سوال اول

۱ – با توجه به اینکه تصمیم نهایی با Majority Vote انجام می‌گیرد، بنابراین حاصل جمع از مقدار $m = \frac{N+1}{2} = 3$ شروع می‌شود و داریم:

$$\mu = \sum_{i=3}^{N=5} \binom{N}{i} (0 \cdot 7)^i (1 - 0 \cdot 7)^{N-i} = 0 \cdot 837$$

۲ – مانند قبل $m = \frac{N+1}{2} = 5$ و داریم:

$$\mu = \sum_{i=5}^{N=9} \binom{N}{i} (0 \cdot 7)^i (1 - 0 \cdot 7)^{N-i} = 0 \cdot 9012$$

۳ – عبارت نهایی برای $q = 0 \cdot 7$ و $N \to \infty$ را می توان به صورت زیر بازنویسی کرد:

$$\mu = \lim_{N\to\infty} \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} q^i (1-q)^{N-i}$$

برای به دست آوردن مقدار عبارت بالا از بسط دو جمله‌ای استفاده می‌نماییم:

$$\begin{aligned}
1 = \lim_{N\to\infty} (q + 1 - q)^N &= \lim_{N\to\infty} \sum_{i=0}^{N} \binom{N}{i} q^i (1-q)^{N-i} \\
&= \lim_{N\to\infty} \left[\sum_{i=0}^{\lfloor \frac{N-1}{2} \rfloor} \binom{N}{i} q^i (1-q)^{N-i} + \sum_{i=\lfloor \frac{N+1}{2} \rfloor}^{N} \binom{N}{i} q^i (1-q)^{N-i} \right] \\
&= \lim_{N\to\infty} \left[\sum_{i=\lfloor \frac{N+1}{2} \rfloor}^{N} \binom{N}{i} (1-q)^i q^{N-i} + \sum_{i=\lfloor \frac{N+1}{2} \rfloor}^{N} \binom{N}{i} q^i (1-q)^{N-i} \right] \\
&= \lim_{N\to\infty} \left[q^N \sum_{i=\lfloor \frac{N+1}{2} \rfloor}^{N} \binom{N}{i} (\frac{1-q}{q})^i + \sum_{i=\lfloor \frac{N+1}{2} \rfloor}^{N} \binom{N}{i} q^i (1-q)^{N-i} \right] \\
&= \lim_{N\to\infty} \left[q^N \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} (\frac{1-q}{q})^i \right] + \mu
\end{aligned}$$

در عبارت نهایی به دست آمده با توجه به اینکه $q = 0 \cdot 7$ و $N \to \infty$، هر دو عبارت q^N و $\sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} (\frac{1-q}{q})^i$ به صفر میل می‌کنند ($\frac{1-q}{q} \leq 1$ و $q \leq 1$) و بنابراین داریم: $\mu = 1$. در عمل نیز

به این مقدار دقت نمی توان رسید، زیرا که با افزایش تعداد طبقه‌بندها احتمال آن هریک مستقل از دیگری عمل نماید، کاهش می‌یابد.

۴ – مانند قسمت اول، با توجه به اینکه تصمیم نهایی با Majority Vote انجام می‌گیرد، بنابراین حاصل جمع از مقدار $m = \frac{N+1}{2} = 3$ شروع می‌شود و داریم:

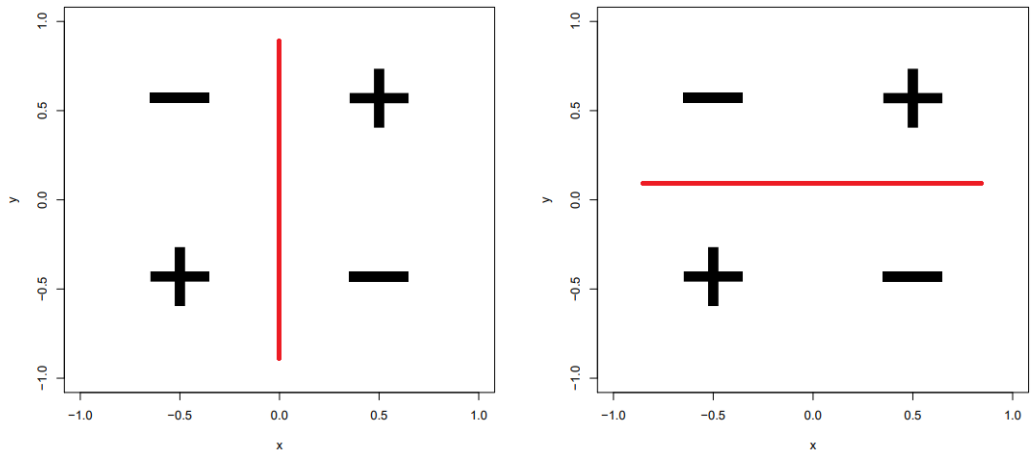
$$\mu = \sum_{i=3}^{N=5} \binom{N}{i} (0 \cdot 5)^i (1 - 0 \cdot 5)^{N-i} = 0 \cdot 5$$

درواقع ترکیب طبقه‌بندها با دقت 50% تاثیری بر مقدار دقت طبقه‌بند نهایی ندارد. ترکیب طبقه‌بندها زمانی مفید خواهد بود که حداقل دقت یکی از آنها بیشتر از $0 \cdot 5$ باشد.

سوال دوم

الف)

۱ - از آنجایی که هر تقسیم از مجموعه دیتا داده شده، خطای ۵۰٪ ایجاد می کند، Adaboost هرگز به دقت طبقه بندی بهتر از ۵۰ درصد در این مجموعه داده نخواهد رسید.



۲ – حد بالای خطای یادگیری طبقه‌بند نهایی، H، به صورت رابطه‌ی زیر به دست می‌آید:

$$\frac{1}{m}\sum_{i=1}^{m}\delta(H(x_i)\neq y_i)\leq \exp(-2\sum_{t=1}^{T}(0\cdot 5-\epsilon_t)^2)$$

بنابراین حد بالای عبارت سمت است برابر $\frac{1}{m}$ خواهد بود. حال باید مقدار T را به گونه‌ای انتخاب کنیم که:

$$\frac{1}{m}\geq \exp(-2\sum_{t=1}^{T}(0\cdot 5-\epsilon_t)^2)$$

از آنجایی که کار کردن با عبارت سمت راست رابطه‌ی بالا دشوار بوده و در صورت سوال ذکر شده که $\epsilon_t \leq \gamma_t$ است، می توان آن را به صورت زیر بازنویس کرد و در عوض مقدار T را به گونه‌ای به دست می‌آوریم که:

$$\frac{1}{m}\geq \exp(-2T(0\cdot 5-\gamma_t)^2)$$

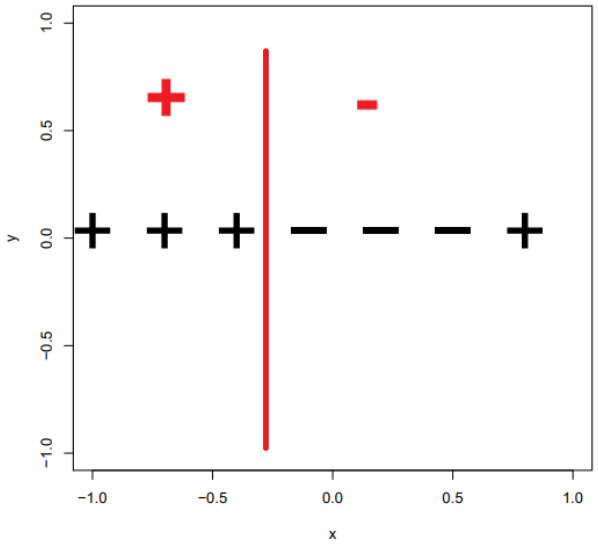
(به عبارت دیگر نامساوی $\frac{1}{m}\geq \exp(-2T(0\cdot 5-\gamma_t)^2)\geq \exp(-2\sum_{t=1}^{T}(0\cdot 5-\epsilon_t)^2)$ برقرار است.)

