CSC411 HW2

Casey Juanxi Li - 998816973

October 2018

1 Information Theory.

a) Prove that entropy is non-negative.

Given

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

We know that $0 \leq p(x) \leq 1$, so $p(x) \geq 0$.

$log_2(p(x))$ for $p(x) \leq 1$ is at most 0, so this term ≤ 0.

So we have:

$$H(X) = -\sum_x (\text{positive or 0 thing}) \cdot (\text{negative or 0 thing}) = -\sum_x (\text{negative or 0 thing}) = (\text{ positive or 0 thing }).$$

b) Prove that KL divergence is non-negative:

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

Since the question hints at us to use Jensen's inequality, we should probably examine the convexity of $f(x) = log_2(x)$:

$$\frac{d^2 f}{dx^2} = \frac{d(\frac{1}{x \ln(2)})}{dx} = \frac{-1}{x^2} \times \frac{1}{\ln(2)}$$

$\frac{1}{\ln(2)}$ is a positive quantity. $\frac{-1}{x^2}$ is a negative quantity for all x. $\frac{d^2 f}{dx^2}$ is therefore ≤ 0 for all x, which also means that $\frac{d^2(-log_2(x))}{dx^2}$ will be ≥ 0 for all x. In other words, $-log_2(x)$ is a convex function, and we can apply Jensen's inequality to it. Let's rewrite D_{KL} using $-log_2(x)$:

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

Assuming p(x) is the pdf for x, this definition is same as:

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

Using Jensen's inequality, we know that:

$$\begin{aligned} D_{KL} = E\left[-log_2(\frac{q(x)}{p(x)}) \right] \geq -log_2(E[\frac{q(x)}{p(x)}]) &= -log_2(\sum_x p(x) \times \frac{q(x)}{p(x)}) \\ &= -log_2(\sum_x q(x)) \\ &= -log_2(1) \qquad \text{(by definition of q(x) being a pdf)} \\ D_{KL} &\geq 0 \end{aligned}$$

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.

$$
\begin{aligned}
KL(p(x,y)||p(x)p(y)) &= H(Y) && - H(Y|X) \\
\sum_x \sum_y p(x,y) log_2(\frac{p(x,y)}{p(x)p(y)}) &= -\sum_y p(y) log_2(p(y)) && - \left(-\sum_x \sum_y p(x,y) log_2(p(y|x)) \right) \\
\sum_x \sum_y p(x,y) log_2(\frac{p(x,y)}{p(x)} \cdot \frac{1}{p(y)}) &= -\sum_y p(y) log_2(p(y)) && + \sum_x \sum_y p(x,y) log_2(p(y|x)) \\
\sum_x \sum_y p(x,y) \left[log_2(p(y|x)) - log_2(p(y)) \right] &= -\sum_y p(y) log_2(p(y)) && + \sum_x \sum_y p(x,y) log_2(p(y|x)) \\
\sum_x \sum_y p(x,y) log_2(p(y|x)) - \sum_y \sum_x p(x,y) log_2(p(y)) &= \sum_x \sum_y p(x,y) log_2(p(y|x)) && - \sum_y p(y) log_2(p(y)) \\
\sum_x \sum_y p(x,y) log_2(p(y|x)) - \sum_y p(y) log_2(p(y)) &= \sum_x \sum_y p(x,y) log_2(p(y|x)) && - \sum_y p(y) log_2(p(y))
\end{aligned}
$$

2 Benefit of averaging.

Again since we get a hint to use Jensen's Inequality, we should probably examine the convexity of the loss function $L(t,y) = \frac{1}{2}(y-t)^2$.

$$
\begin{aligned}
\frac{\partial^2 L}{\partial y^2} &= \frac{\partial (y-t)}{\partial y} = 1 \\
\frac{\partial^2 L}{\partial t^2} &= \frac{\partial (t-y)}{\partial t} = 1
\end{aligned}
$$

Both second partial derivatives are always > 0, so this loss function is convex for all y and t.
Using Jensen's Inequality and $y_i = h_i(x)$, we can then say for all y and t:

$$
\begin{aligned}
& L(E[y], t) \leq E\left(L(y,t)\right) && \\
& L(\bar{h}(x), t) \leq E\left(L(y,t)\right) && \text{(by definition of } \bar{h}(x) \text{ given in the question text)} \\
& L(\bar{h}(x), t) \leq \frac{1}{m} \sum_{i=1}^{m} L(h_i(x), t) && \text{(by definition of expected value)}
\end{aligned}
$$

This last line is the claim given in the question text.

3 AdaBoost.

First let's get rid of the exponential in the weight update. Using change of base we know that in general:

$$
e^{log_{10} x} = e^{\frac{log_e x}{log_e 10}} = e^{log_e x - log_e 10} = \frac{e^{log_e x}}{e^{log_e 10}} = \frac{x}{\text{some constant}}
$$

So we can evaluate the exponential in the expression for w_i'. For math purposes we can disregard that bottom constant since w_i' appears in both the numerator and the denominator sum of err_t'.

$$
\begin{aligned}
w_i' &= w_i e^{-\alpha t^{(i)} h_t(x^{(i)})} \\
&\propto w_i \cdot (\frac{1 - err_t}{err_t})^{-\frac{1}{2} \cdot t^{(i)} h_t(x^{(i)})}
\end{aligned}
$$

$t^{(i)} h_t(x^{(i)}) = 1$ if the classifier was correct ($-1 \cdot -1$, or $1 \cdot 1$), and -1 otherwise.

