ب)

خب میدانیم که:

$$E = \frac{1}{N}\sum_{i=1}^{N} exp(-y_i H_T(x_i))$$

و همچنین میدانیم که

$$H_T(x) = \sum_{t=1}^{T} \alpha_t h_T(x)$$

پس اگر این دو را در هم جاگذاری کنیم داریم:

$$E = \frac{1}{N}\sum_{i=1}^{N} exp(-y_i \sum_{t=1}^{T} \alpha_t h_T(x_i)) = \sum_{i=1}^{N} \frac{1}{N} exp(\sum_{t=1}^{T} -\alpha_t y_i h_t(x_i))$$

که همان حکم است.

ت)

برای این ابتدا میاییم $D_{T+1}(i)$ رو نگاه میکنیم. داریم:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} exp(-\alpha_t y_i h_t(i))$$

$$\Rightarrow D_{T+1}(i) = \frac{exp(-\alpha_T y_i h_T(i))}{Z_T} D_T(i) = \frac{exp(-\alpha_T y_i h_T(i))}{Z_T}\frac{exp(-\alpha_{T-1} y_i h_{T-1}(i))}{Z_{T-1}} D_T(i) = \dots$$

$$= \frac{exp(-\alpha_T y_i h_T(i))}{Z_T}\frac{exp(-\alpha_{T-1} y_i h_{T-1}(i))}{Z_{T-1}} \dots \frac{exp(-\alpha_1 y_i h_1(i))}{Z_1} D_0(i)$$

$$= \frac{1}{N}\frac{exp(-\sum_{t=1}^{T} \alpha_t y_i h_t(i))}{\Pi_{t=1}^{T} Z_t} = \frac{1}{N}\frac{exp(-y_i H_T(x_i))}{\Pi_{t=1}^{T} Z_t}$$

خب حالا چون D_{T+1} توزیع است مجموعش ۱ میشود. پس:

$$\Rightarrow \sum_{i=1}^{N} D_{T+1}(i) = 1 \Rightarrow \sum_{i=1}^{N} \frac{1}{N}\frac{exp(-y_i H_T(x_i))}{\Pi_{t=1}^{T} Z_t} = 1 \Rightarrow E = \frac{1}{N}\sum_{i=1}^{N} exp(-y_i H_T(x_i)) = \Pi_{t=1}^{T} Z_t$$

ث)

$i=1$

خب میدونیم که

$$Z_t = \sum_{i=1}^{N} D_t(i)exp(-\alpha_t y_i h_t(x_i))$$

است. حالا اگر $y_i = h_t(x_i)$ باشد، $y_i h_t(x_i) = 1$ میشود و اگر $y_i \neq h_t(x_i)$ باشد، $y_i h_t(x_i) = -1$ میشود. پس عبارت بالا رو میتونیم به صورت زیر بنویسیم:

$$Z_t = \sum_{y_i \neq h_t(x_i)} D_t(i)exp(\alpha_t) + \sum_{y_i = h_t(x_i)} D_t(i)exp(-\alpha_t)$$

$$= exp(\alpha_t) \sum_{y_i \neq h_t(x_i)} D_t(i) + exp(-\alpha_t) \sum_{y_i = h_t(x_i)} D_t(i)$$

خب اگر دقیق نگاه کنیم، میبینیم که $\epsilon_t = \sum_{y_i \neq h_t(x_i)} D_t(i)$ است. از طرفی چون D_t توزیع است داریم:

$$\sum_{y_i \neq h_t(x_i)} D_t(i) + \sum_{y_i = h_t(x_i)} D_t(i) = 1 \Rightarrow \sum_{y_i = h_t(x_i)} D_t(i) = 1 - \epsilon_t$$

پس:

$$\Rightarrow Z_t = exp(\alpha_t)\epsilon_t + exp(-\alpha_t)(1 - \epsilon_t)$$

ج)

خب این هم ساده است. کافیه از Z_t بر حسب α_t مشتق بگیریم و مساوی صفر قرار بدیم.

$$Z_t = exp(\alpha_t)\epsilon_t + exp(-\alpha_t)(1 - \epsilon_t) \Rightarrow \frac{\partial Z_t}{\partial \alpha_t} = exp(\alpha_t)\epsilon_t - exp(-\alpha_t)(1 - \epsilon_t) = 0$$

$$\Rightarrow exp(2\alpha_t) = \frac{1 - \epsilon_t}{\epsilon_t} \Rightarrow \alpha_t = \frac{1}{2}ln(\frac{1 - \epsilon_t}{\epsilon_t})$$

