CSC411H1 L0101, Fall 2018
Assignment 2

Name: JingXi, Hao
Student Number: 1000654188

Due Date: 03 October, 2018 11:59pm

1. **Information Theory.** The goal of this question is to help you become more familiar with the basic equalities and inequalities of information theory. They appear in many contexts in machine learning and elsewhere, so having some experience with them is quite helpful. We review some concepts from information theory, and ask you a few questions.

   Recall the definition of the entropy of a discrete random variable X with probability mass function, $p : H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$. Here the summation is over all possible values of $x \in X$, which (for simplicity) we assume is finite. For example, X might be $1, 2, ..., N$.

   (a) Prove that the entropy H(X) is non-negative.

   Solution:

   By the definition of probability, $p(x)$, where $x \in X$, is quantified as a number between 0 and 1 inclusive. Therefore, we are able to say that $\frac{1}{p(x)} \geq 1$, which indicates that the value of $log_2(\frac{1}{p(x)})$ is non-negative. Then, we have that the value for the product of $p(x)$ and $log_2(\frac{1}{p(x)})$, which is $p(x)log_2(\frac{1}{p(x)})$, is greater than or equal to 0. Hence, the sum of the product for each $x \in X$, $\sum_x p(x)log_2(\frac{1}{p(x)})$, is non-negative since we just sum up the numbers that are either 0 or greater than 0. Consequently, we can conclude that the entropy $H(X) = \sum_x p(x)log_2(\frac{1}{p(x)})$ is non-negative.

   (b) Prove that $KL(p||q)$ is non-negative. Hint: you may want to use Jensen's Inequality, which is described in the Appendix.

   Solution:

   The relative entropy or the KL-divergence of two distributions p and q is defined as, $KL(p||q) = \sum_x p(x) log_2 \frac{p(x)}{q(x)}$. Then, we have that

$$\begin{aligned} KL(p||q) &= \sum_x p(x) log_2 \tfrac{p(x)}{q(x)} \\ &= \sum_x p(x)(-log_2 \tfrac{q(x)}{p(x)}) \text{ \# By law of logarithm} \\ &= \sum_x (-log_2 \tfrac{q(x)}{p(x)}) p(x) \\ &= E(-log_2 \tfrac{q(x)}{p(x)}) \text{ \# By definition of expectation} \\ &= -E(log_2 \tfrac{q(x)}{p(x)}) \end{aligned}$$

$\geq -log_2(E(\frac{q(x)}{p(x)}))$ # since $q(x)$ and $p(x)$ are probabilities, thus, both $q(x)$ and $p(x)$ are numbers between 0 and 1 inclusive. This implies that $\frac{q(x)}{p(x)}$ produces a positive real number. Therefore, $log_2 \frac{q(x)}{p(x)}$ is concave on $\frac{q(x)}{p(x)}$. By Jensen's inequality, we have that $\phi(E(X)) \geq E(\phi(X))$ if $\phi(x)$ is a concave function of x. Then, we have that $log_2(E(\frac{q(x)}{p(x)})) \geq E(log_2 \frac{q(x)}{p(x)})$, which indicates that $-E(log_2 \frac{q(x)}{p(x)}) \geq -log_2(E(\frac{q(x)}{p(x)}))$

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

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

$= -log_2(1)$ # By property of probability, $\sum_x q(x) = 1$, where $x \in X$

$= 0$

Therefore, we are able to conclude that $KL(p||q) \geq 0$ is non-negative.

(c) The Information Gain or Mutual Information between X and Y is $I(Y;X) = H(Y) - H(Y|X)$. Show that

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

, where $p(x) = \sum_y p(x,y)$ is the marginal distribution of X.

