Brendan Neal 1001160236 | nealbre1 2018-09-29

CSC411 Homework 2

Question 1: Information Theory

Question 1a: Proof of Non-Negative Entropy

Definition of entropy

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

Note that $p(x) = 0 \Rightarrow p(x) \log p(x) = 0$. Therefore, the above summation effectively ignores $p(x) = 0$. Also note that all probabilities belong in the inclusive range $0 \leq p(x) \leq 1$.

$$(p(x) \neq 0) \wedge (0 \leq p(x) \leq 1) \Rightarrow 0 < p(x) \leq 1$$

Taking the log of the above:

$$\Rightarrow \log(0) < \log p(x) \leq \log(1)$$

$$\Rightarrow -\infty < \log p(x) \leq 0$$

Since $p(x)$ is positive, multiplying by $p(x)$ means that the range stays the same:

$$\Rightarrow -\infty < p(x) \log p(x) \leq 0$$

Since all $p(x)$ in the summation have the same properties as the generic $p(x)$ which was just analyzed:

$$(\forall x \in X : -\infty < p(x) \log p(x) \leq 0)$$

$$\Rightarrow -\infty < \sum_{x \in X} p(x) \log p(x) \leq 0$$

Negating the summation to obtain the definition of entropy

$$\Rightarrow 0 \leq -\sum_{x \in X} p(x) \log p(x) < \infty$$

$$\Rightarrow 0 \leq H(X) < \infty$$

Therefore, $H(X)$ is non-negative.

Question 1b: Proof of Non-Negative KL-Divergence

Definition of KL-Divergence

$$K(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Since $\log(x^{-1}) = -\log(x)$

$$\Rightarrow -K(p||q) = \sum_x p(x) \log \frac{q(x)}{p(x)}$$

Note that the above is an expectation

$$\Rightarrow \; -K(p||q) = \sum_x p(x) \log \frac{q(x)}{p(x)} = E\left(\log \frac{q(x)}{p(x)}\right)$$

Jensen's inequality for **concave** function f

$$E\big(f(x)\big) \leq f\big(E(x)\big)$$

Since log is **concave** over the positive real numbers and $\frac{q(x)}{p(x)}$ is positive, Jensen's inequality can be applied

$$\Rightarrow E\left(\log \frac{q(x)}{p(x)}\right) \leq \log E\left(\frac{q(x)}{p(x)}\right)$$

Expanding the second expectation

$$\Rightarrow E\left(\log \frac{q(x)}{p(x)}\right) \leq \log\left(\sum_x p(x)\left(\frac{q(x)}{p(x)}\right)\right) = \log\left(\sum_x q(x)\right)$$

Since the sum of probabilities over the entire set of events is 1

$$\Rightarrow E\left(\log \frac{q(x)}{p(x)}\right) \leq \log 1 = 0$$

Rewriting the expectation in the inequality as negative KD-divergence

$$\Rightarrow -K(p||q) \leq 0$$

Negating the inequality

$$\Rightarrow 0 \leq K(p||q)$$

Therefore, KD-divergence is non-negative.

Question 1c: KL-Divergence / Information Gain Equivalence

Definition of KL-Divergence with $p = p(x,y)$ and $q = p(x)p(y)$. Call this $A(X,Y)$ for brevity.

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

Applying Bayes' theorem to the above:

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

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

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

Logarithm of division can be expressed as subtraction of logarithms

$$\log\frac{a}{b} = \log a - \log b$$

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

Substituting first term with definition of conditional entropy

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

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

Switching order of summation, and bringing $\log p(y)$ outside the summation over X

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

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

Applying the definition of marginal distribution

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

$$\Rightarrow A(X,Y) = -H(Y|X) - \sum_{y\in Y} \log p(y)\, p(y)$$

Question 1c continued…

Substituting second term with entropy of Y

$$H(Y) = -\sum_{y \in Y} p(y) \log p(y)$$

$$\Rightarrow A(X,Y) = -H(Y|X) + H(Y)$$

$$\Rightarrow A(X,Y) = H(Y) - H(Y|X)$$

The above is equivalent to information gain of Y given X

$$K(p(x,y)||p(x)p(y)) = A(X,Y) = H(Y) - H(Y|X) = I(Y;X)$$

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

Question 2: Benefit of Averaging

Jensen's inequality definition: https://en.wikipedia.org/wiki/Jensen%27s_inequality#Finite_form

