

## مدرس سوم درس داده کاری - ناطه عالم زاده - ۹۵۳۱.۷

سوال ۱

classification accuracy: مکانی از تائید چندین مدل (متا کلاس) در پیش‌نمایش (ensemble) ensemble مدل (متا کلاس) با هم ترکیب شوند به این صورت که بیک مدل classification بیور یافته حاصل شود.

$$M_1, M_2, \dots, M_K \rightarrow M_*$$

در روش boosting به هر تاپل داده training می‌باشد وزن نسبت داده شود. جو مدل‌های از classifier که طور تسلیم (learn) شوند. پس از اینکه مدل  $M_i$  آموزش دید، وزن‌ها آپدیت می‌شوند تا به تاپل می‌باشند که توسط  $M_i$  به اشتایه کلاس (نیز) شده باشند و وزن بیشتری داده شود یا به عبارتی همان مدل‌ها بیشتر توجه می‌شوند!

مدل کنایی ( $M_*$ ) را (که) هر classifier را می‌تیرد و آن مدل را با هم تجمع می‌کند به گونه‌ای که وزن هر classifier تابعی از دقت (accuracy) آن است.

یکی از الگوریتم‌های معروف boosting اسیدویتم Ada Boot می‌باشد. در این اگوریتم هر دست boosting این‌طور به هر تاپل وزنی دهیم و وزن‌ها را آپدیت می‌کنیم باعث افزایش دست تصور می‌شوند که مدل misclassified را در صورت توجه قراری دهیم.

یادهای سوال ۱:

gradient boosting: این تکنیک تعداد زیادی از عمل‌های راهنمایی (gradual) اترابینه (additive) و پیوستی (sequential) آموزش می‌دهد. تفاوت اساسی میان این دو روش adaboost است که این دو، نقضیت‌های خود را تغییر می‌دهند.

در adaboost این عمل با استفاده از نقطه ای که وزن بالایی درین صورتی نیز ایجاد نماید. این روش با استفاده از گرادیان (gradient) loss می‌باشد که معمولاً loss می‌باشد. تابع loss بیکار است که شانسی دارد ضرایب عمل چقدر برای fit شون به داده‌ی اصلی خوب هستند.

این اصلی این دستوریستم این است که base\_learner چندی بسازد که بینشیدن هم‌ستگی را با گرادیان منتهی تابع loss داشته باشد. تابع loss می‌تواند به طور دکتره‌انهای انتخاب شود و به حوزه‌ای که researcher کاری کند مرتبط است.

CART (Classification And Regression Tree) می‌باشد gradient boosting معمولاً از رختهای ریختی (Decision Tree) به عنوان base learner استفاده می‌کند.

$$\begin{aligned}
 \text{accuracy} &= \frac{TP + TN}{P + N} = \frac{TP}{P+N} + \frac{TN}{P+N} \\
 &= \frac{TP}{P+N} \times \frac{P}{P} + \frac{TN}{P+N} \times \frac{N}{N} \\
 &= \underbrace{\frac{TP}{P} \times \frac{P}{P+N}}_{\text{recall}} + \underbrace{\frac{TN}{N} \times \frac{N}{P+N}}_{\text{specificity}} \\
 &= \text{recall} \left( \frac{P}{TP+TN} \right) + \text{specificity} \left( \frac{N}{TP+TN} \right)
 \end{aligned}$$

A	B	C	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	0

$$\begin{aligned}
 \text{Gain}(A) &= \text{Info}(D) - \text{Info}_A(D) \quad \text{مسئلہ} \\
 \text{Info}(D) &= -\sum_{i=1}^m p_i \log_p(p_i) = -\left(\frac{1}{4} \lg \frac{1}{4} + \frac{3}{4} \lg \frac{3}{4}\right) \approx 0.94 \\
 \text{Info}_A(D) &= \sum_{j=1}^r \frac{|D_j|}{|D|} \times \text{Info}(D_j) = \frac{1}{4} \left(\frac{1}{4} \lg \frac{1}{4} + \frac{1}{4} \lg \frac{1}{4}\right) \\
 &\quad - \frac{1}{4} \left(\frac{1}{3} \lg \frac{1}{3} + \frac{2}{3} \lg \frac{2}{3}\right) \approx 0.9
 \end{aligned}$$