Solution:

Since $I(Y;X) = H(Y) - H(Y|X)$, we have that

$I(Y;X) = H(Y) - H(Y|X)$

$= H(Y) - (H(X,Y) - H(X))$ # By chain rule, $H(X,Y) = H(Y|X) + H(X)$

$= H(Y) - H(X,Y) + H(X)$

$= -\sum_y p(y) log_2 p(y) - (-\sum_x \sum_y p(x,y) log_2 p(x,y)) + (-\sum_x p(x) log_2 p(x))$

$= -\sum_y p(y) log_2 p(y) + \sum_x \sum_y p(x,y) log_2 p(x,y) - \sum_x p(x) log_2 p(x)$

$= -\sum_y \sum_x p(x,y) log_2 p(y) + \sum_x \sum_y p(x,y) log_2 p(x,y) - \sum_x \sum_y p(x,y) log_2 p(x)$

$= -\sum_x \sum_y p(x,y) log_2 p(y) + \sum_x \sum_y p(x,y) log_2 p(x,y) - \sum_x \sum_y p(x,y) log_2 p(x)$

$= \sum_x \sum_y (p(x,y) log_2 p(x,y) - p(x,y) log_2 p(x) - p(x,y) log_2 p(y))$

$= \sum_x \sum_y (p(x,y)(log_2 p(x,y) - log_2 p(x) - log_2 p(y)))$

$= \sum_x \sum_y (p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)})$

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

Hence, we have proved that $I(Y;X) = KL(p(x,y)||p(x)p(y))$.

2. **Benefit of Averaging.** Consider m estimators $h_1, \ldots, h_m$, each of which accepts an input x and produces an output y, i.e., $y_i = h_i(x)$. These estimators might be generated through a Bagging procedure, but that is not necessary to the result that we want to prove. Consider the squared error loss function $L(y,t) = \frac{1}{2}(y-t)^2$. Show that the loss of the average estimator

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

is smaller than the average loss of the estimators. That is, for any x and t, we have

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

Hint: you may want to use Jensen's Inequality, which is described in the Appendix.

Solution:

We show that the loss of average estimator is smaller than the average loss of the estimators. Then, we have that

$L(\bar{h}(x), t) = L(\frac{1}{m}\sum_{i=1}^{m} h_i(x), t)$ # Substitute the $\bar{h}(x)$ into the equation

$= L(E(h_i(x), t)$ # Since expected value is just mean (average value), thus, $E(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} x_i$

$= \frac{1}{2}(E(h_i(x)) - t)^2$ # Since $L(y,t) = \frac{1}{2}(y-t)^2$

$= \frac{1}{2}(E(h_i(x)) - E(t))^2$

$= \frac{1}{2}(E(h_i(x) - t))^2$

$\leq \frac{1}{2}E((h_i(x) - t)^2)$ # Let $\phi(z)$ be z^2 and z_i be $h_i(x) - t$. Thus, $\phi(z)$ is a convex function of z, which indicates that we are able to use Jensen's Inequality, $\phi(E(X)) \leq E(\phi(X))$. Then, we have that $(E(h_i(x) - t))^2 \leq E((h_i(x) - t)^2)$. After multiplying $\frac{1}{2}$ for both sides, we obtain that $\frac{1}{2}(E(h_i(x) - t))^2 \leq \frac{1}{2}E((h_i(x) - t)^2)$

$= \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m}(h_i(x) - t)^2$ # Since $E(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} x_i$

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

$= \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t)$ # Since $L(y,t) = \frac{1}{2}(y-t)^2$

Hence, we have proved that the loss of average estimator is smaller than the average loss of the estimators, $L(\bar{h}(x), t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t)$.

3. **AdaBoost.** The goal of this question is to show that the AdaBoost algorithm changes the weights in order to force the weak learner to focus on difficult data points. Here we consider the case that the target labels are from the set $1, +1$ and the weak learner also returns a classifier whose outputs belongs to $1, +1$ (instead of $0, 1$). Consider the t^{th} iteration of AdaBoost, where the weak learner is

$$h_t \leftarrow \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\} \neq t^{(i)})$$

, the w-weighted classification error is

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

, and the classifier coefficient is

$$\alpha_t = \frac{1}{2} log \frac{1 - err_t}{err_t}$$

. (Here, log denotes the natural logarithm.) AdaBoost changes the weights of each sample depending on whether the weak learner h_t classifies it correctly or incorrectly. The updated weights for sample i is denoted by w_i' and is

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

. Show that the error w.r.t. $(w_i' \ldots w_N')$ is exactly $\frac{1}{2}$. That is, show that

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

. Note that here we use the weak learner of iteration t and evaluate it according to the new weights, which will be used to learn the $t+1$-st weak learner. What is the interpretation of this result?

