מערכות לומדות - 046195 הפקולטה להנדסת חשמל

אביב תשע״ט (2019) הטכניון - מכון טכנולוגי לישראל

תרגיל בית 4 :

אופטימיזציה, פרספטרון, עצים

תאריך הגשה : **11.06.19**

מוריאל בן משה 304832512, דניאל אנגלסמן 300546173

פרספטרון

שאלה 1

א. הוכח כי $\left\langle w^*, w^{(t+1)} \right\rangle \geq t$. הנחיה : השתמש בטור טלקסופי עבור כל האיטרציות עד t.

ב. נסמן : $R = \max_i \|x_i\|$. הוכח כי $\left\|w^{(t+1)}\right\|^2 \leq \left\|w^{(t)}\right\|^2 + R^2$

ג. הראה כי מתקיים $\left\|w^{(t+1)}\right\|^2 \leq tR^2$

1

1.1

Let dot product satisfy $y_i\langle w^*,\ x_i\rangle \geq 1$, and the subtraction :

$$\langle w^*,\ w^{t+1}\rangle - \langle w^*,\ w^t\rangle = \langle w^*,\ w^{t+1} - w^t\rangle$$

$$\text{Hence,} \qquad \langle w^*,\ w^{t+1}\rangle = \langle w^*,\ w^{t+1} - w^t\rangle + \langle w^*,\ w^t\rangle$$

We'll take $t = [1,\ t]$ to be the number of iterations until reaching convergence :

$$\begin{aligned} \langle w^*,\ w^{t+1}\rangle &= \Big(\langle w^*,\ w^{t+1} - w^t\rangle + \langle w^*,\ w^t\rangle\Big)_{1:t} = \\ &= \Big(\langle w^*,\ w^2 - \cancel{w^1}\rangle + \langle w^*,\ \cancel{w^1}\rangle\Big)_{t=1} + \Big(\langle w^*,\ w^3 - \cancel{w^2}\rangle + \langle w^*,\ \cancel{w^2}\rangle\Big)_{t=2} + \dots \\ &\dots + \Big(\langle w^*,\ w^{t+1} - \cancel{w^t}\rangle + \langle w^*,\ \cancel{w^t}\rangle\Big)_{t=t} = \sum_{t=2}^{t+1}\langle w^*,\ w^t\rangle = \sum_{t=1}^{t}\langle w^*,\ w^{t+1}\rangle \geq \underline{\underline{\sum_{t=1}^{t} 1 = t}} \end{aligned} \tag{1.1}$$

1.2

Express the norm function such as (Recall - $x_i \in \mathbb{R}^n,\ y_i \in \mathbb{R}$):

$$\begin{aligned}
\left\|w^{t+1}\right\|^2 = \left\|w^t + y_i x_i\right\|^2 = \langle w^t + y_i x_i,\ w^t + y_i x_i\rangle &= (w^t)^T w + 2 y_i w^t x_i + (y_i x_i)^T (y_i x_i) \\
y_i^2 = \{-1,\ 1\}^2 = 1 \quad \forall\ i \quad \Rightarrow \qquad & \left\|w^t\right\|^2 + 2\langle w^t,\ y_i x_i\rangle + \|x_i\|^2 \\
y_i\langle w^t,\ x_i\rangle \le 0, \qquad \|x_i\|^2 \Leftrightarrow \max\|x_i\| \equiv R \qquad \Rightarrow \qquad & \left\|w^{t+1}\right\|^2 \le \underline{\underline{\left\|w^t\right\|^2 + R^2}}
\end{aligned} \tag{1.2}$$

1.3

Using the previous proof recursively :

$$\begin{aligned}
\left\|w^2\right\|^2 \le \left\|w^1\right\|^2 + R^2 \quad \rightarrow \quad \left\|w^3\right\|^2 &\le \left\|w^2\right\|_{w_1}^2 + R^2 \ \le\ \left\|w^1\right\|^2 + 2R^2 \\
\left\|w^4\right\|^2 &\le \left\|w^3\right\|_{w_2}^2 + R^2 \ \le\ \left\|w^1\right\|^2 + 3R^2 \\
\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\
\left\|w^t\right\|^2 &\le \left\|w^1\right\|^2 + (t-1)R^2 \\
\text{Recall,} \quad \bar{\mathbf{w}}_1 = 0 \ \Rightarrow \quad \left\|w^{t+1}\right\|^2 &\le \left\|w^t\right\|^2 + R^2 \le \cancelto{0}{\left\|w^1\right\|^2} + \underline{\underline{tR^2}}
\end{aligned} \tag{1.3}$$