Denote the following for brevity (note that t is fixed)

$$y_i = h_i(x)$$

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

$$J(y) = 2L(y,t) = (y-t)^2 = y^2 - 2ty + t^2$$

Jensen's inequality for **convex** function f, values $x_{1\ldots n}$ and weights $a_{1\ldots n}$

$$(f\ convex) \wedge x_{1\ldots n} \subset dom(f) \wedge \forall i \in \{1 \ldots n\}: a_i > 0$$

$$\Rightarrow f\left(\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i}\right) \le \frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i}$$

If all weights $a_{1\ldots n} = 1$, Jensen's inequality for **convex** functions can be applied to averages of **convex** functions

$$(f\ convex) \wedge x_{1\ldots n} \subset dom(f) \Rightarrow f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \le \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Note that $J(y)$ quadratic with respect to y, so it is convex $\forall y \in \mathbb{R}$. Applying Jensen's inequality for convex functions

$$J\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) \le \frac{1}{m}\sum_{i=1}^{m} J(y_i)$$

Substituting definition of $\bar{y}$

$$\Rightarrow J(\bar{y}) \le \frac{1}{m}\sum_{i=1}^{m} J(y_i)$$

Substituting definition of J

$$\Rightarrow 2L(\bar{y},t) \le \frac{2}{m}\sum_{i=1}^{m} L(y_i,t)$$

Recalling definition of $\bar{y}$, y_i and cancelling the 2's

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

Question 3: AdaBoost

Notes:

- Log refers to the natural logarithm, i.e. $\exp(\log x) = 1$
- To avoid confusion with the training data (denoted as t), analysis will be conducted using k as the iteration number.

Denote the following for brevity:

$$c_k = err_k$$

$$d_k = err_k{}'$$

$$x_i = x^{(i)}$$

$$t_i = t^{(i)}$$

$$u_i = w_i{}'$$

$$(x \neq y) = I\{x \neq y\}$$

$$n = N$$

Rewriting the given definitions in terms of the above

$$h_k = \operatorname*{argmin}_{h \in H} \sum_{i=1}^{n} w_i (h(x_i) \neq t_i)$$

$$c_k = \frac{\sum_{i=1}^{n} w_i (h_k(x_i) \neq t_i)}{\sum_{i=1}^{n} w_i}$$

$$\alpha_k = \frac{1}{2} \log\left(\frac{1 - c_k}{c_k}\right)$$

$$u_i = w_i \exp(-\alpha_k t_i h_k(x_i))$$

$$d_k = \frac{\sum_{i=1}^{n} u_i (h_k(x_i) \neq t_i)}{\sum_{i=1}^{n} u_i}$$

Want to Show: $d_k = \frac{1}{2}$

Denote the following sets

$$E_k = \{i : (h_k(x_i) \neq t_i)\}$$

$$G_k = \{i : (h_k(x_i) = t_i)\}$$

Question 3 continued…

Both $h_k(x_i)$ and t_i take values in the domain $\{-1, 1\}$. Therefore, within each set E_k and G_k, the product $h_k(x_i)t_i$ is identical for all $i \in E_k$ or all $i \in G_k$

$i \in E_k \Rightarrow h_k(x_i)t_i = -1$	$i \in G_k \Rightarrow h_k(x_i)t_i = 1$
Proof	
$i \in E_k \Rightarrow (h_k(x_i), t_i) \in \{(-1,1),(1,-1)\}$ $\Rightarrow h_k(x_i)t_i \in \{-1 * 1, 1 * -1\}$ $\Rightarrow h_k(x_i)t_i \in \{-1, -1\} = \{-1\}$ $\Rightarrow h_k(x_i)t_i = -1$	$i \in G_k \Rightarrow (h_k(x_i), t_i) \in \{(-1,-1),(1,1)\}$ $\Rightarrow h_k(x_i)t_i \in \{-1 * -1, 1 * 1\}$ $\Rightarrow h_k(x_i)t_i \in \{1, 1\} = \{1\}$ $\Rightarrow h_k(x_i)t_i = 1$

The above result can be used to create consistent definitions of u_i for $i \in E_k$ and $i \in G_k$