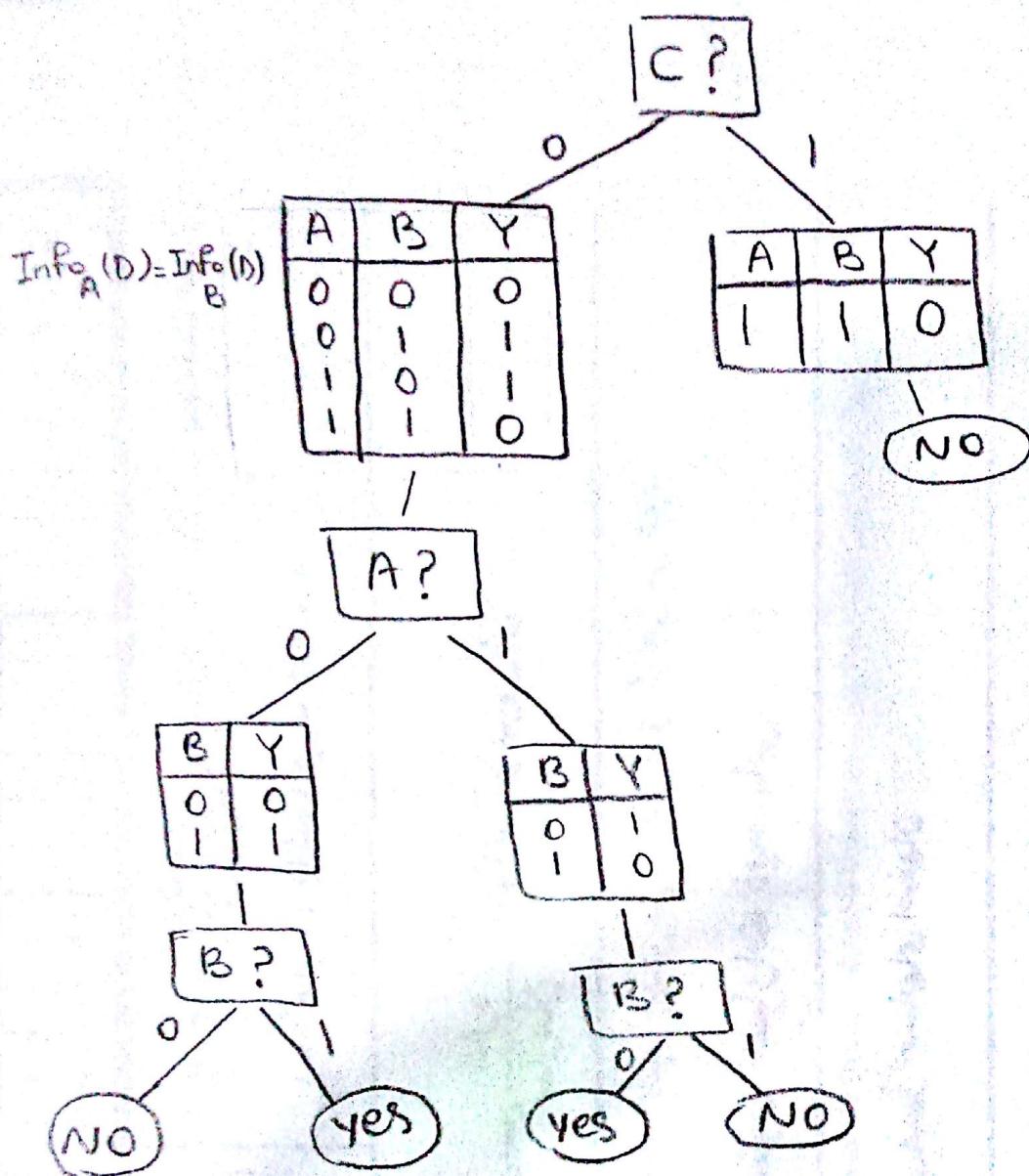
$$\text{Info}_B(D) = -\frac{1}{2} \left(\frac{1}{2} \lg \frac{1}{2} + \frac{1}{2} \lg \frac{1}{2}\right) - \frac{1}{2} \left(\frac{1}{2} \lg \frac{1}{2} + \frac{1}{2} \lg \frac{1}{2}\right) \approx 0.9$$

$$\text{Info}_C(D) = -\frac{1}{3} \left(\frac{1}{3} \lg \frac{1}{3} + \frac{1}{3} \lg \frac{1}{3} + \frac{1}{3} \lg \frac{1}{3}\right) - \frac{1}{3} \left(\frac{1}{2} \lg \frac{1}{2} + \frac{1}{2} \lg \frac{1}{2}\right) = -\frac{1}{3} \times \lg \frac{1}{2} = \frac{1}{3} \approx 0.1$$