و با حل آن داریم:

$$\frac{1}{m} \geq \exp(-2T(0\cdot 5-\gamma_t)^2) \Rightarrow \ln(m) \leq 2T(0\cdot 5-\gamma_t)^2 \Longrightarrow \frac{\ln(m)}{2(0\cdot 5-\gamma_t)^2} \leq T$$

ب)

۳ – مرز تصمیم بین ۳ نقطه + و ۳ نقطه – قرار گرفته و کلاس + در سمت چپ آن واقع شده است.



۴ – $\epsilon_1 = \frac{1}{7}$ و $\alpha_1 = \frac{1}{2}\ln\frac{6/7}{1/7} = \frac{1}{2}\ln(6) = 0\cdot 8959$

دقت طبقه‌بند برابر با $\frac{6}{7}$ خواهد بود.

۵ –

$$\omega_i^{(1)} = e^{-y_i F_i(x_i)} = e^{-y_i \alpha_1 h_1(x_i)}$$

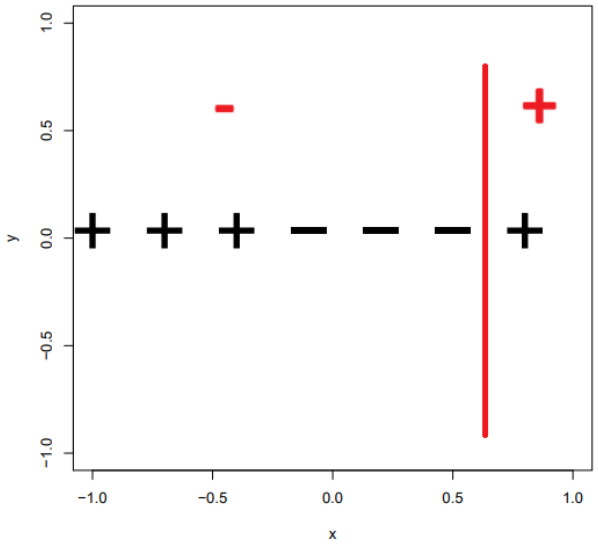
$$\omega_i^{(1)} = \begin{cases} e^{-\alpha_1} = e^{-0\cdot 89} & i = 1\cdots 6 \\ e^{\alpha_1} = e^{0\cdot 89} & i = 7 \end{cases}$$

$$Z_1 = \frac{6}{7}exp(-0\cdot 8959) + \frac{1}{7}\exp(0\cdot 8959) = 0\cdot 6998542$$

$$D_2(i) = \frac{(1/7)\exp(-0\cdot 8959)}{Z_1} = 0\cdot 083333 \;\forall i = 1\cdots 6$$

$$D_2(7) = \frac{(1/7)\exp(0\cdot 8959)}{Z_1} = 0\cdot 5$$

۶ - مرز تصمیم بین تک نقطه + در سمت راست و ۳ نقطه – است که کلاس + در سمت راست آن واقع شده است.



۷ – پس از تکرار دوم، ۳ نقطه – کمترین وزن را خواهند داشت، زیرا توسط هر دو طبقه‌بند h_1 و h_2 به درستی طبقه‌بندی شده‌اند.

۸ - خیر، دقت طبقه‌بندی بهبود نمی‌یابد. زیرا، تمام نقاط پس از تکرارهای اول و دوم به طور یکسان طبقه بندی شده‌اند.

سوال پنجم

در این سوال با استفاده از معیار بهره اطلاعات درخت تصمیم را تشکیل می دهیم و سپس برای داده های آزمون با توجه به درخت در دست خروجی را پیش بینی میکنیم.

الف)

در جدول ۹ نفر دچار انسداد شریان هستند و ۵ نفر دچار این عارضه نیستند .

