CSC2515 Fall 2018
Homework 2

Soon Chee Loong
999295793
cheeloong.soon@mail.utoronto.ca
Last Name: Soon
First Name: Chee Loong
cdf markus: soonchee

1. Information Theory

Entropy of Discrete Random Variable

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

a) Prove entropy $H(x)$ is non-negative.

Since $P(x)$ is a probability distribution over discrete random variable X,

$\Rightarrow P(x) \in [0,1] \ \forall x$, $\sum_x p(x) = 1$

Properties of log

$\log(x) < 0 \iff x < 1$

$\log(x) = 0 \iff x = 1$

$\log(x) > 0 \iff x > 1$

$\log\left(\frac{1}{x}\right) = \log(1) - \log(x) = -\log(x)$

Entropy of Discrete Random Variable

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

Since $p(x) \in [0,1]$, $\Rightarrow \log_2(p(x)) \in (-\infty, 0)$

$\Rightarrow p(x) \geq 0$, $\log_2(p(x)) < 0$

$$\Rightarrow H(X) = \underbrace{\underbrace{-}_{<0} \underbrace{\sum_x \underbrace{p(x)}_{\geq 0} \underbrace{\log_2(p(x))}_{<0}}_{\leq 0}}_{\geq 0} \geq 0$$

$\therefore$, proven that $H(x) \geq 0$ since $\forall \left(-p(x) \log_2(p(x))\right) \geq 0$ in the summation.

Relative Entropy = KL-divergence

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

b) Prove that $KL(p\|q)$ is non-negative. Assume $p(x) > 0$, $q(x) > 0$

Hint: Use Jensen's Inequality $\phi(E[X]) \leq E[\phi(X)]$

Show $-\log(x)$ is a convex function

$f(x) = \log(x)$

$$\frac{d f(x)}{dx} = \frac{1}{x}$$

$$\frac{d^2 f(x)}{dx} = -\frac{1}{x^2} < 0 \text{, since } x^2 \geq 0$$

$\Rightarrow \log(x)$ is a concave function

$\Rightarrow -\log(x)$ is a convex function since $\frac{d^2(-\log(x))}{dx^2} = \frac{1}{x^2} > 0$, $x^2 \geq 0$

For convex function,

Jensen's Inequality

$\phi(E[X]) \leq E[\phi(X)]$, $\phi(x)$ is a convex function of x

Relative Entropy = KL-divergence

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

b) Prove that $KL(p\|q)$ is non-negative. Assume $p(x) > 0$, $q(x) > 0$

Hint: Use Jensen's Inequality $\phi(E(x)) \leq E[\phi(x)]$

Show $-\log(x)$ is a convex function

$f(x) = \log(x)$

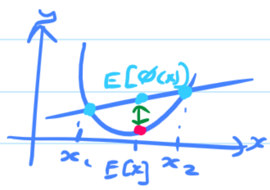
$\frac{d f(x)}{dx} = \frac{1}{x}$

$\frac{d^2 f(x)}{dx} = -\frac{1}{x^2} < 0$, since $x^2 \geq 0$

$\Rightarrow \log(x)$ is a concave function

$\Rightarrow -\log(x)$ is a convex function since $\frac{d^2(-\log(x))}{dx^2} = \frac{1}{x^2} > 0$, $x^2 \geq 0$

For convex function,



Jensen's Inequality

$\phi(E(x)) \leq E[\phi(x)]$, $\phi(x)$ is a convex function of x

Relative Entropy

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

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

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

$$\geq -\log_2\left(E_{p(x)}\left(\frac{q(x)}{p(x)}\right)\right)$$

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

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

$$= -\log_2(1)$$

$$= 0$$

$\therefore$, $KL(p\|q) \geq 0$

proven that relative entropy is non-negative

c) Information Gain = Mutual Information between X and Y

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

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

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

$H(Y) = -\sum_y p(y)\log(p(y))$

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

$$= -\sum \sum p(x, y)\log(p(y))$$

c) Information Gain = Mutual Information between X and Y

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

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

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

$H(Y) = -\sum_y p(y) \log(p(y))$

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

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

Conditional Entropy

$H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log(p(y|x))$

$= -\sum_x \sum_y p(x)p(y|x) \log(p(y|x))$, $p(x)$ is constant w.r.t. y

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

$= -\sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)}\right)$, $p(y|x) = \frac{p(x,y)}{p(x)}$