ברצוננו להוכיח שהאלגוריתם מתכנס ל- w^* , כלומר שמתקיים:

$$\cos\theta_{t+1} = \frac{\left\langle w^*, w^{(t+1)}\right\rangle}{\left\|w^*\right\|_2 \left\|w^{(t+1)}\right\|_2} \xrightarrow{t\to\infty} 1$$

ד. הסבר במילים את המשמעות הגאומטרית של תנאי זה.

1.4

In the Euclidean space, dot ($\subset$ inner) product is defined as the product of their norms by the cosine angle between them - $cos\ (w^*,\ w^{t+1})\ = cos\ \theta_{t+1}$. Therefore, vectors pointing the same angle will perform similar classification due to ($y_i = \{\pm 1\}$).

(1.4)

נסמן : $B = \min\left\{\|w\| : \forall 1 \le i \le m \;\; y_i \langle w, x_i \rangle \ge 1\right\}$

ונסמן את w^* להיות הוקטור אשר משיג את המינימום של B.

ה. חבר את הסעיפים הקודמים יחדיו. מה החסם התחתון שמצאתם על $\cos\theta_{t+1}$? מהו החסם העליון הטריוויאלי על $\cos\theta_{t+1}$?

ו. הסבר איך החסמים שהוכחתם מוכיחים את התכנסות האלגוריתם. מצא חסם על מספר האיטרציות עד להתכנסות.

1.5

$$cos\ \theta_{t+1} = \frac{\langle w^*,\ w^{t+1}\rangle}{\|w^*\|_2\ \|w^{t+1}\|_2} \le \frac{t}{B\ \sqrt{tR^2}} = \frac{\sqrt{t}}{BR} \quad \Rightarrow \quad \frac{\sqrt{t}}{BR} \le \underline{\underline{cos\ \theta_{t+1}}} \le\ 1 \tag{1.5}$$

1.6

Convergence is obtained for monotonic increasing t index, such that :

$$\frac{\sqrt{t}}{BR} \le\ 1 \quad \Rightarrow \quad t \le (BR)^2 \tag{1.6}$$

2

Using the entropy criterion for all $H(S)$ and each $H(S \mid prop.)$ of the samples :

$$\hat{p}_j = \frac{1}{N}\sum_{k=1}^{N} I\{d_k \ = \ c_j\} = \frac{4}{8}$$

$$H(S) = -\sum_j \hat{p}_j \ log_2(\hat{p}_j) = -\frac{1}{2} \ log_2(\frac{1}{2}) - \frac{1}{2} \ log_2(\frac{1}{2}) = 1$$

Let us concentrate the data, ignoring the *Sometimes* option when calculating :

ID	Family Heart attack events	Male	Smokes	Exercises	Result (Heart Attack)
1	Yes	Yes	No	Yes	No
2	Yes	Yes	Yes	No	Yes
3	No	No	Yes	No	No
4	No	Yes	Sometimes	Yes	No
5	Yes	No	Yes	Yes	Yes
6	No	No	Yes	Yes	No
7	Yes	No	Sometimes	No	Yes
8	No	Yes	Yes	Yes	Yes

Index

	Yes	No	Some.
Yes			
No			
Some.			