برای رابطه انتروپی داریم:

$$E(s) = -P_+ \log_2 P_+ - p_- \log_2 P_-$$

برای گره اولی با توجه به تعداد خروجی های مثبت و منفی میتوان گفت :

$$E(s) = -(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}) = 0.94$$

برای محاسبه بهره اطلاعات باید انتروپی خروجی را محاسبه کنیم و سپس تا جای امکان به ازای تمام ویژگی ها بهره اطلاعات متناظر را محاسبه و سپس ویژگی با بیشترین بهره اطلاعات را انتخاب می کنیم.

رابطه گین به صورت زیر است :

$$Gain(S, attribute) = E(s) - \sum \frac{S_v}{S} E(S_v)$$

حال باید به سراغ بهره اطلاعات برای هر ویژگی برویم. داریم:

$$Gain(S, Blood\ pressure) = 0.94 - \frac{6}{14}(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - 1/2\log_2(\frac{1}{2}))$$
$$-\frac{8}{14}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) = 0.0481$$

$$Gain(S, cholesterol) = 0.94 - \frac{5}{14}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) - \frac{4}{14}(-\log_2(1))$$
$$-\frac{5}{14}(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) = 0.246$$

$$Gain(S, smoke) = 0.94 - \frac{7}{14}\left(-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}\right)$$

$$-\frac{7}{14}\left(-\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)\right) = 0.151$$

$$Gain(S, weight) = 0.94 - \frac{6}{14}(-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2(\frac{4}{6}))$$
$$-\frac{4}{14}(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2(\frac{1}{4})) - \frac{4}{14}\left(-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 0.029$$

با توجه به مقادیر بهره اطلاعات در اولین مرحله کلسترول را انتخاب میکنیم . برای انتروپی در این مرحله داریم:

$$E(s) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

اگر در مرحله اول کلسترول را انتخاب کرده باشیم آنگاه اگر کلسترول وضعیت بحرانی داشت میتوانیم به طور قطع بگوییم فرد مد نظر مبتلا به انسداد شریان هاست و همچنین انتروپی این شاخه صفر است و نیاز به بررسی بیشتر نیست .

در صورتی که مقدار کلسترول نرمال باشد آنگاه میتوانیم بگوییم :

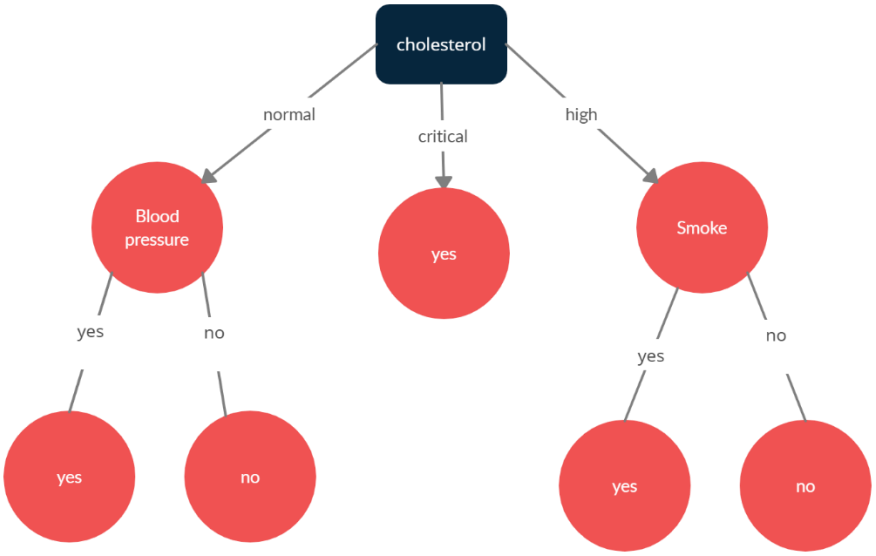
$$Gain(S_{cholesterol}, Blood\ pressure) = 0.97$$
$$-\frac{2}{5}\left(-\frac{2}{2}\log_2 1 - 0\log_2 0\right) - \frac{3}{5}\left(-\frac{3}{3}\log_2 1 - 0\log_2 0\right) = 0.97$$

با توجه به اینکه در این مرحله فشار خون به طرز ایده آلی خروجی را طبقه بندی کرد پس برای شاخه کلسترول نرمال شاخص فشار خون را برای بررسی انسداد شریان ها در نظر می گیریم.