$i \in E_k \Rightarrow u_i = w_i \sqrt{\frac{1 - c_k}{c_k}}$	$i \in G_k \Rightarrow w_i \sqrt{\frac{c_k}{1 - c_k}}$
Proof	
$i \in E_k \Rightarrow h_k(x_i)t_i = -1$ $\Rightarrow u_i = w_i \exp\big(-\alpha_k t_i h_k(x_i)\big) = w_i \exp(\alpha_k)$	$i \in G_k \Rightarrow h_k(x_i)t_i = 1$ $\Rightarrow u_i = w_i \exp\big(-\alpha_k t_i h_k(x_i)\big) = w_i \exp(-\alpha_k)$
$\alpha_k = \frac{1}{2}\log\left(\frac{1 - c_k}{c_k}\right)$ (Recalling definition of α_k)	$-\alpha_k = \frac{1}{2}\log\left(\frac{c_k}{1 - c_k}\right)$ (Negation of logarithm is inversion of operand)
$\Rightarrow u_i = w_i \exp\left(\frac{1}{2} log \frac{1 - c_k}{c_k}\right)$ $\Rightarrow u_i = w_i \sqrt{\exp\left(log \frac{1 - c_k}{c_k}\right)}$ $\Rightarrow u_i = w_i \sqrt{\frac{1 - c_k}{c_k}}$	$\Rightarrow u_i = w_i \exp\left(\frac{1}{2} log \frac{c_k}{1 - c_k}\right)$ $\Rightarrow u_i = w_i \sqrt{\exp\left(log \frac{c_k}{1 - c_k}\right)}$ $\Rightarrow u_i = w_i \sqrt{\frac{c_k}{1 - c_k}}$

Question 3, continued…

Summation over the whole set $\{1 \ldots n\}$ is equal to adding the summations over E_k and G_k

$$\sum_{i=1}^{n} a_i = \sum_{i \in E_k} a_i + \sum_{i \in G_k} a_i$$

Summation over all weights where $h_k(x_i) \neq t_i$ is equal to summation over all weights for which $i \in E_k$

$$\sum_{i=1}^{n} u_i(h_k(x_i) \neq t_i) = \sum_{i \in E_k} u_i$$

Using the previous two results to rewrite d_k

$$d_k = \frac{\sum_{i=1}^{n} u_i(h_k(x_i) \neq t_i)}{\sum_{i=1}^{n} u_i} = \frac{\sum_{i \in E_k} u_i}{\sum_{i \in E_k} u_i + \sum_{i \in G_k} u_i}$$

Substituting the results for u_i

$$\Rightarrow d_k = \frac{\sum_{i \in E_k} w_i \left(\sqrt{\frac{1-c_k}{c_k}}\right)}{\sum_{i \in E_k} w_i \left(\sqrt{\frac{1-c_k}{c_k}}\right) + \sum_{i \in G_k} w_i \left(\sqrt{\frac{c_k}{1-c_k}}\right)}$$

Moving the radicals of c_k outside the summations

$$\Rightarrow d_k = \frac{\left(\sqrt{\frac{1-c_k}{c_k}}\right) \sum_{i \in E_k} w_i}{\left(\sqrt{\frac{1-c_k}{c_k}}\right) \sum_{i \in E_k} w_i + \left(\sqrt{\frac{c_k}{1-c_k}}\right) \sum_{i \in G_k} w_i}$$

Rewriting c_k in terms of E_k and G_k

$$c_k = \frac{\sum_{i \in E_k} w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i \in E_k} w_i}{\sum_{i \in E_k} w_i + \sum_{i \in G_k} w_i}$$

For brevity, let

$$W_E = \sum_{i \in E_k} w_i$$

$$W_G = \sum_{i \in G_k} w_i$$

$$W_n = W_E + W_G = \sum_{i \in E_k} w_i + \sum_{i \in G_k} w_i = \sum_{i=1}^{n} w_i$$

Rewriting the fractions of c_k that appear in d_k using the above

$$c_k = \frac{W_E}{W_E + W_G}$$

$$\Rightarrow 1 - c_k = 1 - \frac{W_E}{W_E + W_G} = \frac{(W_E + W_G) - W_E}{(W_E + W_G)} = \frac{W_G}{W_n}$$

$$\Rightarrow \frac{1 - c_k}{c_k} = \frac{W_G}{W_n} \times \frac{W_n}{W_E} = \frac{W_G}{W_E}$$

$$\Rightarrow d_k = \frac{\sqrt{\frac{W_G}{W_E}} \times W_E}{\sqrt{\frac{W_G}{W_E}} \times W_E + \sqrt{\frac{W_E}{W_G}} \times W_G}$$

$$\Rightarrow d_k = \frac{\sqrt{\frac{W_G W_E^2}{W_E}}}{\sqrt{\frac{W_G W_E^2}{W_E}} + \sqrt{\frac{W_E W_G^2}{W_G}}}$$

$$\Rightarrow d_k = \frac{\sqrt{W_G W_E}}{\sqrt{W_G W_E} + \sqrt{W_E W_G}}$$

$$\Rightarrow d_k = \frac{\sqrt{W_G W_E}}{2\sqrt{W_G W_E}}$$

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

$$\Rightarrow err_k' = \frac{1}{2}$$