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

$= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

Mutual Information

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

$= -\sum_x \sum_y p(x,y) \log(p(y)) - \left(-\sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)}\right)\right)$

$= \sum_x \sum_y p(x,y) \left(\log\left(\frac{p(x,y)}{p(x)}\right) - \log(p(y))\right)$

$= \sum_x \sum_y p(x,y) \left(\log\left(\left(\frac{p(x,y)}{p(x)}\right)\left(\frac{1}{p(y)}\right)\right)\right)$

$= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

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

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

2. Benefit of Averaging

Consider m estimators $h_1, \dots, h_m$, each which accepts an input x and produces an output y, $y_i = h_i(x)$

Consider squared error loss function $L(y,t) = \frac{1}{2}(y-t)^2$

Average Estimator

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

2. Benefit of Averaging

Consider m estimators $h_1, \dots, h_m$, each which accepts an input x and produces an output y, $y_i = h_i(x)$

Consider squared error loss function $L(y,t) = \frac{1}{2}(y-t)^2$

Average Estimator

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

Show loss of Average Estimator is smaller than average loss of the estimators

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

Hint: Use Jensen's Inequality

Show Loss Function is convex w.r.t. y to use Jensen's Inequality

We need to show that the loss function $L(x,t)$ is convex.

Since t remains a fixed constant,

we can think of the loss function as a function of a single variable

$$L(x,t) = L(x) \quad \text{since } t \text{ is constant}$$
$$= \frac{1}{2}(x-t)^2$$

Need to show $L(x)$ is convex

$$\frac{dL(x)}{dx} = (x-t)$$

$$\frac{d^2L(x)}{dx^2} = 1 > 0 \Rightarrow L(x) \text{ is convex}$$

Now that we have showed the loss function is indeed convex, we can use Jensen's Inequality.

Use Jensen's Inequality to show

loss of Average Estimator is smaller than average loss of the estimators

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

$= E[h_i(x)]$, expectation is uniformly distributed and use Law of Large Numbers

From Jensen's Inequality $\phi(E[x]) \le E[\phi(x)]$, $\phi(x)$ is convex

and because $L(x)$ is convex

$$\Rightarrow L(E[h_i(x)]) \le E[L(h_i(x))]$$

$$L(E[h_i(x)]) = L(\bar{h}(x)) = L(\bar{h}(x), t)$$

$$E[L(h_i(x))]$$
$$= \frac{1}{m}\sum_{i=1}^{m} L(h_i(x)) \quad \text{, since expectation is uniformly distributed}$$
$$= \frac{1}{m}\sum_{i=1}^{m} L(h_i(x), t)$$

$\therefore$, proven

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

3. AdaBoost

Show AdaBoost algorithm changes weights in order to force weak learner to focus on more difficult points.

Consider case that target labels are from set $\{-1, +1\}$ and weak learner also returns a classifier whose output belongs to $\{-1, +1\}$ (instead of $\{0, 1\}$)

Consider t-th iteration of AdaBoost,

weak learner

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

w-weighted classification error

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

Classifier Coefficient

$$\alpha_t = \frac{1}{2} \log_e\left(\frac{1 - err_t}{err_t}\right)$$

Updated weights for sample i

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

Show error w.r.t. updated weights $(w_1', \dots, w_N')$ is exactly $\frac{1}{2}$.