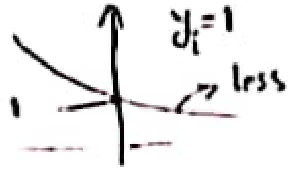
الف)

$$E = \frac{1}{N}\sum_{i=1}^{N} exp(-y_i H_t(x_i)) = \frac{1}{N}\sum_{\hat{y}_i = y_i} exp(-y_i H_t(x_i)) + \frac{1}{N}\sum_{y_i \neq \hat{y}_i} exp(-y_i H_t(x_i))$$

$$\geq \frac{1}{N}\sum_{y_i \neq \hat{y}_i} \underbrace{exp(-y_i H_t(x_i))}_{\geq 1} \Rightarrow$$

زیرا loss به شکل نمایی است که زمانی که اشتباه دسته‌بندی کنیم جریمه‌ای بیشتر از 1 دارد.

$$\boxed{\geq \frac{1}{N}\sum_{y_i \neq \hat{y}_i} 1 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i \neq y_i)}$$



ب)

ب) فرض کنید $D_t(i)$ توزیعی وزنی داده ها در مرحله t است. $D_{t+1}(i)$ را بر حسب x_i, y_i, α_t, D_t و دسته بند h_t بیابید. T گام نهایی است و به یاد داشته باشید Z_t عامل نرمال کننده D_{t+1} است:

$$Z_t = \sum_{i=1}^{N} D_t(i) exp(-\alpha_t y_i h_t(x_i))$$

خب همان طور که ازین عامل نرمال کننده معلوم است، D_{t+1} برابر است با:

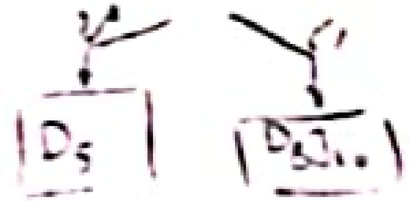
$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t} = \frac{D_t(i) exp(\alpha_t y_i h_t(x_i))}{\sum_{i=1}^{N} D_t(i) exp(-\alpha_t y_i h_t(x_i))}$$

جنسیت انتخاب می‌شود.

وضعیت ... اوهایو :

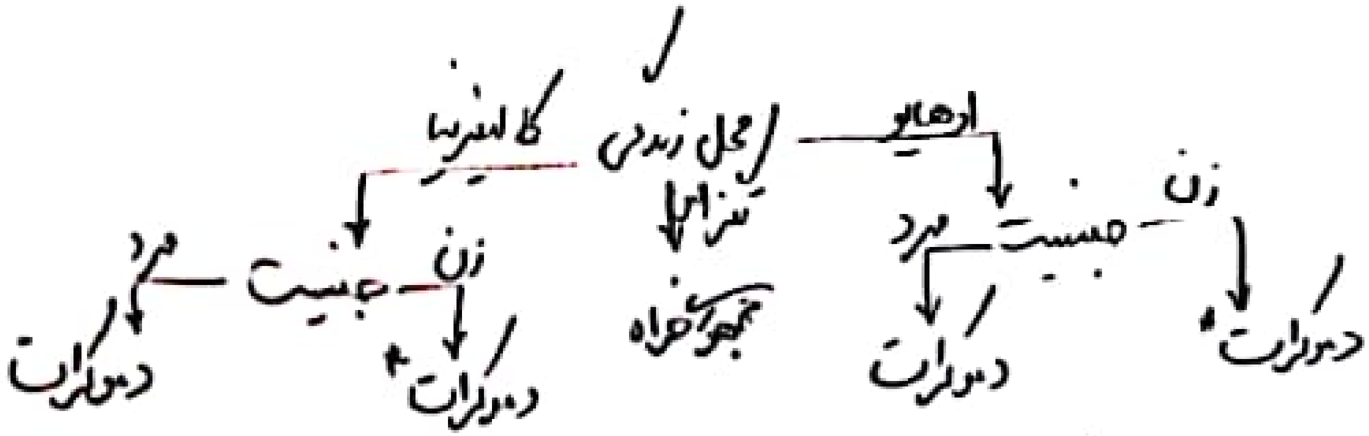
GI (اوهایو) ، مانند برابر GI (تگزاس)

GI (جنسیت) $: \frac{2}{3}\left(\frac{1}{2}(\frac{1}{2})+\frac{1}{2}(\frac{1}{2})\right) = \frac{1}{3}$



جنسیت انتخاب می‌شود

گره (۲) هم مشخص است بنابراین تا عمق ۲ درخت رسم می‌شود



دموکرات* : می‌توان هم دموکرات و هم جمهوری‌خواه در نظر گرفت چون که احتمال برابری دارند.

