HW#5

Problem 2

Part 1 $H(p) = -q \log_2 q - (1-q)\log_2(1-q)$

$P(X=1) = q$

$P(X=0) = p \qquad p+q=1$

$H(s)$ of the set $H(\frac{p}{p+n})$

$H(s) \le 1 \qquad H(s) = 1$ when $p=n$

$\frac{\partial H}{\partial q} = \frac{\partial}{\partial q}\left(-q\log_2 q - (1-q)\log_2(1-q)\right)$

$\frac{\partial H}{\partial q} = -\log q + \cancel{1} + \log(1-q) - 1 = 0$

$-\log q + \log(1-q) = 0$

$\log_2(1-q) = \log_2 q$

$1-q = q$

$1 = 2q$

$q = \frac{1}{2}$ at the optimum

$q = \frac{p}{p+n} = \frac{1}{2}$ when $p=n$

$H(\frac{1}{2}) = -\frac{1}{2}\log_2\frac{1}{2} - (1-\frac{1}{2})\log_2(1-\frac{1}{2})$

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

$= 1$

At both sides above + below

$\frac{p}{p+n} = \frac{1}{4}$ and $\frac{3}{4}$ ($p=0.25$ and $p=0.75$)

$p = 3n$ and $3p = n$

$H(\frac{1}{4}) = -\frac{1}{4}\log_2\frac{1}{4} - (1-\frac{1}{4})\log_2(1-\frac{1}{4})$

$= -\frac{1}{4}(-2) - \frac{3}{4}(-0.415)$

$= 0.81$

$H(\frac{3}{4}) = -\frac{3}{4}\log_2\frac{3}{4} - (1-\frac{3}{4})\log_2(1-\frac{3}{4})$

$= -\frac{3}{4}(-0.415) - \frac{1}{4}(-2)$

$= 0.81$

$\boxed{\text{Both sides decrease, thus } q=\tfrac{1}{2} \text{ is max, } H(s) \le 1 \text{ and } H(s)=1 \text{ when } p=n}$

Part 2 D: 800 pts

C_1: 400 pts

C_2: 400 pts

Split A (300, 100) (100, 300)

Split B (200, 400) (200, 0)

Reduction in cost on (j,t) = $cost(D) - \left[\frac{|D_{left}|}{|D|} cost(D_{left}) + \frac{|D_{right}|}{|D|} cost(D_{right})\right]$

Misclassification Rate: $cost(D) = \frac{1}{|D|}\sum_{(x,y)\in D} I(y \ne \hat{y})$ $\quad \hat{y}$ = majority label in D

Entropy: $cost(D) = -p\log_2 p - (1-p)\log_2(1-p)$ $\quad$ P = fraction of positive examples in D

Gini Index: $cost(D) = 2p(1-p)$

Misclassification Rate

A: $cost(D) = \frac{1}{800}\sum_{(x,y)} I(y \ne \hat{y}) = 0.5$

$cost(D_{left}) = \frac{1}{400}\sum_{(x,y)} I(y \ne \hat{y}) = 0.25$

$cost(D_{right}) = \frac{1}{400}\sum_{(x,y)} I(y \ne \hat{y}) = 0.25$

reduction = $cost(D) - \left[\frac{400}{800}(0.25) + \frac{400}{800}(0.25)\right]$

$= 0.5 - 0.25 = 0.25$

B: $cost(D_{left}) = \frac{1}{600}\sum_{(x,y)} I(y \ne \hat{y}) = 0.33$

$cost(D_{right}) = \frac{1}{200}\sum_{(x,y)} I(y \ne \hat{y}) = 0$

reduction = $cost(D) - \left[\frac{600}{800}(0.33) + 0\right]$

$= 0.5 - 0.2475 = 0.2525$

entropy

$cost(D) = -p\log_2 p - (1-p)\log_2(1-p)$

$= -0.5\log_2(\frac{1}{2}) - 0.5\log_2(\frac{1}{2})$

$= -0.5(-1) - 0.5(-1)$

$= 1$

A. $cost(D_{left}) = -(\frac{3}{4})\log_2(\frac{3}{4}) - (\frac{1}{4})\log_2(\frac{1}{4}) = 0.811$

$cost(D_{right}) = -(\frac{3}{4})\log_2(\frac{3}{4}) - (\frac{1}{4})\log_2(\frac{1}{4}) = 0.811$

reduction $= 1 - [\frac{400}{800}(0.811) + \frac{400}{800}(0.811)] = 0.19$

B. $cost(D_{left}) = -\frac{4}{6}\log_2\frac{4}{6} - (\frac{2}{6})\log_2(\frac{2}{6}) = 0.918$

$cost(D_{right}) = -1\log_2 1 - (0)\log_2(0) = 0$

reduction $= 1 - [\frac{600}{800}(0.918)] = 0.3115$

Gini Index

$cost(D) = 2p(1-p) = 2(0.5)(0.5) = 0.5$

A. $cost(D_{left}) = 2(\frac{3}{4})(\frac{3}{4}) = 0.375$

$cost(D_{right}) = 2(\frac{3}{4})(\frac{3}{4}) = 0.375$ reduction $= 0.5 - (0.5(0.375) + 0.5(0.375))$

$= 0.125$

B. $cost(D_{left}) = 2(\frac{2}{3})(\frac{1}{3}) = \frac{4}{9}$

$cost(D_{right}) = 2(1)(0) = 0$ reduction $= 0.5 - (\frac{6}{8}(\frac{4}{9}))$

$= 0.17$

Split B is the superior split (preffered) according to every error measure as its reduction in cost is the larger value for every error measure compared to split A

Part 3

Misclassification rate can never ~~never~~ increase when splitting (cost will stay the same at worst)

$mr = \frac{1}{|D|}\sum_{(x,y)\in D} I(y \neq \hat{y})$ $\hat{y}$ = majority label

Highest $MR = \frac{1}{2}$ ($\frac{1}{2}, \frac{1}{2}$ split between classes), if ↑ split for any class, majority label changes, decreasing MR

Take example dataset with 500 pts in c_1, 100 pts in c_2

split into (100, 100), (400, 0) $MR = \frac{1}{600}(100) = \frac{1}{6}$

↑ highest error

$\frac{1}{200}(100) = 0.5$; 0 ⇒ cost reduction $= \frac{1}{6} - \frac{1}{3}(0.5) = 0$

↑ the MR for this part of the split is higher but overall in the dataset, the amount of points that are misclassified can only go down or stay the same when splitting on a feature and thus the MR can only go down when considering the entire dataset D. A portion of the dataset (1 side of the split) can have higher MR but when put back in the context of all data points, the MR can only stay the same at worst.

problem 3.

part 1

$$E_{bag} = E_x[\epsilon_{bag}(x)^2] = E_x[\{\frac{1}{L}\sum_{l=1}^{L} h_l(x) - f(x)\}^2]$$

$$E_{av} = \frac{1}{L}\sum_{l=1}^{L} E(\epsilon_l(x)^2] = \frac{1}{L}\sum_{l=1}^{L} E_x[\epsilon_l(x)^2]$$

$$E_{bag} = E_x[\{\frac{1}{L}\sum_{l=1}^{L} f(x) + \epsilon_l(x) - f(x)\}^2]$$

$$= E_x[\{\frac{1}{L}\sum_{l=1}^{L} \epsilon_l(x)\}^2]$$

$$= (\frac{1}{L})^{2}\sum_{l=1}^{L} E_x[\epsilon_l(x)^2]$$

$$E_{bag} = (\frac{1}{L})\sum_{l=1}^{L} E_x[\epsilon_l(x)^2]$$