The numerator of err'_t is the sum of all the updated weights for the predictions which were wrong. In other words, it equals:

$$\sum_{i \in E} w_i \cdot (\frac{1 - err_t}{err_t})^{\frac{1}{2}} \tag{1}$$

The denominator of err'_t, aka $\sum_{i=1}^{N} w'_i$, is equal to:

$$\sum_{i=1}^{N} w'_i = \sum_{i \in E} w_i \cdot (\frac{1 - err_t}{err_t})^{\frac{1}{2}} + \sum_{i \in E^c} w_i \cdot (\frac{1 - err_t}{err_t})^{\frac{-1}{2}} \tag{2}$$

$$= \sum_{i \in E} w_i \cdot (\frac{(1 - err_t)^{1/2}}{(err_t)^{1/2}}) + \sum_{i \in E^c} w_i \cdot (\frac{(err_t)^{1/2}}{(1 - err_t)^{1/2}}) \tag{3}$$

Now taking our handy-dandy hint:

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

Note that also,

$$1 - err = \frac{\sum_{i \in E^c} w_i}{\sum_{i=1}^{N} w_i}$$

Let's make some convenient definitions to make the math look cleaner:

$$\sigma_r \equiv \sum_{i \in E} w_i \quad \text{(r, for right predictions)}$$

$$\sigma_w \equiv \sum_{i \in E^c} w_i \quad \text{(w, for wrong predictions)}$$

$$\sigma_N \equiv \sum_{i=1}^{N} w_i \quad \text{(N, for all predictions)}$$

We can then rewrite err'_t using (1) and (3):

$$err'_t = \frac{\sum_{i \in E} w_i \cdot (\frac{1 - err_t}{err_t})^{\frac{1}{2}}}{\sum_{i \in E} w_i \cdot (\frac{(1 - err_t)^{1/2}}{(err_t)^{1/2}}) + \sum_{i \in E^c} w_i \cdot (\frac{(err_t)^{1/2}}{(1 - err_t)^{1/2}})}$$

Sub in our nice definitions:

$$err'_t = \frac{\sigma_w \cdot \sigma_r^{1/2} \cdot \sigma_w^{-1/2} \cdot \sigma_N^{1/2} \cdot \sigma_N^{-1/2}}{\sigma_w \cdot \sigma_r^{1/2} \cdot \sigma_w^{-1/2} \cdot \sigma_N^{1/2} \cdot \sigma_N^{-1/2} + \sigma_r \cdot \sigma_w^{1/2} \cdot \sigma_r^{-1/2} \cdot \sigma_N^{1/2} \cdot \sigma_N^{-1/2}}$$

The $\sigma_N^{1/2} \cdot \sigma_N^{-1/2}$'s are just 1:

$$\begin{aligned} err'_t &= \frac{\sigma_w \cdot \sigma_r^{1/2} \cdot \sigma_w^{-1/2}}{\sigma_w \cdot \sigma_r^{1/2} \cdot \sigma_w^{-1/2} + \sigma_r \cdot \sigma_w^{1/2} \cdot \sigma_r^{-1/2}} \\ &= \frac{\sigma_w^{1/2} \cdot \sigma_r^{1/2}}{\sigma_w^{1/2} \cdot \sigma_r^{1/2} + \sigma_r^{1/2} \cdot \sigma_w^{1/2}} \\ &= \frac{1}{1+1} = \frac{1}{2} \end{aligned}$$

Intuitive interpretation:

w_i is the weight, or importance, of each data point from $i = 1$ *to* N when we are calculating err_t. When we increase a certain w_i, we are saying, "for the purpose of evaluating how well this estimator is doing, we want to penalize an error on item i more than we were penalizing it before".

The given definition of err_t' updates the weight of each item, **while keeping the old estimator** h_t, such that the weighted error becomes 50%. In other words, if we kept our current classifier, but apply our new "importances", this classifier would now be doing no better than random chance. This necessarily means boosting the weights of items we got wrong because...

AdaBoost requires a weak learner h which does slightly better than chance, and we can see why this is the case. "Doing slightly better than chance" means that on the first iteration, if we start with equal weights, err should be "small", roughly speaking, in the sense that it is < 0.5.

The weight update for a "wrong" item is then, intuitively speaking:

$$w'_{i,wrong} = w_i \cdot \sqrt{\frac{1 - err}{err}} = w_i \cdot \sqrt{\frac{1 - \text{small thing}}{\text{small thing}}} = w_i \cdot \sqrt{\text{something} > 1} = w_i \cdot (something > 1)$$

In other words, w_i' necessarily grows for an item that was misclassified, while the opposite is true for an item that we got correct:

$$w'_{i,right} = w_i \cdot \sqrt{\frac{err}{1 - err}} = w_i \cdot \sqrt{\frac{\text{small thing}}{\text{1 - small thing}}} = w_i \cdot \sqrt{\text{something} < 1} = w_i \cdot (something < 1)$$

Going back to the "$\frac{1}{2}$" result, it's now evident that the only way to go from having an error rate < 0.5 (because we are using a weak learner) to an error of 0.5 after the weight update is to increase the weights of the mistakes, while decreasing the weights of the correct classifications.