Gini Impurity for (نژاد) $= \frac{1}{2}\left[\frac{2}{5}\left(\frac{3}{5}\right)+\frac{3}{5}\left(\frac{2}{5}\right)\right]+\frac{1}{2}\left[\frac{3}{5}\left(\frac{2}{5}\right)+\frac{2}{5}\left(\frac{3}{5}\right)\right] = \frac{12}{25}$

GI(درآمد) $= \frac{4}{10}\left[\frac{1}{4}\left(\frac{3}{4}\right)+\frac{3}{4}\left(\frac{1}{4}\right)\right]+\frac{3}{10}\left[0\right]+\frac{3}{10}\left[\frac{1}{3}\left(\frac{2}{3}\right)+\frac{2}{3}\left(\frac{1}{3}\right)\right] = \frac{3}{20}+\frac{2}{15}$

GI(جنسیت) $= \frac{1}{2}\left[\frac{2}{5}\left(\frac{3}{5}\right)+\frac{3}{5}\left(\frac{2}{5}\right)\right]+\frac{1}{2}\left[\frac{3}{5}\left(\frac{2}{5}\right)+\frac{2}{5}\left(\frac{3}{5}\right)\right] = \frac{12}{25} = 0.48$

GI(کل نژاد) کمینه است پس بعنوان ریشه انتخاب می‌شود

محاسبات فرعی گره (۱)
[illegible] :

GI(نژاد) $= \frac{1}{4}[0]+\frac{3}{4}\left[\frac{2}{3}\left(\frac{1}{3}\right)+\frac{1}{3}\left(\frac{2}{3}\right)\right] = \frac{1}{3}$

سیاه → D4
سفید → D1 D2 D3

GI(جنسیت) $= \frac{1}{2}[0]+\frac{1}{2}\left[\frac{1}{2}\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right] = \frac{1}{4}$

زن → D1 D2
مرد → D3 D4

جنسیت انتخاب می‌شود.

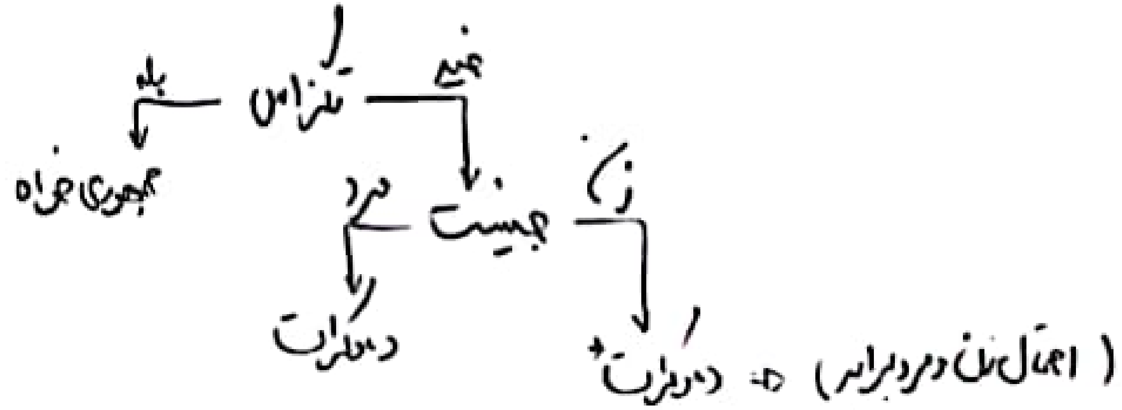
$D_1 D_2 J_6 D_7 \xleftarrow{\text{بله}}$ $GI($کالیفرنیا$) = \frac{4}{7}\left[\frac{3}{4}\left(\frac{1}{4}\right) + \frac{1}{4}\left(\frac{3}{4}\right)\right] + \frac{3}{7}\left[\frac{2}{3}\left(\frac{1}{3}\right) + \frac{1}{3}\left(3\right.\right.$

$D_5 D_3 J_{10} \xleftarrow{\text{خیر}}$

$D_5 D_3 J_{10} \xleftarrow{\text{بله}}$ $GI($اوهایو$) = \frac{3}{7}\left[\frac{2}{3}\left(\frac{1}{3}\right) + \frac{1}{3}\left(\frac{2}{3}\right)\right] + \frac{4}{7}\left[\frac{3}{4}\left(\frac{1}{4}\right) + \frac{1}{4}\left(\frac{3}{4}\right)\right] = 0.405$

