سینا نمازی
شماره دانشجویی: ۴۰۰۱۱۰۱۵۴

۱-

الف)

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

ب)

$$P(x_1=T)=\frac{6}{10},\quad P(x_1=I)=\frac{4}{10},\quad P(y=w\mid x_1=T)=\frac{1}{3},\quad P(y=w\mid x_1=I)=\frac{3}{4}$$

$$IG(y,x_1) = 1-\left(\frac{6}{10}\times -\left(\frac{1}{3}\log\frac{1}{3}+\frac{2}{3}\log\frac{2}{3}\right)+\frac{4}{10}\times -\left(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4}\right)\right) \simeq 0,12$$

$$P(x_2=M)=\frac{4}{10},\quad P(x_2=P)=\frac{6}{10},\quad P(y=w\mid x_2=M)=\frac{3}{4},\quad P(y=w\mid x_2=P)=\frac{1}{3}$$

$$IG(y,x_2) = 1-\left(\frac{6}{10}\times -\left(\frac{1}{3}\log\frac{1}{3}+\frac{2}{3}\log\frac{2}{3}\right)+\frac{4}{10}\times -\left(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4}\right)\right) \simeq 0,12$$

$$P(x_3=S)=\frac{5}{10},\quad P(x_3=C)=\frac{5}{10},\quad P(y=w\mid x_3=S)=\frac{4}{5},\quad P(y=w\mid x_3=C)=\frac{1}{5}$$

$$IG(y,x_3) = 1-\left(\frac{5}{10}\times -\left(\frac{4}{5}\log\frac{4}{5}+\frac{1}{5}\log\frac{1}{5}\right)+\frac{5}{10}\times -\left(\frac{1}{5}\log\frac{1}{5}+\frac{4}{5}\log\frac{4}{5}\right)\right) \simeq 0,28$$

این معیار باید ویژگی x_3 انتخاب شود چون Information Gain بیشتری دارد.

(الف 1

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(M_T(x_i) - y_i)^2 \xrightarrow{M_T(x_i)=\sum_{t=1}^{T}\beta_t G_t(x_i)} MSE = \frac{1}{n}\sum_{i=1}^{n}(\sum_{t=1}^{T}\beta_t G_t(x_i) - y_i)^2$$

(ب

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\beta_T G_T(x_i) + \sum_{t=1}^{T-1}\beta_t G_t(x_i) - y_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_T G_T(x_i) + M_{T-1}(x_i) - y_i)^2$$

$$\frac{\partial MSE}{\partial \beta_T} = \frac{1}{n}\sum_{i=1}^{n} 2(\beta_T G_T(x_i) + M_{T-1}(x_i) - y_i) \times G_T(x_i) = 0$$

$$\Rightarrow \beta_T \sum_{i=1}^{n} G_T^2(x_i) = \sum_{i=1}^{n}(y_i G_T(x_i) - G_T(x_i) M_{T-1}(x_i))$$

$$\Rightarrow \beta_T = \frac{\sum_{i=1}^{n}(y_i G_T(x_i) - G_T(x_i) M_{T-1}(x_i))}{\sum_{i=1}^{n} G_T^2(x_i)} = \frac{\sum_{i=1}^{n} G_T(x_i)(y_i - M_{T-1}(x_i))}{\sum_{i=1}^{n} G_T^2(x_i)}$$

_3

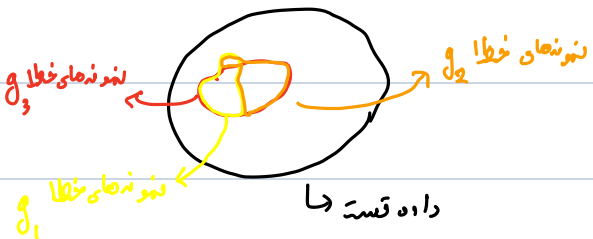
الف)

احتمال انتخاب نشدن یک نمونه خاص $= (1-\frac{1}{N})^{PN}$

چون N داده داریم $\Rightarrow N(1-\frac{1}{N})^{PN} \xrightarrow{\text{چون N خیلی بزرگ است}} N \times \lim_{N\to\infty}(1-\frac{1}{N})^{PN} = N \times e^{-P}$

ب) ابتدا lower bound را بدست می‌آوریم. برای این‌کار می‌توان در نظر گرفت هرکدام از درخت های تصمیم ما روی داده‌های متفاوتی اشتباه دسته بندی می‌کنند، به عبارت دیگر هرگاه نمونه‌ای توسط یکی از درخت ها اشتباه دسته بندی می‌شود، توسط دو درخت دیگر به درستی دسته بندی می‌شود. چون جمع خطاهای 3 درخت از 1 کمتر است $(0.15+0.25+0.35=0.75)$، این حالت ممکن است و در این حالت خطای مدل کلی ما یعنی G برروی داده تست 0 می‌شود.

حال upper bound را بدست می‌آوریم. برای این‌کار باید تعداد و بخشی از داده تست که حداقل توسط 2 تا از درخت ها به اشتباه دسته بندی می‌شوند را پیشنهاد کنیم. برای این‌کار به درخت نگاه می‌کنیم که بیشترین خطا را دارد یعنی g_3 که این میزان 0.35 است حال این 0.35 را به دو قسمت با اندازه‌های 1.25, 2.25 تقسیم می‌کنیم. حال اگر فرض کنیم تمام قسمت اول توسط g_2 نیز اشتباه دسته بندی شوند و تمام قسمت دوم نیز توسط g_1 اشتباه دسته بندی شوند، همچنین 0.25 از هرکدام از g_1 و g_2 باقی‌مانده این بخش از این دو نیز با یکدیگر overlap دارند. مدل G ما هم این 0.375 از داده تست را اشتباه دسته بندی می‌کند بنابراین در این حالت خطای کل 0.375 خواهد بود.



