Question ٦:

Random forest در واقع مجموعه ای از درخت تصمیم هستند که در
پایان فرایند کلاسیک یک میانگین یا قوانین اکثریت
ترکیب شده اند

در Gradient boosting machine نیز درخت های تصمیم
ترکیب می شوند با این تفاوت که فرایند ترکیب از
ابتدا (بجای پایان فرایند) انجام می شود

Random forest نیز مانند Gradient boosting machine مجموعه ای از
درخت تصمیم است با دو تفاوت اصلی:

۱- تفاوت در نحوه ی ساخت درخت: Random forest
هر درخت را به صورت مستقل می سازد در حالی که
Gradient boosting هر بار یک درخت را می سازد.

۲- ترکیب نتایج: Random forest نتایج را در انتهای
فرایند (با میانگین یا قوانین اکثریت) ترکیب می کند
در حالی که Gradient boosting نتایج را در طی مسیر
انجام می شود

از پارامترهای بیشتری برای تنظیم در gradient boosting نسبت به
بیشتری از random forest می‌دهد. بنابراین حساسیت به noise باید
باشد. gradient boosting مناسب‌تر نیست
ممکنه باعث overfitting شه. همچنین تنظیم
پارامترهای gradient boosting نسبت به random forest
سخت‌تر است.

random forest و gradient boosting هر دو
نقاط قوت خود را دارند. random forest در
طبقه‌بندی و تشخیص objectهای چندکلاسه که
در آنها آمار بالایی noise همراه است و gradient boosting
برای داده‌های نامتعادل مثل ارزیابی ریسک استفاده
می‌شود

سوال ۲

درخت روی داده های آموزشی overfit شده

(شاخه های زیادی داره الان)

این داده ای که ندیده پس دیگه دقت پایینی داره

دو تا راه برای جلوگیری از overfitting وجود داره

۱. prepruning ← که gain(A) از یه حدی کمتر شد ادامه ندیم

۲. postpruning ← درخت رو که ساختیم احتمالا سعی میکنیم node های انتهایی رو حذف کنیم و تاثیرش رو روی دقت (accuracy) ببینیم اگر خیلی فرقی نکرد هرسش کنیم

سوال ۳

A/p	C	¬C
C	10 (TP)	15 (FN)
¬C	25 (FP)	100 (TN)

Accuracy = 73.3%

A/p	C	¬C
C	0	25
¬C	0	125

Accuracy = 83.3%

زمانی که $TP < FP$ باشد و دسته منفی را بیشتر پیش بینی
می کند دقت زیاد می شود
یا زمانی که $TN > FN$ باشد

$$recall = \frac{TP}{TP+FN} \quad , \quad precision = \frac{TP}{TP+FP}$$

↓

از بین همه مواردی که C1 تشخیص داده
چه تعداد شان درست بوده

یا از بین تمام C1 ها
چه تعدادش را درست تشخیص
داده

یه معیار خیلی خوب

$$F_1\text{-}measure = \frac{2\,precision \times recall}{precision + recall}$$

سوال ۴

تعداد حالاتی که بیش از نیمی از classifier ها اشتباه کنند شمرده می‌شود

پس، در روش ensemble اینطور حساب میشود

$$e_{ensemble} = \sum_{i=13}^{25} \binom{25}{i} (0.35)^i (1-0.35)^i = 0.06$$

سوال ۵

$$\text{Info}(D) = -\frac{5}{14}\log\frac{5}{14} - \frac{9}{14}\log\frac{9}{14} = 0.94$$

$$\text{Info}_{Habitat}(D) = \frac{5}{14}\times 0.97 + \frac{5}{14}\times 0.97 = 0.692$$

$$D_{woods} = D_{Leaves} = [3+, 2-] : E = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$$

$$D_{grasses} = [4+, 0] : E = 0$$

$\hookrightarrow$ $\text{Gain}(D, \text{Habitat}) = 0.247$ ✓ max

$$\text{Info}_{Cap\ Color}(D) = \frac{4}{14}\times 1 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.811 = 0.91$$

$D_{Red}[2+, 2-] : E = 1$ $D_{Green}[4+, 2-] : E = 0.918$

$D_{white}[3+, 1-] : E = 0.811$

$\hookrightarrow$ $\text{Gain}(D, \text{Cap Color}) = 0.029$

$\text{Info}_{\text{cap shape}}(D) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.591 = 0.788$