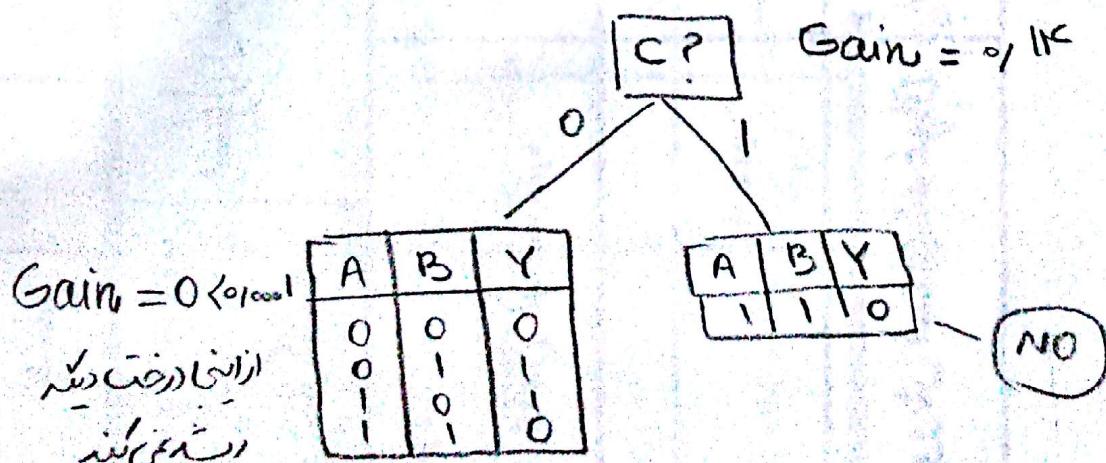
$$\text{Info}_B(D) = \text{Info}_A(D) > \text{Info}_C(D)$$

از آن جایی که  $\text{Info}(D)$  برای همه سه متغیر متساوی است، مفهوم  $\text{Info}_A(D)$  کمتر باشد. دادا (Gain بیشتر) خواهد بود بنابراین  $ID_3$  را درین مرحله دنبال کرده است.

(C)

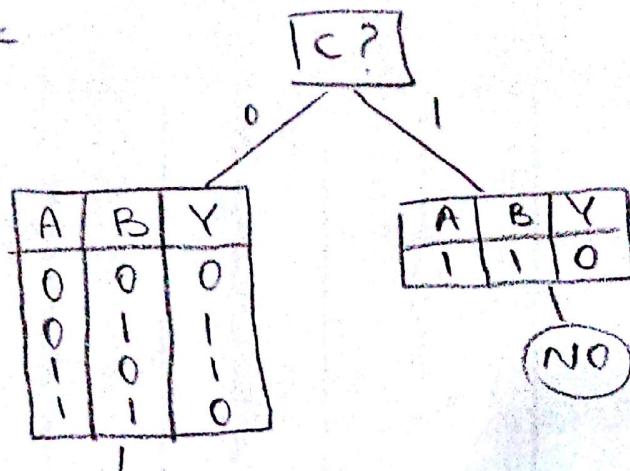


(E)

