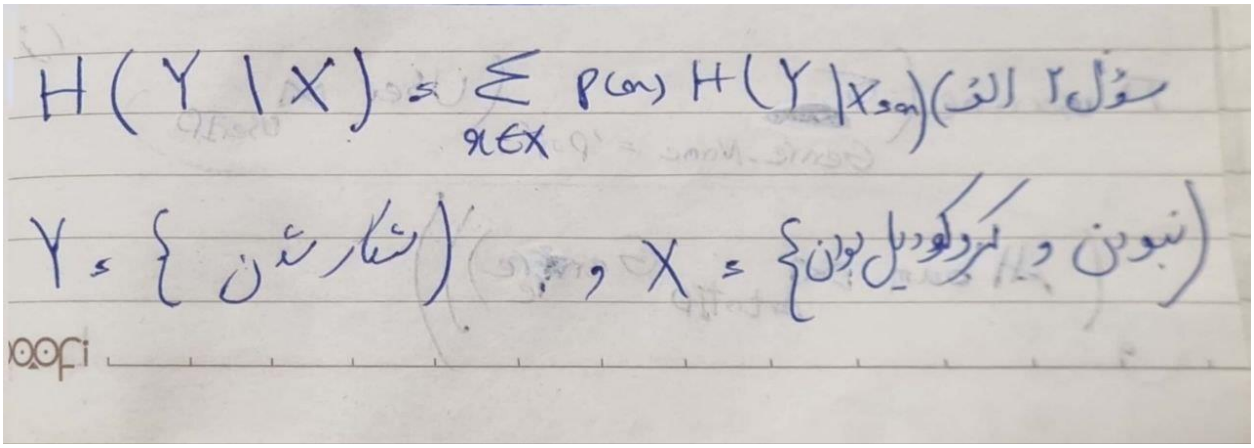
Mahan Veisi – ML HW1 – 400243081

1- میدانیم که مشکل اورفیت زمانی رخ میدهد که مدل ما به دلیل وابستگی زیاد به داده های موجود خودش (به دلیل کمبود دیتا یا حتی مشکل ساختاری مدل) نتواند روی داده های دیده نشده از قبل (تست دیتا) پردیکشن درستی داشته باشد.
با توجه به تعریف ارائه شده، میدانیم در رندوم فورست اول از همه به نوعی از یک زیرمجموعه از کل داده های موجود استفاده میکنیم زیرا داده ها برای train کردن هر درخت میتوانند تکراری(with replacement) و با شانس برابر باشند و پس از پایان ساختن درخت ها و ساخت جنگل، امکان داشته باشد که هر درخت داده هایی را بررسی کند که دیده نشده باشند (bagging). از طرفی میدانیم معمولا یکی از نواقص ساخت تک درخت ها (درخت تصمیم)، اورفیت شدن میباشد.
پس اگر ما در رندوم فورست بتوانیم دقت train را بالا ببریم، نسبت به گذشته (که از درخت تصمیم استفاده میکردیم) اطمینان بیشتری داریم که در مقابل داده های دیده نشده، به overfit دچار نشویم (اختلاف بین دقت train و test کمتر است). در نهایت برای پیشبینی کردن از همه درخت ها استفاده میکنیم که این خود نیز یک نوع تکنیک برای regularization است.

2- الف)



$$H(Y|X) = \frac{64}{100} H(Y|\text{کروکودیل}) + \frac{36}{100} H(Y|\text{بقیه‌ی کروکودیل})$$

$$H(Y|\text{کروکودیل}) = -P(Y|\text{کروکودیل}) \cdot \log_2 (P(Y|\text{کروکودیل})) - P(\neg Y|\text{کروکودیل}) \cdot \log_2 (P(\neg Y|\text{کروکودیل}))$$