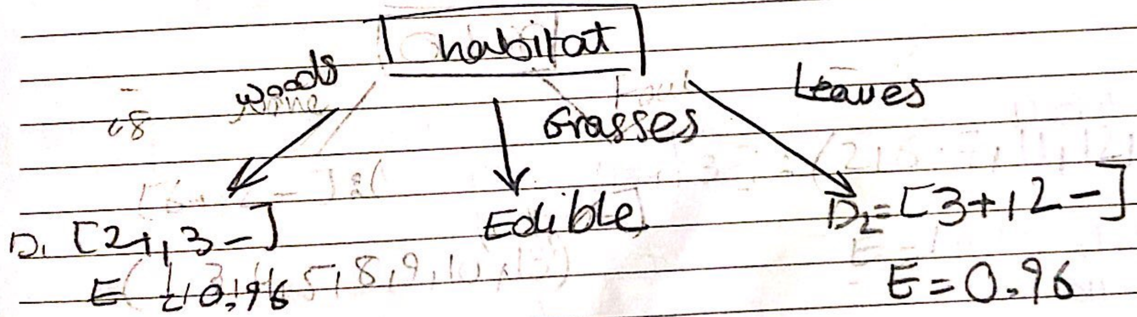
$D_{\text{Flat}}[3+,4-]: E = 0.985$ $\quad D_{\text{Conven}}[6+,1-]: E = 0.591$

↳ $\text{Gain}(D, \text{Cap Shape}) = 0.152$

$\text{Info}_{\text{odor}}(D) = \frac{8}{14} \times 0.811 + \frac{6}{14} \times 1 = 0.474$

$D_{\text{None}} = [6+,2-]: E = 0.811$ $\quad D_{\text{Foul}} = [3+,3-]: E = 1$

↳ $\text{Gain}(D, \text{odor}) = 0.0485$ ✓



$\text{Info}_{\text{Cap Color}}(D_1) = \frac{2}{5} \times 1 = 0.4$

$D_{\text{Red}} = [0,2-]: E = 0$ $\quad D_{\text{Green}} = [1+,1-]: E = 1$

$D_{\text{white}} = [1+,0]: E = 0$

↳ $\text{Gain}(D_1, \text{CapColor}) = 0.96 - 0.4 = 0.56$

$\text{Info}_{\text{cap shape}}(D_1) = 0$

$D_{\text{Flat}} = [0,3-]: E = 0$

$D_{\text{Conven}} = [2+,0]: E = 0$

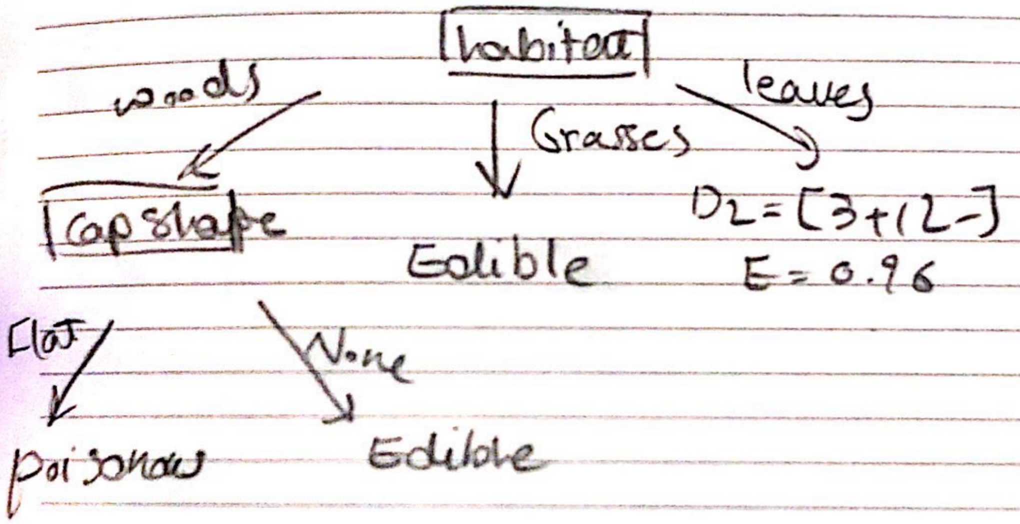
↳ $\text{Gain}(D_1, \text{Color Shape}) = 0.96$ ✓ max

$\text{Info}_{\text{odor}}(D_1) = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 1 = 0.95$

$D_{\text{None}} = [1+, 2-] : E = 0.918$

$D_{\text{Foul}} = [1+, 1-] : E = 1$

↳ $\text{Gain}(D_1, \text{odor}) = 0.96 - 0.95 = 0.01$



$\text{Info}_{\text{o}\ \text{Cap Color}}(D_2) = \frac{2}{5} \times 1 + \frac{3}{5} \times 0.918 = 0.95$