Solution:

In order to show that the error w.r.t. $(w_i' \ldots w_N')$ is exactly $\frac{1}{2}$, we first divide the samples into two sets of $E = \{i : \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}\}$ and its complement $E^c = \{i : \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}\}$. Then, we have that

$$\begin{aligned}
err_t' &= \frac{\sum_{i=1}^{N} w_i' \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i'} \\
&= \frac{\sum_{i \in E} w_i'}{\sum_{i \in E} w_i' + \sum_{j \in E^c} w_j'} \\
&= \frac{\sum_{i \in E} w_i exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)}))}{\sum_{i \in E} w_i exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})) + \sum_{j \in E^c} w_j exp(-\alpha_t t^{(j)} h_t(\mathbf{x}^{(j)}))} \quad \text{\# Substitute } w_i' \text{ and } w_j' \text{ into the equation} \\
&= \frac{\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\})}{\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}) + \sum_{j \in E^c} w_j exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(j)}) \neq t^{(j)}\})} \quad \text{\# since } w_i' \leftarrow w_i exp(-\alpha_t t^{(i)} h_t(\mathbf{x}^{(i)})) \equiv \\
&\qquad w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}) \\
&\qquad \text{and same for } w_j'
\end{aligned}$$

For the term $\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\})$, $\mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}$ always evaluates to be 1 since $i \in E$. Therefore, we are able to simplify this term as follows,

$$\begin{aligned}
\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}) &= \sum_{i \in E} w_i exp(2\alpha_t) \\
&= \sum_{i \in E} w_i exp((2)(\tfrac{1}{2} log \tfrac{1 - err_t}{err_t})) \quad \text{\# Substitue } \alpha_t \text{ into the equation} \\
&= \sum_{i \in E} w_i exp(log \tfrac{1 - err_t}{err_t}) \\
&= \sum_{i \in E} w_i \tfrac{1 - err_t}{err_t} \\
&= \sum_{i \in E} w_i \frac{(1 - \frac{\sum_{k \in E} w_k}{\sum_{k=1}^{N} w_k})}{(\frac{\sum_{k \in E} w_k}{\sum_{k=1}^{N} w_k})} \quad \text{\# By hint, } \frac{\sum_{i \in E} w_i}{\sum_{i=1}^{N} w_i} = err_t
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i \in E} w_i \frac{(\frac{\sum_{k=1}^{N} w_k - \sum_{k \in E} w_k}{\sum_{k=1}^{N} w_k})}{(\frac{\sum_{k \in E} w_k}{\sum_{k=1}^{N} w_k})} \\
&= \sum_{i \in E} w_i \frac{\sum_{k=1}^{N} w_k - \sum_{k \in E} w_k}{\sum_{k \in E} w_k} \\
&= (\frac{\sum_{k=1}^{N} w_k - \sum_{k \in E} w_k}{\sum_{k \in E} w_k}) \sum_{i \in E} w_i \\
&= \sum_{k=1}^{N} w_k - \sum_{k \in E} w_k \quad \# \text{ Since } \sum_{k \in E} w_k = \sum_{i \in E} w_i \\
&= \sum_{j \in E^c} w_j
\end{aligned}
$$

Similarly, we can simplify the term $\sum_{j \in E^c} w_j exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(j)}) \neq t^{(j)}\})$. Since $j \in E^c$, $\mathbb{I}\{h(\mathbf{x}^{(j)}) \neq t^{(j)}\}$ always equals to 0. Then, we have that $\sum_{j \in E^c} w_j exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(j)}) \neq t^{(j)}\}) = \sum_{j \in E^c} w_j exp(0) = \sum_{j \in E^c} w_j$.

Consequently, the equation, $err'_t = \frac{\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\})}{\sum_{i \in E} w_i exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(i)}) \neq t^{(i)}\}) + \sum_{j \in E^c} w_j exp(2\alpha_t \mathbb{I}\{h(\mathbf{x}^{(j)}) \neq t^{(j)}\})}$, can be simplified as $err'_t = \frac{\sum_{j \in E^c} w_j}{\sum_{j \in E^c} w_j + \sum_{j \in E^c} w_j} = \frac{1}{2}$. Then, we have shown that the error w.r.t. $(w'_i \ldots w'_N)$ is exactly $\frac{1}{2}$.