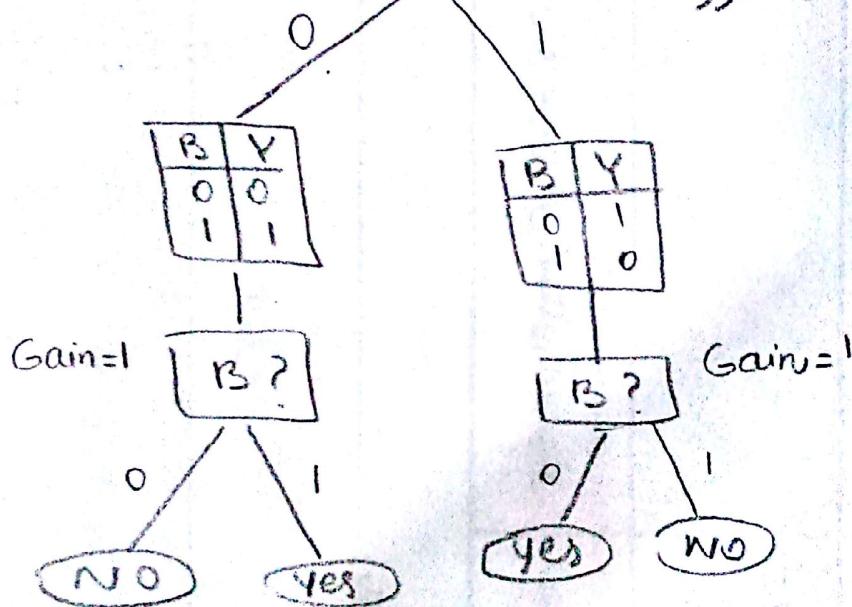


Gain = 0.1K

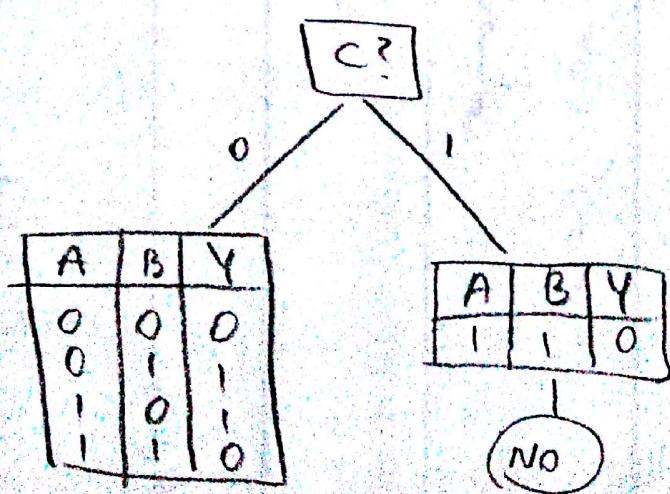
(P)



Gain = 0 → prune X  
یا شور



مساءلہ پر یقینی



ه) در روش prepruning بدلیل دقت بوسایل توقف ساخت آن هرگز می‌شود و بونه کمتر  
تر از آن تراویح اتفاق نمی‌افتد. بفرمایشی شود. بلکه از طبقاتی این روش انتساب  
بیل threshold می‌نماید است. آن threshold بالا باید باشد تا بتوان در حفظ تمامی  
سمت‌اللسانی (نمایندگان threshold) کمیاباتی توان ساده‌سازی کمی بتوان درستی  
درخت به خوبی هرس می‌شود.

(روش postpruning) زیردرخت هایی را از درخت کامل شده حذف می‌نمایم. بدلیل زیردرخت با  
حذف شاخه‌های و مطابق‌داری بدلیل فرمایشی آنها، هرس می‌شود. این روش بیاز به  
عایقاب می‌شود اما بمقابله prepruning دارای این دلیل مطیع نیست.

سوال ۱۴) خیز. ID3 تفکین علی کنترلر های عواید چندی (کنترلر اسپرسی) کرده.

این الگوریتم به کمینه‌سازی علی قدر تراویح شود. از یک استراتژی غریبان (greedy) استفاده می‌کند زیرا در هر مرحله local attribute نسبت صورت در آن جمله بین است را انتخاب می‌کند تا dataset را برای این split کند.  
بینگ این الگوریتمی تواند توسط backtracking افزایش یابد اما حذف یک علاوه بر این داشت.

سوال ۱۵) واحد اندازه‌گیری می‌تواند برآن نشود اگر این اندازه‌گیری طول سبدی واحد از متربه اینچ یا برای اندازه‌گیری وزن، سیوکرم به بوند، می‌تواند نتیجه متفاوت را نتیجه بدهد  
به طور مخصوص بین بیک دیگری (attribute) با واحد های (attribute) کوچک تر باشد می‌شود که آن range از attribute بزرگتر شود و آن attribute وزن یا تاثیرگذاری تعلق نیافردد. برای دوری از از مسئله دستیابی normalize شود به گونه‌ای که بعد از خود نشان بدهد [۱،۱]

۱) [۱،۰] نتیجه شود.

نحوی کردن دستیابی دارد که به علی attribute صادر از نیافرده. همچنین نحوی کردن دستیابی می‌شود دستیابی و در دری رستار نیافرده از خود نشان بدهد  
(رسانی classification) اگر دستیابی را نحوی نکنیم ای از دارای scale attribute

بزرگتری است برنتیه نتیجه نهاد و آن را به شدت تحت تاثیر قرار گرفته.