exp(-\alpha_t) + \sum_{i: t^{(i)} \neq h_t(x^{(i)})} w_i \exp(\alpha_t) \quad \rightarrow (1)$$

Now, Reducing and putting $\alpha_t = \frac{1}{2} \log \frac{1-err_t}{err_t}$ in (1)

$$w_i' = (1-err_t)\exp(-\alpha_t) + err_t \exp(\alpha_t)$$

$$\boxed{w_i' = 2\sqrt{err_t(1-err_t)}} \quad (A)$$

(P.T.O) →

Now basically

$$err_t' = \frac{\sum_{i=1}^{N} w_i' \mathbb{I}\{h_t(X_i^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'}$$

from (A)

$$err_t' = \frac{err_t \sqrt{\frac{1-err_t}{err_t}}}{2\sqrt{err_t}\sqrt{1-err_t}}$$

$$= \frac{\sqrt{err_t}\sqrt{1-err_t}}{2\sqrt{err_t}\sqrt{1-err_t}}$$

$$= \frac{1}{2}$$

Ans

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

hence proved.