CSC411 Fall 2018: Homework 2

Niloufar Afsariardchi

March 9, 2020

1 Problem 1

a) We know $0 \leq p(x) \leq 1$, therefore $1 \leq (1/p(x))$, hence $0 \leq \log_2(1/p(x))$. Since $p(x)$ is non-negative and $\log_2(1/p(x))$ is also non-negative, the multiplication of two term is also non-negative, we therefore have $0 \leq p(x)\log_2(1/p(x))$. Since for each x, the term $p(x)\log_2(1/p(x))$ is non-negative, the summation over all possible x is also non-negative and $0 \leq \sum_x p(x)\log_2(1/p(x))$

b) Let's define $\phi(x) = -\log(x)$, note that since $\log(x)$ is a concave function, $-\log(x)$ is convex. Let's define a random variable $U(X) = \frac{q(X)}{p(X)}$, then from Jensen's inequality we know:

$$\begin{aligned}
\phi(E_X[U(X)]) &\leq E_X[\phi(U(X))] && (1)\\
-\log_2\big(\sum_x p(x)\frac{q(X)}{p(X)}\big) &\leq -\sum_x p(x)\log_2\big(\frac{q(X)}{p(X)}\big) && (2)\\
-\log_2\big(\sum_x q(x)\big) &\leq KL(p||q) && (3)\\
-\log_2 1 &\leq KL(p||q) && (4)\\
0 &\leq KL(p|q) && (5)
\end{aligned}$$

where in equation (1) we used Jensen's inequality for convex function ϕ, equation (2) is from the definition of the expectation for discrete distribution, and in equation (3) we know the sum of the probabilities over all outcomes is 1.

c) From definition we know:

$$\begin{aligned}
& KL(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} && (6)\\
= \;& \sum_{x,y} p(x,y)\log_2 p(x,y) - \sum_{x,y} p(x,y)\log_2 p(x) - && (7)\\
& -\sum_{x,y} p(x,y)\log_2 p(y) && (8)\\
= \;& -H(X,Y) + H(X) + H(Y) && (9)\\
= \;& H(Y) - H(Y|X) && (10)\\
= \;& H(X) - H(X|Y) && (11)\\
= \;& IG(X;Y) && (12)
\end{aligned}$$

where equation (10) is the result of the chain rule discussed in the class.

2 Problem 2

a) Define convex function $\phi(y) = 0.5(y-t)^2$ and random variable $\hat{y} \in \hat{Y} = \{y_1, y_2, .., y_m\}$. From Jansen's inequality we have:

$$\begin{aligned}
\phi(E_{\hat{Y}}[\hat{Y}]) &\leq E_{\hat{Y}}[\phi(U(\hat{y}))] && (13)\\
0.5\big(t - \sum_{y_i \in \hat{Y}} p(y_i) y_i\big)^2 &\leq \sum_{y_i \in \hat{Y}} p(y_i) 0.5(t-y_i)^2 && (14)\\
0.5\big(t - \frac{1}{m}\sum_{y_i \in \hat{Y}} y_i\big)^2 &\leq \frac{1}{m}\sum_{y_i \in \hat{Y}} 0.5(t-y_i)^2 && (15)\\
L(t, h(\bar{x})) &\leq \frac{1}{m}\sum_{i=1}^{m} L(t, h_i(x)) && (16)
\end{aligned}$$

note that I assume uniform distribution for random variable $\hat{Y}$. There for $p(y_i) = 1/m$ in equation (15).

b) Let's define the set $E = \{i : h_t(x^{(i)}) \neq t^{(i)}\}$. Therefore,

$$\begin{aligned}
\text{err}'_t &= \frac{\sum_{i\in E} w'_i}{\sum_{i=1}^{N} w'_i} && (17)\\
&= \frac{\sum_{i\in E} w_i \exp(-\alpha_t h(x^{(i)}) t^{(i)})}{\sum_{i\in E} w_i \exp(-\alpha_t h(x^{(i)}) t^{(i)}) + \sum_{i\in E^c} w_i \exp(-\alpha_t h(x^{(i)}) t^{(i)})} && (18)\\
&= \frac{\sum_{i\in E} w_i \exp(\alpha_t)}{\sum_{i\in E} w_i \exp(\alpha_t) + \sum_{i\in E^c} w_i \exp(-\alpha_t)} && (19)\\
&= \frac{\sum_{i\in E} w_i \exp(0.5 \log \frac{1-\text{err}_t}{\text{err}_t})}{\sum_{i\in E} w_i \exp(0.5 \log \frac{1-\text{err}_t}{\text{err}_t}) + \sum_{i\in E^c} w_i \exp(-0.5 \log \frac{1-\text{err}_t}{\text{err}_t})} && (20)\\
&= \frac{\sum_{i\in E} w_i (\frac{1-\text{err}_t}{\text{err}_t})^{0.5}}{\sum_{i\in E} w_i (\frac{1-\text{err}_t}{\text{err}_t})^{0.5} + \sum_{i\in E^c} w_i (\frac{1-\text{err}_t}{\text{err}_t})^{-0.5}} && (21)\\
&= \frac{\sum_{i\in E} w_i}{\sum_{i\in E} w_i + \sum_{i\in E^c} w_i (\frac{\text{err}_t}{1-\text{err}_t})} && (22)\\
&= \frac{1}{1 + \frac{\sum_{i\in E^c} w_i}{\sum_{i\in E} w_i} (\frac{\text{err}_t}{1-\text{err}_t})} && (23)\\
&= \frac{1}{1 + (\frac{\sum_{i\in E^c} w_i}{\sum_{i\in E} w_i} + 1 - 1)(\frac{\text{err}_t}{1-\text{err}_t})} && (24)\\
&= \frac{1}{1 + (\frac{\sum_{i=1}^{N} w_i}{\sum_{i\in E} w_i} - 1)(\frac{\text{err}_t}{1-\text{err}_t})} && (25)\\
&= \frac{1}{1 + (\frac{1}{\text{err}_t} - 1)(\frac{\text{err}_t}{1-\text{err}_t})} && (26)\\
&= \frac{1}{2} && (27)
\end{aligned}$$

equation (19) is the result of $h(x^{(i)})t^{(i)}) = -1$ for false and $h(x^{(i)})t^{(i)}) = 1$ for correct estimation.

The interpretation of this result is that using new weights the old weak learner is as good as a random guess. Therefore, with the new weights we need to update the weak learner, and we should do a better job at estimation than the previous round so that the error drop below 1/2. This highlights the need to increase the complexity of our estimator at each iteration (e.g., by having a new weak learner added to the ensemble of the weak learners from previous iterations), so that bias decreases over iteration.