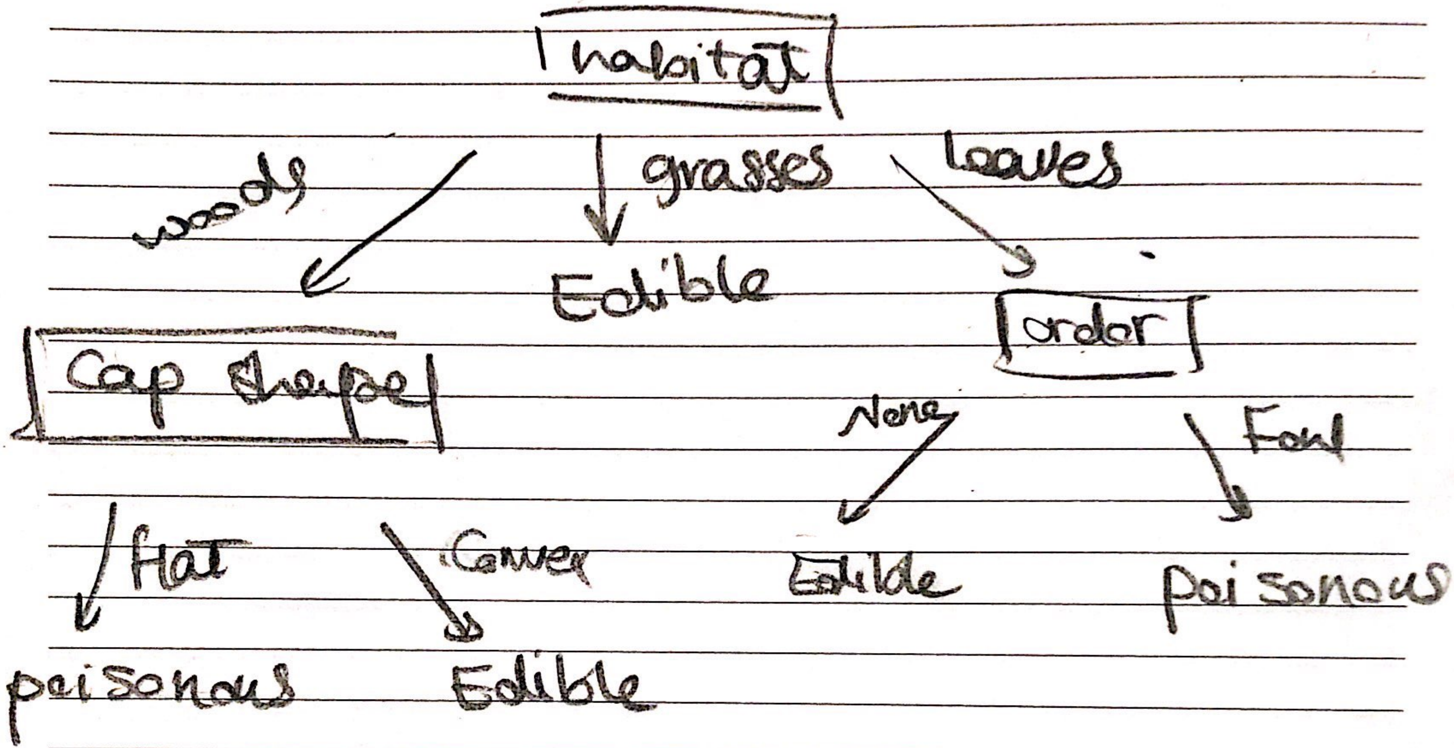
$D_{\text{white}} = [1+, 1-] : E = 1 \quad D_{\text{Green}} = [2+, 1-] : E = 0.918$

↳ $\text{Gain}(D_2, \text{Cap Color}) = 0.01$

$\text{Info}_{\text{n}}(D_2) = 0$

$D_{\text{None}} = [3+, 0) : E = 0 \quad D_{\text{Foul}} = [0, 2-] : E = 0$

$\text{Gain}(D_2, \text{Ordor}) = 0.96$


habitat
woods
grasses
leaves
Edible
ordor
cap shape
None
Foul
Edible
poisonous
Flat
Convex
poisonous
Edible

حل ۴

$$p(y|a,b,c) = \frac{p(y)}{p(a,b,c)} \times p(a|y) \times p(b|y) \times p(c|y)$$ ①

②

$p(c_i)$: $p(Y = \text{"ok"}) = \frac{5}{8}$

$p(Y = \text{"bad"}) = \frac{3}{8}$

$p(A = \text{"false"} | Y = \text{"ok"}) = \frac{4}{5}$

$p(A = \text{"false"} | Y = \text{"bad"}) = \frac{1}{3}$

$p(B = \text{"true"} | Y = \text{"ok"}) = \frac{2}{5}$

$p(B = \text{"true"} | Y = \text{"bad"}) = \frac{2}{3}$

$p(C = \text{"false"} | Y = \text{"ok"}) = \frac{4}{5}$

$p(C = \text{"false"} | Y = \text{"bad"}) = \frac{2}{3}$

$p(x|c_i)$: $p(x | Y = \text{"ok"}) = \frac{4}{5} \times \frac{2}{5} \times \frac{4}{5} = \frac{32}{125}$

$p(x|y = \text{"bad"}) = \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} = \frac{4}{27}$

0.165

$p(x|c_i) \times p(c_i)$:

$p(x | Y = \text{"ok"}) \times p(y = \text{"ok"}) = \frac{32}{125} \times \frac{5}{8} = 0.16$ ✓

$p(x|y = \text{"bad"}) \times p(y = \text{"bad"}) = \frac{4}{27} \times \frac{3}{8} = 0.05$

[illegible] case OK [illegible] ۳

posterior = liklihood x perior / evidence

evidence p(a,b,c) $= \frac{2}{8} = \frac{1}{4}$

posterior $= \frac{32}{125} \times \frac{5}{8} / \frac{1}{4} \cong 0.64$