در نهایت برای کلسترول بالا و متغیر کشیدن سیگار داریم :

$$Gain(S_{cholesterol}, Smoke) = 0.97$$
$$-\frac{2}{5}\left(-\frac{2}{2}\log_2 1 - 0\log_2 0\right) - \frac{3}{5}\left(-\frac{3}{3}\log_2 1 - 0\log_2 0\right) = 0.97$$

همانطور که دیده می شود برای کلسترول بالاهم متغیر سیگار کشیدن به طور ایده آل دسته بندی را انجام می دهد . در نهایت با توجه به اینکه در تمام شاخه ها به طور کامل دسته بندی انجام شده است پس کار ساخت درخت تصمیم را تمام شده در نظر میگیریم. درنهایت درخت تصمیم مشابه شکل ۱-۱ میشود.



شکل۱-۱ : درخت تصمیم مناسب برای داده های قسمت الف

ب)

برای پیش بینی خروجی متناظر با هر ورودی باید از ریشه درخت شروع کرد و آنقدر ادامه داد تا در نهایت به پاسخ متناظر برسیم. برای موارد داده شده در صورت سوال داریم:

جدول ۱-۱ : ورودی و خروجی حاصل از مدل درخت تصمیم به دست آمده

فشار خون	سطح کلسترول	مصرف سیگار	وزن	انسداد شرایین	دسته پیشبینی شده
بله	نرمال	بله	چاق	بله	بله
بله	بالا	بله	چاق	بله	بله
بله	بالا	نه	نرمال	نه	نه
بله	نرمال	نه	نرمال	نه	بله
نه	نرمال	بله	اضافه وزن	بله	نه

با توجه به جدول بالا ماتریس آشفتگی به شکل زیر خواهد بود :

جدول ۲-۱ : ماتریس آشفتگی برای مدل به دست آمده و داده های قسمت ب

	مثبت	منفی
مثبت	۲	۱
منفی	۱	۱

با توجه به ماتریس آشفتگی میتوان گفت مدل روی داده های تست عملکرد خیلی خوبی ندارد و صرفا ۶۰ درصد داده ها را به درستی طبقه بندی کرده است . البته باید توجه داشت که داده های آموزش صرفا ۱۴ عدد بودند و نمیتوان از مدل به این سادگی انتظار بالایی داشت . در مجموع میتوان گفت با توجه به ابعاد درخت و حجم داده آموزش عملکرد مدل قابل قبول است .

ج)

در درخت های تصمیم ما به طور مداوم حالت های خاصتر را در نظر میگیریم تا زمانی که به طور کامل خروجی ها از هم جدا شده باشند . مشخصا زمانی که تعداد ویژگی[۱]های داده ها زیاد باشد ممکن است که عمق درخت زیاد شود که در این حالت داده های نویزی حتی اگر به تعداد کم موجود باشند ممکن تاثیر زیادی داشته باشند .

برای رفع این مشکل اولین راه حلی که به ذهن میرسد این است که عمق درخت[۲] را محدود کنیم به این صورت داده های نویزی نمیتوانند تاثیر خیلی زیادی بر روند تشکیل شاخه های درخت بگذارند . راه حل دیگری که در این موارد مطرح میشود این است که داده های آموزش و ارزیابی[۳]تشکیل دهیم و پس از اموزش شروع به هرس کردن درخت به دست آمده کنیم . هرس کردن به این صورت خواهد بود که تاثیر حذف کردن شاخه های مدل به دست آمده را بر روی داده های ارزیابی بررسی کنیم و در صورت داشتن مثبت آنها را حذف کنیم .

[۱] feature
[۲] Max depth
[۳] validation