CSC 411 H1 FALL 2018

LINO LASTELLA 1001237654

1

a. Let X be a discrete random variable with probability mass function p.

Then, by definition, $H(X) = -\sum_x p(x) \log_2(p(x))$, where $x \in \mathcal{X}$, for some finite $\mathcal{X}$.

Also by definition, $0 < p(x) \leq 1$. We exclude the single value $p(x) = 0$ since $\log_2(0)$ is not defined.

Then, $H(x) = -\sum_x p(x) \log_2(p(x))$ is a

non-negative sum, since $p(x)$ is non-negative and the function $\log_2$ outputs values less than or equal to zero for any input between 0 and 1, 0 not included (the two minus signs cancel).

b. Let X be a discrete random variable with probability mass functions p and q, and expectation p.

Then, by definition,

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

Now notice that since $\log_2(x)$ is concave for any $x > 0$, $-\log_2(x)$ is convex.

Therefore, by Jensen's inequality,

$$KL(p\|q) = \sum_x \left[p(x) \cdot \left(-\log\left(\frac{q(x)}{p(x)} \right) \right) \right]$$

$$= E\left[-\log(X') \right], \text{ for some r.v. } X'(X) = \frac{q(x)}{p(x)}$$

$$\geq -\log(E[X'])$$

$$= -\log\left(\sum_x \cancel{p(x)} \cdot \frac{q(x)}{\cancel{p(x)}} \right)$$

$$= -\log\left(\sum_x q(x) \right) = -\log(1) = 0$$

because $q(x)$ is still a valid pmf

This means that $KL(p\|q)$ is non-negative.

Note: I assumed that both distributions p and q are ≥ 0 and that $KL(p\|q) = 0$ whenever $p = q$, which includes the case $p = q = 0$.

New Note: Instructor said to ignore the case $p = q = 0$.

c. Let X, Y be two discrete random variables, where $p(x) = \sum_y p(x,y)$ is the marginal distribution of X.

Define $I(Y;X)$ to be $H(Y) - H(Y|X)$.

Then,

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

$$= -\sum_y p(y)\log_2(p(y)) - \sum_x p(x) H(Y|X=x)$$

from marginal dis. $$= -\sum_y \left(\log_2(p(y)) \sum_x p(x,y)\right) - \sum_x p(x) H(Y|X=x)$$

from $H(Y|X=x)$ formula $$= -\sum_{x,y} \log_2(p(y)) \cdot p(x,y) - \sum_x p(x)\left(-\sum_y p(y|x) \cdot \log_2(p(y|x))\right)$$

$$= -\sum_{x,y} p(x,y)\log_2(p(y)) + \sum_{x,y} p(x)p(y|x)\log_2(p(y|x))$$

$$= \sum_{x,y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)}\right) - \sum_{x,y} p(x,y)\log_2(p(y))$$

$$= \sum_{x,y} p(x,y) \cdot \left(\log_2\left(\frac{p(x,y)}{p(x)}\right) + \log_2\left(\frac{1}{p(y)}\right)\right)$$

$$= \sum_{x,y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

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

② Let $L(y,t) = \frac{1}{2}(y-t)^2$. Let $\bar{h}(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(x)$.

Claim: L is a convex function.

proof:

Let L be a function as above.

then,

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial y} \\ \frac{\partial L}{\partial t} \end{pmatrix} = \begin{pmatrix} y-t \\ t-y \end{pmatrix}.$$

$$\nabla^2 L = \begin{pmatrix} \frac{\partial^2 L}{\partial y \partial y} & \frac{\partial^2 L}{\partial y \partial t} \\ \frac{\partial^2 L}{\partial t \partial y} & \frac{\partial^2 L}{\partial t \partial t} \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Let $v \in \mathbb{R}^2$.

Then $v^T \nabla^2 L\, v = (v_1, v_2)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$

$$= v_1^2 - v_1 v_2 - v_2 v_1 + v_2^2 = (v_1 - v_2)^2 \geq 0$$

It then follows that the Hessian matrix of L is positive-semidefinite, which implies L is convex.

We can now apply Jensen's Inequality to the estimators $h_1, \ldots, h_m$.

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

$$= E[L(h_i(x), t)]$$

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

(Notice: in this case outputs of each estimator are all equally likely, so average and expected value are equivalent)

$\square$

③

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

Let E, E^c be as described in the handout.

Then, $err'_t = \dfrac{\sum_{i\in E} w_i' \cdot 1 + \sum_{i \in E^c} w_i' \cdot 0}{\sum_{i\in E} w_i' + \sum_{i\in E^c} w_i'} =$

$$= \frac{\sum_{i\in E} w_i \cdot \exp(2\alpha I\{h_t(x^{(i)}) \neq t^{(i)}\})}{\sum_{i\in E} w_i \cdot \exp(2\alpha_t I\{h_t(x^{(i)}) \neq t^{(i)}\}) + \sum_{i\in E^c} w_i \cdot \exp(2\alpha_t I\{h_t(x^{(i)}) \neq t^{(i)}\})}$$

$$= \frac{\sum_{i\in E} w_i e^{2\alpha_t}}{\sum_{i\in E} w_i \cdot e^{2\alpha_t} + \sum_{i\in E^c} w_i} =$$

$$= \frac{\sum_{i\in E} w_i \cdot e^{\log \frac{1-err_t}{err_t}}}{\sum_{i\in E} w_i \cdot e^{\log\frac{1-err_t}{err_t}} + \sum_{i\in E^c} w_i} =$$ Since $2\alpha_t = \log \frac{1-err_t}{err_t}$

$$= \frac{\frac{1-err_t}{err_t} \cdot \sum_{i\in E} w_i}{\frac{1-err_t}{err_t} \cdot \sum_{i\in E} w_i + \sum_{i\in E^c} w_i}$$ Since $\frac{1-err_t}{err_t}$ does not depend on summation

= (Next Page)

Notice that any fraction of the form $\frac{x}{x+y}$ can be rearranged to be $\frac{x}{x+y} = \frac{x}{x(1+\frac{y}{x})} = \frac{1}{1+\frac{y}{x}}$.

So in our case, to show that $err'_t = \frac{1}{2}$ it is sufficient to show that $\frac{\sum_{i\in E^c} w_i}{\frac{1-err_t}{err_t} \cdot \sum_{i\in E} w_i} = 1$.

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

$$= \frac{\frac{1}{err_t} - 1}{\frac{1-err_t}{err_t}} \qquad \text{Since} \left(\frac{\sum_{i\in E} w_i}{\sum_{i\in N} w_i}\right)^{-1} = \left(err_t\right)^{-1}$$

$$= \frac{1}{\cancel{err_t}} \cdot \frac{\cancel{err_t}}{1-err_t} - \frac{err_t}{1-err_t} = \frac{1-err_t}{1-err_t} = 1.$$

$$\text{I.E. } err'_t = \frac{1}{1+\frac{\sum_{i\in E^c} w_i}{\frac{1-err_t}{err_t}\sum_{i\in E} w_i}} = \frac{1}{1+1} = 1/2.$$

The interpretation of this result is that on any new iteration, the error w.r.t. the new weights is constant, which means that a large number of classifiers will <u>not</u> cause overfitting, only improve.