CSC411 A2

Zhihong Wang 1002095207

2018-10-03

Q1:
(a). Prove that the entropy $H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$ is non-negative.
Note: $p(x) \in [0, 1]$. If $p(x) = 0$, then $H(x) = 0$ by definition.
We can only consider $p(x) \in (0, 1]$ now.

Proof: Since $p(x) \in (0, 1]$, we can get $\frac{1}{p(x)} \in [1, \infty)$,
then $log_2(\frac{1}{p(x)}) \in [0, \infty)$,
then $p(x) log_2(\frac{1}{p(x)}) \in [0, \infty)$,
then $\sum_x p(x) log_2(\frac{1}{p(x)}) \in [0, \infty)$,
Therefore $H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$ is non-negative.

We can also do the proof by contradiction:
Assume $H(x)$ is negative,
then $log_2(\frac{1}{p(x)})$ is negative (since $p(x) \in (0, 1]$),
then $\frac{1}{p(x)} < 1$,
then $p(x) > 1$, $\sharp Contradiction!$
Therefore $H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$ is non-negative.

(b). Prove that $KL(p||q)$ is non-negative.
Note: $\sum_x f(x)p(x) = E[f(x)], f(x) = log\frac{p(x)}{q(x)}$, by definition of expectation for random variable (i.e. $f(x)$) function.

$KL(p||q) = \sum_x p(x) log\frac{p(x)}{q(x)}$
$= -\sum_x p(x) log\frac{q(x)}{p(x)}$
$= E[-log\frac{q(x)}{p(x)}]$
$\geq -log E[\frac{q(x)}{p(x)}]$
(By Jensen's Inequality: $\phi(E[X]) \leq E[\phi(X)]$, $\phi(x)$ is convex (e.g. $-log$))
$= -log \sum p(x)\frac{q(x)}{p(x)}$
$= -log \sum q(x)$
$= -log 1 = 0$
Therefore $KL(p||q) = \sum_x p(x) log\frac{p(x)}{q(x)} \geq 0$.

(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 $p(x) = \sum_y p(x,y), p(y|x) = \frac{p(x,y)}{p(x)}$ is the marginal distribution of X.

$$\begin{aligned}
I(Y;X) &= H(Y) - H(Y|X) \\
&= \textstyle\sum_y p(y)log(\frac{1}{p(y)}) - \sum_x p(y|x)log(\frac{1}{p(y|x)}) \\
&= \textstyle\sum_y p(y)log(\frac{1}{p(y)}) - \sum_x p(x)H(Y|X=x) \\
&= \textstyle\sum_y p(y)log(\frac{1}{p(y)}) + \sum_x \sum_y p(x,y)log(p(y|x))
\end{aligned}$$

($H(Y|X)$ equations can be found at Lec03 slides)

$$\begin{aligned}
KL(p(x,y)||p(x)p(y)) &= \textstyle\sum_x \sum_y p(x,y)log\frac{p(x,y)}{p(x)p(y)} \\
&= \textstyle\sum_x \sum_y p(x,y)log\frac{p(y|x)}{p(y)} \\
&= \textstyle\sum_x \sum_y p(x,y)log(p(y|x)) - \sum_y p(y)log(p(y)) \\
&= \textstyle\sum_y p(y)log(\frac{1}{p(y)}) + \sum_x \sum_y p(x,y)log(p(y|x))
\end{aligned}$$

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

Q2:
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
$L(\bar{h}(x),t) \leq \frac{1}{m}\sum_{i=1}^{m} L(h_i(x),t)$

$$\begin{aligned} L(\bar{h}(x),t) &= \tfrac{1}{2}(\bar{h}(x)-t)^2 \\ &= \tfrac{1}{2}((\bar{h}(x))^2 - 2t\bar{h}(x) + t^2) \\ &= \tfrac{1}{2}(\bar{h}(x))^2 - t\bar{h}(x) + \tfrac{1}{2}t^2 \\ &= \tfrac{1}{2}(\tfrac{1}{m}\textstyle\sum_{i=1}^{m} h_i(x))^2 - \tfrac{t}{m}\textstyle\sum_{i=1}^{m} h_i(x) + \tfrac{1}{2}t^2 \\ &= \tfrac{1}{2m^2}(\textstyle\sum_{i=1}^{m} h_i(x))^2 - \tfrac{t}{m}\textstyle\sum_{i=1}^{m} h_i(x) + \tfrac{1}{2}t^2 \end{aligned}$$

$$\begin{aligned} \tfrac{1}{m}\textstyle\sum_{i=1}^{m} L(h_i(x),t) &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \tfrac{1}{2}(h_i(x)-t)^2 \\ &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m}(\tfrac{1}{2}h_i(x)^2 - th_i(x) + \tfrac{1}{2}t^2) \\ &= \tfrac{1}{m}(\textstyle\sum_{i=1}^{m}\tfrac{1}{2}h_i(x)^2 - \textstyle\sum_{i=1}^{m} th_i(x) + \textstyle\sum_{i=1}^{m}\tfrac{1}{2}t^2) \\ &= \tfrac{1}{m}(\tfrac{1}{2}\textstyle\sum_{i=1}^{m} h_i(x)^2 - t\textstyle\sum_{i=1}^{m} h_i(x) + m\tfrac{1}{2}t^2) \\ &= \tfrac{1}{2m}\textstyle\sum_{i=1}^{m} h_i(x)^2 - \tfrac{t}{m}\textstyle\sum_{i=1}^{m} h_i(x) + \tfrac{1}{2}t^2 \end{aligned}$$