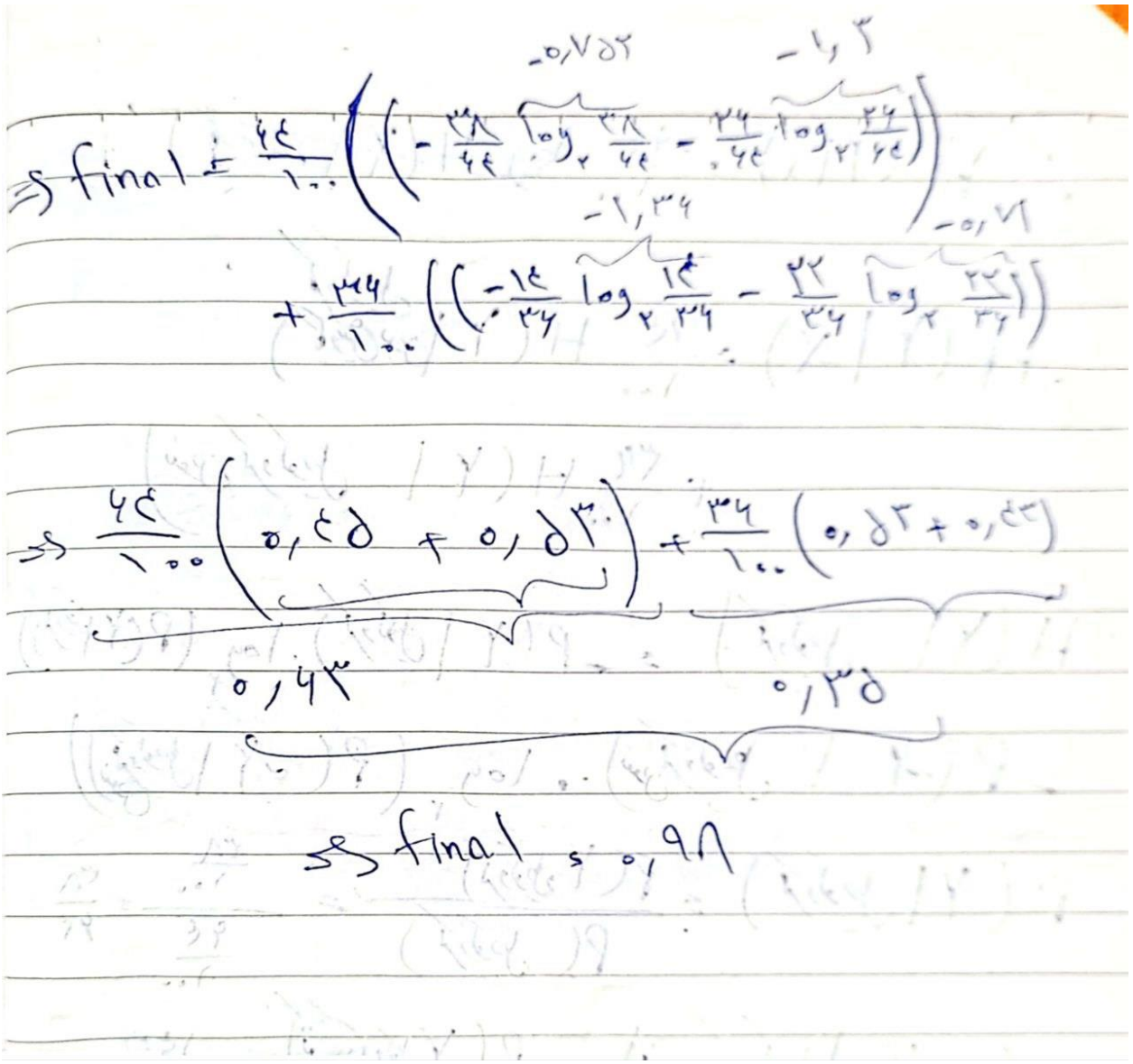
$$P(Y|\text{کروکودیل}) = \frac{P(Y \text{ و کروکودیل})}{P(\text{کروکودیل})} = \frac{\frac{38}{100}}{\frac{64}{100}} = \frac{38}{64}$$

$$P(\neg Y|\text{کروکودیل}) = \frac{P(\neg Y \text{ و کروکودیل})}{P(\text{کروکودیل})} = \frac{\frac{26}{100}}{\frac{64}{100}} = \frac{26}{64}$$

