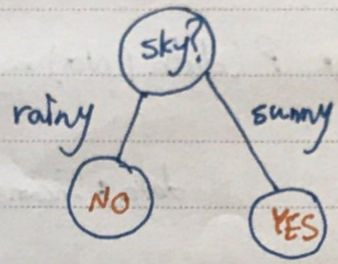
Subject
Date

سوال ۱-

a)

برای معیار خلوص از Gini Index استفاده می‌کنیم.

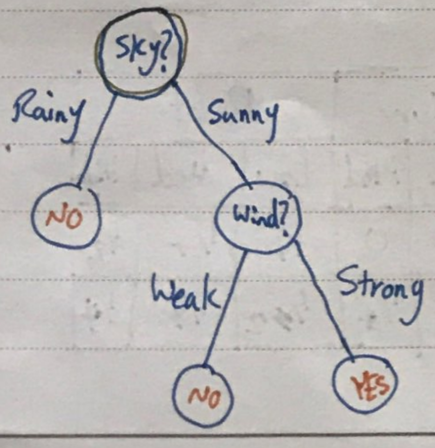
با توجه به کوچک بودن dataset مشخص است که Sky و AirTemp مقدار Gini split صفر دارند. پس این دو متغیر (feature) بهترین تصمیم برای گره‌ی اولند:



و با توجه به الگوریتم هانت، چون برگ‌ها خالص هستند، الگوریتم پایان می‌یابد.

b)

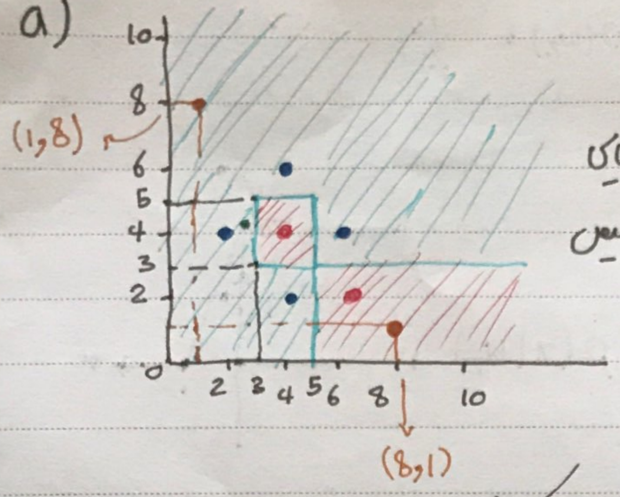
این بار هم featureهای Sky و AirTemp، Wind و Forecast باید ناخالصی بیشترین Gini را دارند.
برای مثال روی درخت را تشکیل می‌دهیم.
در مرحله‌ی بعدی می‌بینیم که اگر Sky=Sunny آنگاه feature چهارم یعنی Wind کمترین Gini split را دارد.
و در این جا درخت کامل می‌شود.



Subject
Date

سوال 2 -

a)



همانطور که در شکل مشخص است، نقطه‌هایی که در فضای هاشور آبی خورده هستند، آبی و نقطه‌هایی که در فضای هاشور خورده‌ی قرمز هستند، قرمز تشخیص داده می‌شوند.

برای محاسبه‌ی این محدوده‌ها، فاصله‌ی هر نقطه در صفحه را از نزدیک‌ترین همسایه‌اش در نظر گرفتم و محدوده‌ها را تعیین کردم.

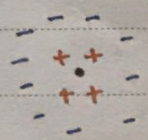
برای این فاصله‌ها هم معیار فاصله‌ی اقلیدسی را در نظر گرفتم.

b) همانطور که در شکل با همین رنگ مشخص است، نقطه‌ی (1,8) متعلق به فضای آبی است.

c) اگر مقدار k بزرگتر از 2 باشد، همه‌ی نقطه‌ها آبی تشخیص داده خواهند شد، زیرا فقط دو نقطه‌ی قرمز داریم.

اگر k را از 4 بزرگتر ($[5,\infty)$) در نظر بگیریم، آنگاه الگوریتم ما کار نخواهد کرد زیرا k نقطه‌ی نزدیک یک رنگ نخواهد یافت. مگر اینکه max کوچکتر از k را به عنوان جواب حساب کند که در آن صورت باز هم جواب آبی خواهد بود.

در حالت کلی و فارغ از این data set خاص، با بزرگ در نظر گرفتن k (نسبتاً)، همسایه‌ها ممکن است از کلاس‌های دیگر (و گاهاً "اشتباه") در نظر گرفته شوند. مثلاً:



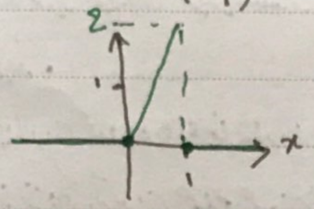
small k → +   large k → −

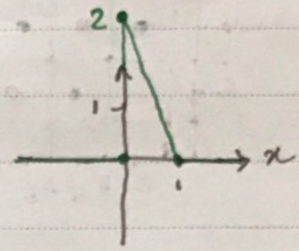
d) بله، ما برای مثال می‌توان مقدار میانگین target را برای k همسایه‌ی نزدیک محاسبه و به عنوان target نقطه‌ی جدید در نظر گرفت (یا هر تابع معقول دیگر مانند میانگین).
بنابراین از لحاظ امکان‌پذیری ما ممکن است.