By Jensen's Inequality: $\phi(E[X]) \leq E[\phi(X)]$, $\phi(x)$ is convex (e.g. x^2)
so $(E[X])^2 \leq E[X^2]$
then $(E[h(x)])^2 \leq E[h(x)^2]$
then $(\sum_{i=1}^{m} h_i(x))^2 \leq \sum_{i=1}^{m} h_i(x)^2$
then $\frac{1}{2m^2}(\sum_{i=1}^{m} h_i(x))^2 \leq \frac{1}{2m}\sum_{i=1}^{m} h_i(x)^2$, since $m \geq 1$

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

Q3:

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

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

$\sum_{i=1}^{N} w_i^{'} || \{h_t(x^{(i)}) \neq t^{(i)}\} = \sum_{i \in E} w_i^{'}$,

$\sum_{i=1}^{N} w_i^{'} = \sum_{i \in E} w_i^{'} + \sum_{i \in E^C} w_i^{'}$

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

When $i \in E$, $t^{(i)} = 1/-1, h_t(x^{(i)}) = -1/1$, (i.e. $t^{(i)}$ and $h_t(x^{(i)})$ are different), so $t^{(i)} \times h_t(x^{(i)}) = -1$

When $i \in E^C$, $t^{(i)} = 1/-1, h_t(x^{(i)}) = 1/-1$, (i.e. $t^{(i)}$ and $h_t(x^{(i)})$ are same), so $t^{(i)} \times h_t(x^{(i)}) = 1$

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

$err_t^{'} = \frac{\sum_{i=1}^{N} w_i^{'} || \{h_t(x^{(i)}) \neq t^{(i)}\}}{\sum_{i=1}^{N} w_i^{'}} = \frac{\sum_{i \in E} w_i^{'}}{\sum_{i \in E} w_i^{'} + \sum_{i \in E^C} w_i^{'}}$

$$\begin{aligned}
\sum_{i \in E} w_i^{'} &= \sum_{i \in E} w_i exp(-\alpha_t t^{(i)} h_t(x^{(i)})) \\
&= \sum_{i \in E} w_i exp(\alpha_t) \\
&= \sum_{i \in E} w_i exp(\tfrac{1}{2} log \tfrac{1-err_t}{err_t}) \\
&= \sum_{i \in E} w_i exp(log(\tfrac{1-err_t}{err_t})^{\frac{1}{2}}) \\
&= \sum_{i \in E} w_i (\tfrac{1-err_t}{err_t})^{\frac{1}{2}}
\end{aligned}$$

$$\begin{aligned}
\sum_{i \in E^C} w_i^{'} &= \sum_{i \in E^C} w_i exp(-\alpha_t t^{(i)} h_t(x^{(i)})) \\
&= \sum_{i \in E^C} w_i exp(-\alpha_t) \\
&= \sum_{i \in E^C} w_i exp(-\tfrac{1}{2} log \tfrac{1-err_t}{err_t}) \\
&= \sum_{i \in E^C} w_i exp(log(\tfrac{1-err_t}{err_t})^{-\frac{1}{2}}) \\
&= \sum_{i \in E^C} w_i (\tfrac{1-err_t}{err_t})^{-\frac{1}{2}}
\end{aligned}$$

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

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

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

Then, $err_t^{'} = \frac{1}{1 + \frac{\sum_{i \in E^C} w_i}{\sum_{i \in E} w_i} (\frac{1-err_t}{err_t})^{-1}}$

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

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

Then, $err_t^{'} = \frac{1}{1+\frac{\sum_{i\in E^C} w_i}{\sum_{i\in E} w_i}(\frac{err_t}{1-err_t})}$

$= \frac{1}{1+\frac{\sum_{i\in E^C} w_i}{\sum_{i\in E} w_i}(\frac{\sum_{i\in E} w_i}{\sum_{i\in E^C} w_i})}$

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

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

Interpretation:

By using a weak learner h and the weight w, we can get the training error $E \leq \frac{1}{2} - \epsilon, \exists \epsilon > 0$. However, if we still fit a weak learner h with the same weight w to the training set, we can only get the same result without any improvement. That's why we have to re-weight the training set. We are able to increase the classification error of h to $\frac{1}{2}$ (i.e. the maximum of error rate, means the prediction is half-wrong), if h is classified wrongly, by using the re-weight formula from previous. Then, the next weak learner can be more focus on the wrongly classified data during iterations.

Because the increasing weight on the points that h classified incorrectly and more wrongly classified data that the weak learner analyzed, at every iteration, we can generate a better weak learner that fitted to the training set to decrease the error.