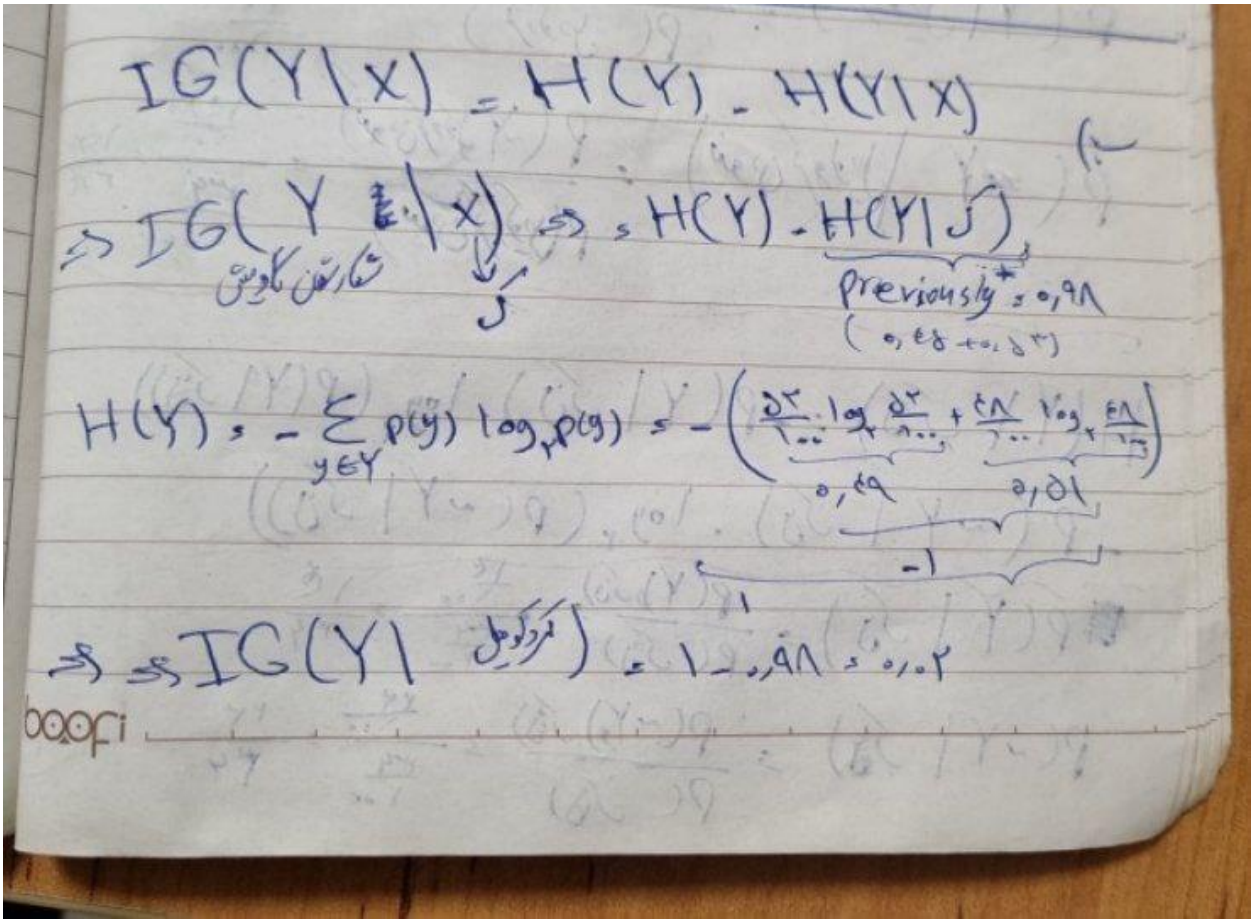
$$H(Y|\neg\text{کروکودیل}) = -P(Y|\neg\text{کروکودیل}) \cdot \log_2 (P(Y|\neg\text{کروکودیل})) - P(\neg Y|\neg\text{کروکودیل}) \cdot \log_2 (P(\neg Y|\neg\text{کروکودیل}))$$

$$P(Y|\neg\text{کروکودیل}) = \frac{P(Y \text{ و } \neg\text{کروکودیل})}{P(\neg\text{کروکودیل})} = \frac{\frac{14}{100}}{\frac{36}{100}} = \frac{14}{36}$$

$$P(\neg Y|\neg\text{کروکودیل}) = \frac{P(\neg Y \text{ و } \neg\text{کروکودیل})}{P(\neg\text{کروکودیل})} = \frac{\frac{22}{100}}{\frac{36}{100}} = \frac{22}{36}$$


-0,٧٥٢
-١,٣
-١,٣٤
-٠,٤٧
final = ٤٤/١٠٠ ((- ٣٨/٤٤ log٢ ٣٨/٤٤ - ٣٤/٤٤ log٢ ٣٤/٤٤))
+ ٣٤٤/١٠٠ (((- ١٤/٣٤٤ log٢ ١٤/٣٤٤ - ٢٢/٣٤٤ log٢ ٢٢/٣٤٤))
٤٤/١٠٠ (٠,٤٥ + ٠,٥٣٠) + ٣٤٤/١٠٠ (٠,٥٣ + ٠,٤٧)
٠,٤٣
٠,٣٥
final = ٠,٩٨

ب)


IG(Y|X) = H(Y) - H(Y|X)
IG(Y|X) = H(Y) - H(Y|X)
previously = ۰,۹۸
H(Y) = - Σ P(y) log₂P(y) = -
۰,۴۹
۰,۵۱
-۱
IG(Y|X) = ۱ - ۰,۹۸ = ۰,۰۲

ج)

ج)

$H($ نک $| \sim Y)$

↓ نگارش

$= -P($ نک $| \sim Y) \log_۲ P($ نک $| \sim Y) -$

$- P($ ک $| \sim Y) \log_۲ P($ ک $| \sim Y)$

$P($ نک $| \sim Y) = \dfrac{P(\text{نک}, \sim Y)}{P(\sim Y)} = \dfrac{\frac{۲۲}{۲۰۰}}{\frac{۴۸}{۲۰۰}} = \dfrac{۲۲}{۴۸}$

$P($ ک $| \sim Y) = \dfrac{P(\text{ک}, \sim Y)}{P(\sim Y)} = \dfrac{\frac{۲۶}{۲۰۰}}{\frac{۴۸}{۲۰۰}} = \dfrac{۲۶}{۴۸}$

$\Rightarrow H($ نک $| \sim Y) = \underbrace{-\dfrac{۲۲}{۴۸} \log_۲ \dfrac{۲۲}{۴۸}}_{۰٫۵۱۶} \underbrace{- \dfrac{۲۶}{۴۸} \log_۲ \dfrac{۲۶}{۴۸}}_{۰٫۴۸} =$

$\Rightarrow H($ نک $| \sim Y) = ۰٫۹۹$