#	History		Σ_1	Male		Σ_2	Smokes		Σ_3	Exercises		Σ_4
Yes	1	3	4	3	1	4	3	2	5	2	3	5
No	1	3	4	1	3	4	0	1	1	2	1	3
Some.	0	0	0	0	0	0	1	1	2	0	0	0

	P(j)											
#	**History**		**Σ/N_j**	**Male**		**Σ/N_j**	**Smokes**		**Σ/N_j**	**Exercises**		**Σ/N_j**
Yes	0.25	0.75	0.5	0.75	0.25	0.5	0.6	0.4	0.625	0.4	0.6	0.625
No	0.25	0.75	0.5	0.25	0.75	0.5	0	1	0.125	0.6667	0.3333	0.375
Some.	0	0	0	0	0	0	0.5	0.5	0.25	0	0	0

#	History		ΣH_1j	Male		ΣH_2j	Smokes		ΣH_3j	Exercises		ΣH_4j
Yes	0.500	0.311	0.811	0.311	0.500	0.811	0.442	0.529	0.971	0.529	0.442	0.971
No	0.500	0.311	0.811	0.500	0.311	0.811	0.000	0.000	0.000	0.390	0.528	0.918
Some.	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.500	1.000	0.000	0.000	0.000
H(S_j)	0.8113			0.8113			**0.6068**			0.9512		
H-ΔH	0.1887			0.1887			**0.3932**			0.0488		

The optimal property (= smoke) is the one that maximizes $\Delta H(S) = H(S) - H(S|A)$.

Calculating the next branch based on the smokes property :

$$\hat{p}_j = \frac{1}{N}\sum_{k=1}^{N} I\{d_k \ = \ c_j\} = \frac{3}{5}$$

$$H(S) = -\sum_j \hat{p}_j \ log_2(\hat{p}_j) = -\frac{3}{5} \ log_2(\frac{3}{5}) - \frac{2}{5} \ log_2(\frac{2}{5}) = 0.971$$

ID	Family Heart attack events	Male	Exercises	Result (Heart Attack)
2	Yes	Yes	No	Yes
3	No	No	No	No
5	Yes	No	Yes	Yes
6	No	No	Yes	No
8	No	Yes	Yes	Yes

	Yes	No
Yes		
No		

As previously seen, when person does not smoke, there's no correlation to cancer.
Therefore, we get maximal *Entropy* and no further split is needed (= criterion is reduced) :

#	History		Σ_1	Male		Σ_2	Exercises		Σ_4
Yes	2	0	2	2	0	2	2	1	3
No	1	2	3	1	2	3	1	1	2
	P(j)								
#	**History**		**Σ/N_j**	**Male**		**Σ/N_j**	**Exercises**		**Σ/N_j**
Yes	1	0	0.4	1	0	0.4	0.6667	0.3333	0.6
No	0.3333	0.6667	0.6	0.3333	0.6667	0.6	0.5	0.5	0.4
#	**History**		**ΣH_1j**	**Male**		**ΣH_2j**	**Exercises**		**ΣH_4j**
Yes	0.000	0.000	**0.000**	0.000	0.000	**0.000**	0.390	0.528	0.918
No	0.528	0.390	0.918	0.528	0.390	0.918	0.500	0.500	1.000
H(S_j)	**0.5510**			**0.5510**			0.9510		
H-ΔH	**0.4200**			**0.4200**			0.0200		

Now we can see that History *(Family heart attack)* and *Male* gained the same maximal entroy ($\Delta H(S|A)$). Both exhibit maximal entropy at the *Yes-Yes* condition, meaning that no further split is needed. Thus we'll reduce each of them and check the next branches :