$D_1 D_2 J_6 D_7 \xleftarrow{\text{خیر}}$

بهترین انتخاب گره «جنسیت» است. بنابراین درخت به صورت زیر است



سه) زمانی که از multiple split استفاده می‌کنیم در مقایسه با Binary split محاسبات کمتری انجام میدهیم اما داده‌ها سریع می‌شکنند و به زیرمجموعه‌های کوچک افراز می‌شوند. بنابراین مدل ما بیشتر به سمت overfit شدن پیش می‌رود. یعنی به جزئی کردن داده‌ها درخت هم تغییر می‌کند و واریانس مدل نسبتاً زیاد است. در مقابل binary split محاسبات بیشتری دارد اما دیرتر به سمت overfit شدن می‌رود و مدل‌های ساده‌تری به دست می‌دهد که واریانس کمتری دارند. در مقابل ممکن است این مدل به اندازه‌ی کافی دقیق نباشد و bias بیشتری نسبت به Multi split داشته باشد.

ج) برای نمایش داده‌ی iام از زیرنویس i استفاده می‌کنیم. داده‌ی دموکرات را با D و جمهوری‌خواه را با R نمایش می‌دهیم

بنابراین به عنوان مثال داده‌ی ۸ام را با D_5 نمایش می‌دهیم.

محل زندگی:
- کالیفرنیا → (1): $D_1\ D_2\ J_6\ D_9$
- تگزاس → (2): $J_3\ J_4\ J_7$
- اوهایو → (3): $D_5\ D_8\ J_{10}$

نژاد:
- سیاه → $J_4\ D_5\ J_{10}\ D_8\ D_9$
- سفید → $D_1\ D_2\ J_3\ J_6\ J_7$

جنسیت:
- مرد → $D_1\ D_2\ J_3\ J_4\ D_5$
- زن → $J_6\ J_7\ D_8\ D_9\ J_{10}$

$$\underset{i \in \text{remaining attr.}}{\arg\max}\ \mathrm{Gain}(S, X_i) = \underset{i \in \text{remaining attr}}{\arg\min}\ GI(Y|X_i$$

۱) می‌دانیم که اگر ویژگی m ها binary باشند تعداد درخت تصمیم برابر با تعداد توابع binary است. برای این کار جدول درستی یک تابع binary با m متغیر را در نظر بگیرید. این جدول 2^m سطر دارد. چون مسئله دو کلاس دارد هر سطر دارای جدول می‌تواند مقدار ۰ یا ۱ باشد. بنابراین 2^{2^m} تا درخت متفاوت داریم. [redacted]

ب) خیر. به طور کلی روش الگوریتم حریصانه جزو روش بهینه هستند و در اینجا هم می‌دانیم مسئله پیدا کردن کوچکترین درخت تصمیم‌گیری NP-complete است. یک مثال نقض برای بهینه نبودن روش حریصانه این مثال است که دو ویژگی اول به تنهایی mutual info صفر دارند اما در کنار هم ماهیت اطلاعات ارائه می‌کنند. روش حریصانه ویژگی سوم را برمی‌گزیند و سپس ویژگی بعدی را.

0,0,0 → -	1,0,1 → +
0,0,0 → -	1,0,1 → +
0,0,1 → -	1,0,0 → +
0,1,1 → +	1,1,0 → -

البته نکته دیگر همانی است که مزیت کوچکترین درخت، بهترین درخت نیست؛ چون که ممکن است overfit کرده باشد.

به علاوه به این دلیل که در فرآیند ساختن درخت از عقبگرد استفاده نمی‌کنیم، درخت بهینه کلی پیدا نخواهد شد.

ت)

در الگوریتم جنگل تصادفی از مفهومی به اسم Bagging استفاده می‌شود. در Bagging ما با استفاده

از تعدادی classifier ضعیف‌تر، classifier ای می‌سازیم که واریانس کمتری داشته باشد. تنها نکته‌ای که

در Bagging وجود دارد این است که classifier های استفاده شده در آن باید با یکدیگر Corrolation

داشته باشند. به همین دلیل از تکنیک bootstrap aggregating استفاده می‌شود که در آن Dataset هایی با ا

عملی از Dataset مرجع با جایگذاری برداشته می‌شود و هر classifier روی یک sample آموزش می‌بیند.