بنابراین محدوده خطای G: $0 \leq E_{out}(G) \leq 0.375$

الف) درست است. در Bagging ، ما داده های آموزشی جدید را از داده های آموزشی اصلی sample می‌کنیم، بنابراین بعد از این مرحله sample کردن و بدست آوردن چند دسته داده آموزشی sample شده، یادگیرنده ها را می‌توان به شکل موازی بر روی هرکدام از این دسته‌ها آموزش داد.

ب) نادرست است. در روش boosting ، یادگیرنده‌ها از هم مستقل نیستند و آموزش هر یادگیرنده به یادگیرنده های قبلی بستگی دارد، بنابراین نمی‌توان آن ها را به شکل موازی آموزش داد.

ج) نادرست است. در روش bagging ، ما برای هر یادگیرنده داده های آموزشی را از داده‌های آموزشی اصلی sample می‌کنیم (با replacement). بنابراین ممکن است از کل داده آموزشی اصلی استفاده نکنیم.

د) درست است. در روش boosting ، هر یادگیرنده پس از قبلی، بر روی کل داده های آموزشی، آموزش داده می‌شود.

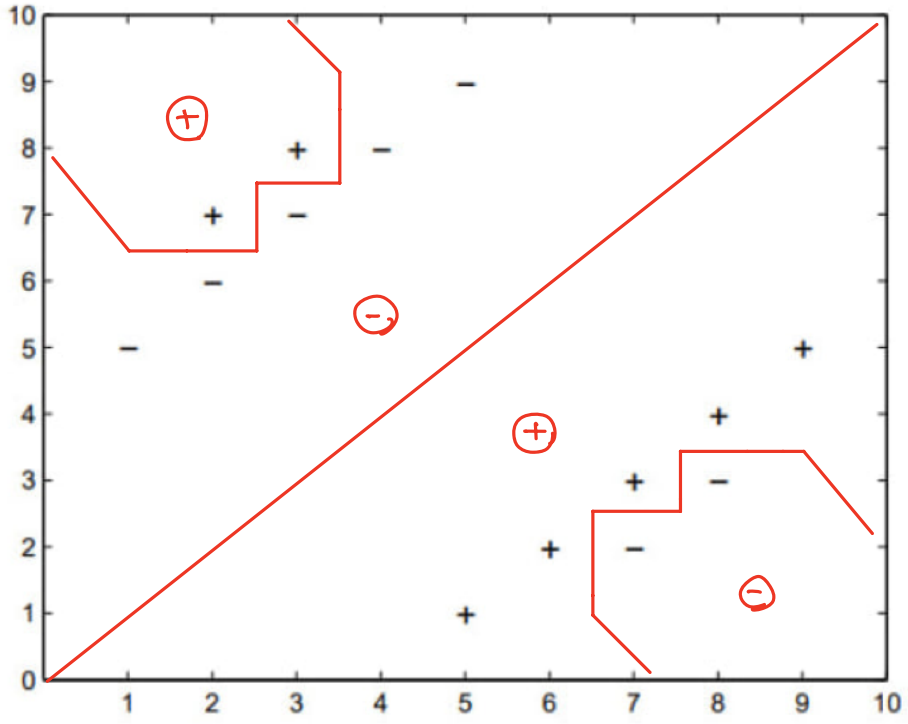
5-

الف) چون هر کدام از این نمونه‌ها در داده آموزش نیز وجود دارند و می‌توانند نزدیک‌ترین همسایه خود باشند، اگر $k=1$ باشد، error کمینه شده و برابر با صفر می‌شود.

ب) اگر k خیلی بزرگ باشد، Resolution مدل ما کم شده و همانطور که مشخص است، اگر $k=13$ باشد، همه نمونه‌ها اشتباه دسته‌بندی می‌شوند (با استفاده از روش leave one out cross validation). اما اگر k خیلی کوچک باشد منجر به overfit شدن مدل ما می‌شود.

ج) اگر مقدار k را برابر با 5 یا 7 قرار دهیم میزان error مدل کمینه می‌شود و مقدار error نیز برابر با $\frac{4}{14}$ می‌شود.

د)



$$H(T) = -\sum_{i=1}^{k} P(c_i)\log P(c_i) = -\sum_{i=1}^{k} \frac{1}{k}\log\frac{1}{k} = \frac{1}{k}\sum_{i=1}^{k}\log k = \log k$$

بدون ویژگی A و T را به m_A دسته disjoint تقسیم می‌کند و توزیع کلاس‌ها در هر دسته به شکل Uniform است:

$$H(T_i^A) = H(T) = \log k \quad , \quad P(A^i) = \frac{|T_i^A|}{|T|}$$

$$\Rightarrow H(T|A) = \sum_{i=1}^{m_A} P(A^i)\, H(T_i^A) = \log k \sum_{i=1}^{m_A} P(A^i) = \log k$$

($H(T_i^A) = \log k$, $\sum_{i=1}^{m_A} P(A^i) = 1$)

$$IG(T, A) = H(T) - H(T|A) = \log k - \log k = 0$$

به عبارت دیگر اگر ویژگی A را بدانیم، اطلاعاتی درباره‌ی کلاس نمونه‌های T نمی‌دهد و عدم قطعیت را کم نمی‌کند.

یا می‌توان گفت کلاس نمونه‌های موجود در T مستقل از ویژگی A هستند.