سؤال 3 -

a) $P(w_1) = P(w_2) \Rightarrow P(w_1) = P(w_2) = \frac{1}{2}$

$$P(x|w_i) = \frac{P(w_i|x)\,P(x)}{P(w_i)}$$

$P(x|w_1) \rightarrow$ 

$P(x|w_2) \rightarrow$ 

b) $P(x) = P(x|w_1)P(w_1) + P(x|w_2)\cancel{w_2}P(w_2)$

$$= \frac{2x}{2} + \frac{2-2x}{2} = 1$$

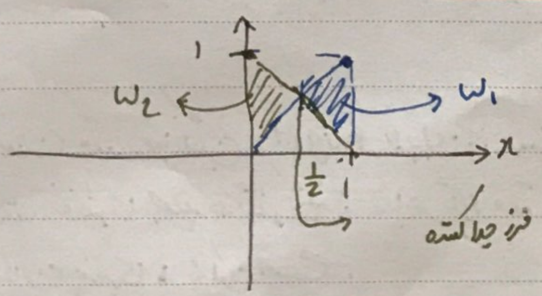
$$P(w_1|x) = \frac{P(x|w_1)P(w_1)}{P(x)} = x \quad 0 \leq x \leq 1$$

$$P(w_2|x) = \frac{P(x|w_2)P(w_2)}{P(x)} = 1-x \quad 0 \leq x \leq 1$$

بعد از به دست آوردن این متغیرها، در هر نقطه‌ی x، اگر مقدار $P(w_1|x)$ بیشتر بود، کلاس w_1 تشخیص داده می‌شود و در غیر این صورت کلاس w_2.

برای قسمت امتیازی:

$P(w_1|x) = P(w_2|x) \Rightarrow x = 1-x \Rightarrow \boxed{x = \frac{1}{2}}$



Subject
Date

سؤال 4-

a) از فرض Naive Bayes بودن استفاده می‌کنیم:

$P(Income = Low \mid D = Yes) = \frac{1}{2}$, $P(Income = Low \mid D = No) = \frac{1}{4}$

$P(Edu = MS \mid D = Yes) = \frac{1}{4}$, $P(Edu = MS \mid D = No) = \frac{3}{4}$

$P(Debt = High \mid D = Yes) = \frac{1}{2}$, $P(Debt = High \mid D = No) = \frac{1}{4}$

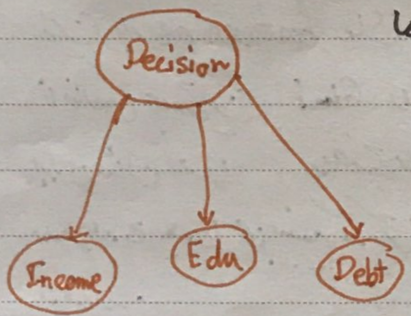
$\Rightarrow P(data \mid Yes) = \frac{1}{16}$, $P(data \mid No) = \frac{3}{64}$

$P(Yes) = P(No) = \frac{1}{2}$ data : Income = low, Edu = MS, Debt = High

$\Rightarrow P(Yes \mid data) > P(No \mid data)$

طبق محاسبات، جواب Yes است.

b) با توجه به اینکه در مدل Naive Bayes، feature ها مستقل در نظر گرفته می‌شوند:



c)

	Income			Education				Debt		
	High	Med	Low	HS	BS	MS	PhD	Low	Med	High
Decision=Yes	1/4	1/4	1/2	1/2	1/4	1/4	0	1/4	1/4	1/2
Decision=No	1/2	1/4	1/4	0	0	3/4	1/4	1/~~4~~2	1/4	1/4

Subject:
Date

سوال ۵-

a) $\dfrac{P(x|H_0)}{P(x|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} T \longrightarrow \dfrac{\frac{1}{\sqrt{2\pi\delta^2}} e^{-x^2/2\delta^2}}{\frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(x-m)^2}{2\delta^2}}} \underset{H_1}{\overset{H_0}{\gtrless}} T$

$\Rightarrow -\dfrac{x^2}{2\delta^2} + \dfrac{(x-m)^2}{2\delta^2} \underset{H_1}{\overset{H_0}{\gtrless}} \log T \Rightarrow \dfrac{m^2 - 2\delta^2 \log T}{2m} \underset{H_1}{\overset{H_0}{\gtrless}} x$

b)

$P_D = P(D_1|H_1) = \int_T^{\infty} D(x|H_1)\, dx$

$P_{FA} = P(D_1|H_0) = \int_T^{\infty} D(x|H_0)\, dx$

سوال ۶-

a)

$Pr(P3) = Pr(P3|P2)\, Pr(P2) + Pr(P3|\overline{P2})\, Pr(\overline{P2})$

$Pr(P2) = Pr(P2|P1)\, Pr(P1) + Pr(P2|\overline{P1})\, Pr(\overline{P1})$

$\Rightarrow Pr(P2) = 0{,}62$

$\Rightarrow Pr(\overline{P3}) = 1 - (0{,}2 \times 0{,}63 + 0{,}3 \times 0{,}68) = 0.67$

b)

$Pr(P2|\overline{P3}) = \dfrac{Pr(\overline{P3}|P2)\, Pr(P2)}{Pr(\overline{P3})} = 0{,}65$

PAPCO