همچنین در جنگل تصادفی برای استقلال بیشتر هر بار یک زیرمجموعه از ویژگی‌ها را انتخاب می‌کنند درخت را روی آن

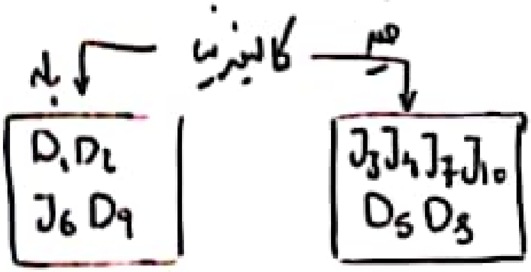
آموزش می‌دهند. دلیل این امر آن است که اگر در بین ویژگی‌ها چند ویژگی خوب وجود داشته باشد

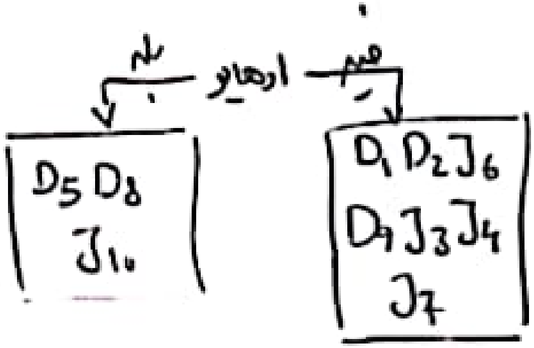
درخت همراه آن‌ها را انتخاب می‌کند و درخت‌ها شبیه هم می‌شوند.

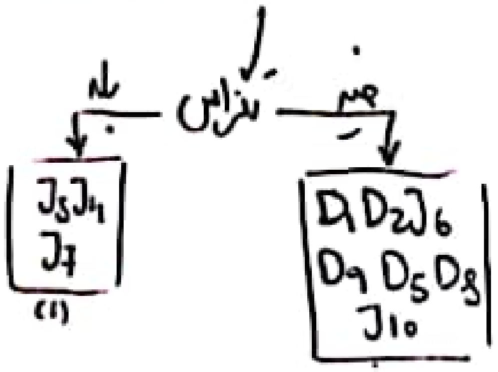
در این الگوریتم بین پایه درخت‌های مختلف رأی‌گیری می‌شود به همین دلیل دقت آن به خوبی دقت تک درخت

باقی می‌ماند.

ج) برای این که مسئله تبدیل به مسئله binary شود می‌توانیم کل حالت زندگی را به صورت ۳ سؤال آیا در (کالیفرنیا / اوهایو / تگزاس) زندگی می‌کند بشکنیم.







$$GI(\text{کالیفرنیا}) = \frac{4}{10}\left(\frac{3}{4}\left(\frac{1}{4}\right) + \frac{1}{4}\left(\frac{3}{4}\right)\right) + \frac{6}{10}\left(\frac{2}{3}\left(\frac{1}{3}\right) + \frac{1}{3}\left(\frac{2}{3}\right)\right) = 0.417$$

$$GI(\text{تگزاس}) = \frac{3}{10}[0] + \frac{7}{10}\left[\frac{5}{7}\left(\frac{2}{7}\right) + \frac{2}{7}\left(\frac{5}{7}\right)\right] = 0.286$$

$$GI(\text{اوهایو}) = \frac{3}{10}\left[\frac{2}{3}\left(\frac{1}{3}\right) + \frac{1}{3}\left(\frac{2}{3}\right)\right] + \frac{7}{10}\left[\frac{3}{7}\left(\frac{4}{7}\right) + \frac{4}{7}\left(\frac{3}{7}\right)\right] = 0.476$$

گره اول (root) گره تگزاس می‌شود.

برای گره b(2) داریم:

$D_5 D_8 D_9$ J_{10} ← سیاه

$$GI(\text{نژاد}) = \frac{4}{7}\left(\frac{3}{4}\left(\frac{1}{4}\right) + \frac{1}{4}\left(\frac{3}{4}\right)\right) + \frac{3}{7}\left(\frac{2}{3}\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right) = 0.405$$

$D_1 D_2 J_6$ ← سفید

$D_1 D_2 D_5$ ← مرد

$$GI(\text{جنسیت}) = \frac{3}{7}(0) + \frac{4}{7}\left[\frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}\left(\frac{1}{2}\right)\right] = 0.286$$

$J_6 D_9 D_8$ J_{10} ← زن

Problem 5.3

- (a) $M = 1$, we only have one hypothesis function.
- (b) Hoeffding bound $P[|E_{out} - E_{in}| > 0.02] \le 2e^{-2\times 0.02^2 \times N} = 0.00067092525580502370$. It says that for a high confidence,the out-of-sample error should be very close to our in-sample error, which is zero.
- (c) It's possible that the first N customers all come from same local area, i.e. we may have sample bias. So the model works on such sample but not on all samples.
- (d) The bank may apply your function to certain type of customers once it knows the implicit assumption in the function.