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

Hint: Start from err_t' and divide the summation to two sets:

$E = \{i : h_t(x^{(i)}) \neq t^{(i)}\}$ ⇒ set of misclassification

$E^c = \{i : h_t(x^{(i)}) = t^{(i)}\}$ ⇒ set of correct classification

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

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

$$= \frac{\sum_{k \in E^c} w_k' (0) + \sum_{i \in E} w_i' (1)}{\sum_{j=1}^{N} w_j'}$$, split numerator

$$= \frac{\sum_{i \in E} w_i'}{\sum_{j=1}^{N} w_j'}$$, correct classifications in numerator disappears due to multiplication with 0

Hint: Start from err_t' and divide the summation to two sets:

$E = \{i : h_t(x^{(i)}) \neq t^{(i)}\}$ ⇒ set of misclassification

$E^c = \{i : h_t(x^{(i)}) = t^{(i)}\}$ ⇒ set of correct classification

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

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

$$= \frac{\sum_{k \in E^c} w_k'(0) + \sum_{i \in E} w_i'(1)}{\sum_{j=1}^{N} w_j'}$$, split numerator

$$= \frac{\sum_{i \in E} w_i'}{\sum_{j=1}^{N} w_j'}$$, correct classifications in numerator disappears due to multiplication with 0

$$= \frac{\sum_{i \in E} w_i \exp(-\alpha_t t^{(i)} h_t(x^{(i)}))}{\sum_{j=1}^{N} w_j \exp(-\alpha_t t^{(j)} h_t(x^{(j)}))}$$, rewrite as weights before update

$$= \frac{\sum_{i \in E} w_i e^{\alpha_t}}{\sum_{j \in E^c} w_j e^{-\alpha_t} + \sum_{k \in E} w_k e^{\alpha_t}}$$, split denominator

$$= \frac{\sum_{i \in E} w_i e^{\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)}}{\sum_{j \in E^c} w_j e^{-\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)} + \sum_{k \in E} w_k e^{\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)}}$$

$$= \frac{\sum_{i \in E} w_i e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{\frac{1}{2}}}}{\sum w_j e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{-\frac{1}{2}}} + \sum w_k e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{1/2}}}$$

$$= \frac{\sum_{i\in E} w_i e^{\alpha_t}}{\sum_{j\in E^c} w_j e^{-\alpha_t} + \sum_{k\in E} w_k e^{\alpha_t}} \quad \text{, split denominator}$$

$$= \frac{\sum_{i\in E} w_i \, e^{\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)}}{\sum_{j\in E^c} w_j e^{-\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)} + \sum_{k\in E} w_k e^{\frac{1}{2}\log_e\left(\frac{1-err_t}{err_t}\right)}}$$

$$= \frac{\sum_{i\in E} w_i \, e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{\frac{1}{2}}}}{\sum_{j\in E_c} w_j e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{-\frac{1}{2}}} + \sum_{k\in E} w_k e^{\log_e\left(\frac{1-err_t}{err_t}\right)^{1/2}}}$$

$$= \frac{\sum_{i\in E} w_i \sqrt{\frac{1-err_t}{err_t}}}{\sum_{j\in E_c} w_j \sqrt{\frac{err_t}{1-err_t}} + \sum_{k\in E} w_k \sqrt{\frac{1-err_t}{err_t}}} \quad , e^{\ln(x)} = x$$

$$= \frac{\sqrt{\frac{1-err_t}{err_t}} \sum_{i\in E} w_i}{\sqrt{\frac{err_t}{1-err_t}} \sum_{j\in E_c} w_j + \sqrt{\frac{1-err_t}{err_t}} \sum_{k\in E} w_k} \quad \text{, normalize}$$

$$= \frac{\left(\sqrt{\frac{1-err_t}{err_t}} \sum_{i\in E} w_i\right)}{\left(\sqrt{\frac{err_t}{1-err_t}} \sum_{j\in E_c} w_j + \sqrt{\frac{1-err_t}{err_t}} \sum_{k\in E} w_k\right)} \quad \frac{\left(\frac{1}{\sum_{m=1}^{N} w_m}\right)}{\left(\frac{1}{\sum_{m=1}^{N} w_m}\right)}$$

$$= \frac{\left(\sqrt{\frac{1-err_t}{err_t}}\right)(err_t)}{\sqrt{\frac{err_t}{1-err_t}}\,(1-err_t) + \sqrt{\frac{1-err_t}{err_t}}\,(err_t)} \quad , err_t = \frac{\sum_{i\in E} w_i}{\sum_{j=1}^{N} w_j}$$

$$1-err_t = \sum w_j$$

$$= \frac{\sqrt{\frac{1-err_t}{err_t}} \sum_{i \in E} w_i}{\sqrt{\frac{err_t}{1-err_t}} \sum_{j \in E_c} w_j + \sqrt{\frac{1-err_t}{err_t}} \sum_{k \in E} w_k} \quad \text{, normalize}$$

$$= \frac{\left(\sqrt{\frac{1-err_t}{err_t}} \sum_{i \in E} w_i\right)}{\left(\sqrt{\frac{err_t}{1-err_t}} \sum_{j \in E_c} w_j + \sqrt{\frac{1-err_t}{err_t}} \sum_{k \in E} w_k\right)} \quad \frac{\left(\frac{1}{\sum_{m=1}^{N} w_m}\right)}{\left(\frac{1}{\sum_{m=1}^{N} w_m}\right)}$$

$$= \frac{\left(\sqrt{\frac{1-err_t}{err_t}}\right)(err_t)}{\sqrt{\frac{err_t}{1-err_t}}(1-err_t) + \sqrt{\frac{1-err_t}{err_t}}(err_t)} \quad , err_t = \frac{\sum_{i \in E} w_i}{\sum_{j=1}^{N} w_j} \quad , 1-err_t = \frac{\sum_{j \in E_c} w_j}{\sum_{i=1}^{N} w_i}$$

$$= \frac{\sqrt{\frac{(1-err_t)(err_t)^2}{(err_t)}}}{\sqrt{\frac{err_t(1-err_t)^2}{(1-err_t)}} + \sqrt{\frac{(1-err_t)(err_t)^2}{(err_t)}}} \quad , \sqrt{x^2} = x$$

$$= \frac{\sqrt{(1-err_t)(err_t)}}{\sqrt{err_t(1-err_t)} + \sqrt{(1-err_t)(err_t)}}$$

$$= \frac{\sqrt{(1-err_t)(err_t)}}{\sqrt{err_t(1-err_t)} + \sqrt{(1-err_t)(err_t)}}$$

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

$$= \frac{\sqrt{(1-err_t)(err_t)}}{\sqrt{err_t(1-err_t)} + \sqrt{(1-err_t)(err_t)}}$$

$$= \frac{\sqrt{(1-err_t)(err_t)}}{\sqrt{err_t(1-err_t)} + \sqrt{(1-err_t)(err_t)}}$$

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

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

$\therefore$ proven

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

We use weak learner of iteration t and evaluate it according to the new weights, which will be used to learn the $(t+1)_{th}$ weak learner.

What is the interpretation of this result?

The interpretation is that the weighted error for the learner at iteration t with respect to the weights for iteration $(t+1)_{th}$ is exactly $\frac{1}{2}$.

You can think of this as a random guess with respect to the updated weighted error.

Because of this, we can always train or select another classifier for next iteration that achieves an error $< \frac{1}{2}$ which means it performs better.

To see why this is true,

if the new classifier has error $\beta > \frac{1}{2}$,

we can simply reverse its predictions

such that its error is $(1-\beta) < \frac{1}{2}$

We use weak learner of iteration t and evaluate it according to the new weights, which will be used to learn the $(t+1)_{th}$ weak learner.

What is the interpretation of this result?

The interpretation is that the weighted error for the learner at iteration t with respect to the weights for iteration $(t+1)_{th}$ is exactly $\frac{1}{2}$.

You can think of this as a random guess with respect to the updated weighted error.

Because of this, we can always train or select another classifier for next iteration that achieves an error $< \frac{1}{2}$ which means it performs better.

To see why this is true,

if the new classifier has error $\beta > \frac{1}{2}$,

we can simply reverse its predictions

such that its error is $(1-\beta) < \frac{1}{2}$

$\therefore$ this allows Adaptive Boosting to continue iterating such that it performs better than the previous combination of ensembles to overfit the training set.