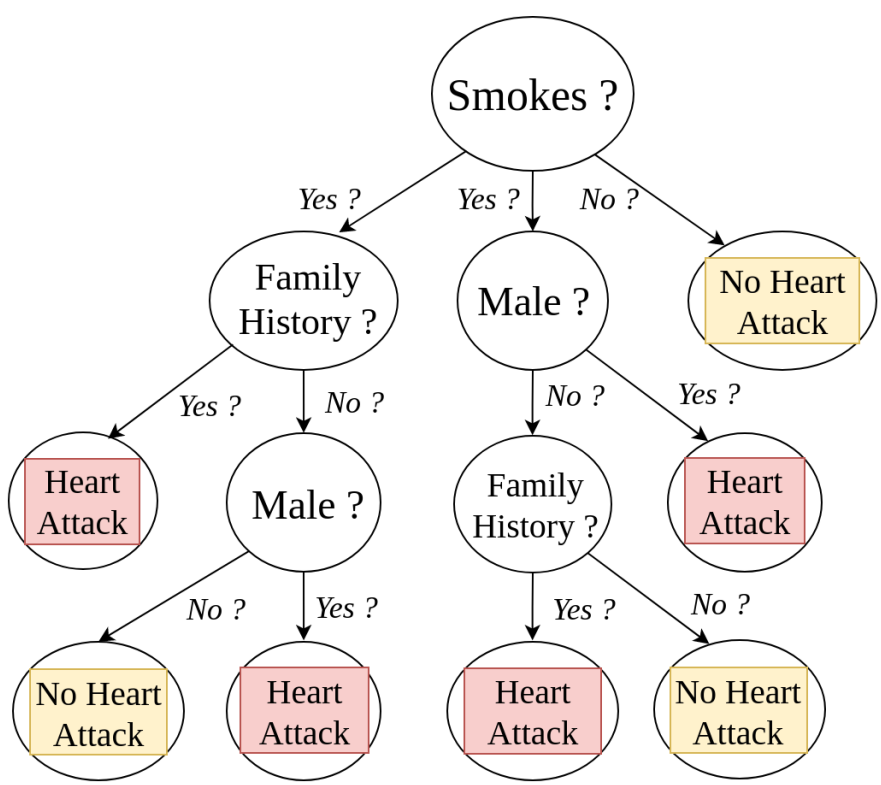
ID	Male	Exercises	Result (Heart Attack)
3	No	No	No
6	No	Yes	No
8	Yes	Yes	Yes

ID	Family Heart attack events	Exercises	Result (Heart Attack)
3	No	No	No
5	Yes	Yes	Yes
6	No	Yes	No

#	Male		Σ_2	Exercises		Σ_4
Yes	1	0	1	1	1	2
No	0	2	2	0	1	1
	P(j)					
#	**Male**		**Σ/N_j**	**Exercises**		**Σ/N_j**
Yes	1	0	0.2	0.5	0.5	0.4
No	0	1	0.4	0	1	0.2
#	**Male**		**ΣH_2j**	**Exercises**		**ΣH_4j**
Yes	0.000	0.000	**0.000**	0.500	0.000	0.500
No	0.000	0.000	**0.000**	0.000	0.000	0.000
H(S_j)	**0.0000**			0.2000		
H-ΔH	**0.9710**			0.7710		

#	History		Σ_2	Exercises		Σ_4
Yes	1	0	1	1	1	2
No	0	2	2	0	1	1
	P(j)					
#	**History**		**Σ/N_j**	**Exercises**		**Σ/N_j**
Yes	1	0	0.2	0.5	0.5	0.4
No	0	1	0.4	0	1	0.2
#	**History**		**ΣH_2j**	**Exercises**		**ΣH_4j**
Yes	0.000	0.000	**0.000**	0.500	0.500	1.000
No	0.000	0.000	**0.000**	0.000	0.000	0.000
H(S_j)	**0.0000**			0.4000		
H-ΔH	**0.9710**			0.5710		

We can see that maximal entropy is gained **either way** such that no further split is needed. Considering 2 maximal entropy cases, the decision tree will be as such :



2.2 Now we get another property (*cholesterol*) which is continuous, and its average value is ($x \sim 220$). Conveniently, it exhibits :

$$H(S|t > 220) = \frac{3}{5} \cdot \frac{2}{5} + \frac{2}{5} \cdot \frac{3}{5} = 0.48, \quad \text{and} \quad H(S|t \leq 220) = 1 \cdot (1 - 1) = \mathbf{0}$$

Which naturally perform the best impurity and thus we can get the following sorted table :

ID	Family Heart attack events	Male	Cholesterol	Blood-pressure	Result (Heart Attack)
1	Yes	Yes	160	High	No
4	No	Yes	170	Normal	No
6	No	Yes	215	Normal	No
5	Yes	No	230	High	Yes
7	Yes	No	235	Normal	No
8	No	Yes	240	High	Yes
3	No	No	245	Normal	No
2	Yes	Yes	260	Normal	Yes

Elaborating only the $x > 220$ condition, we'll divide the data and get 5 options :

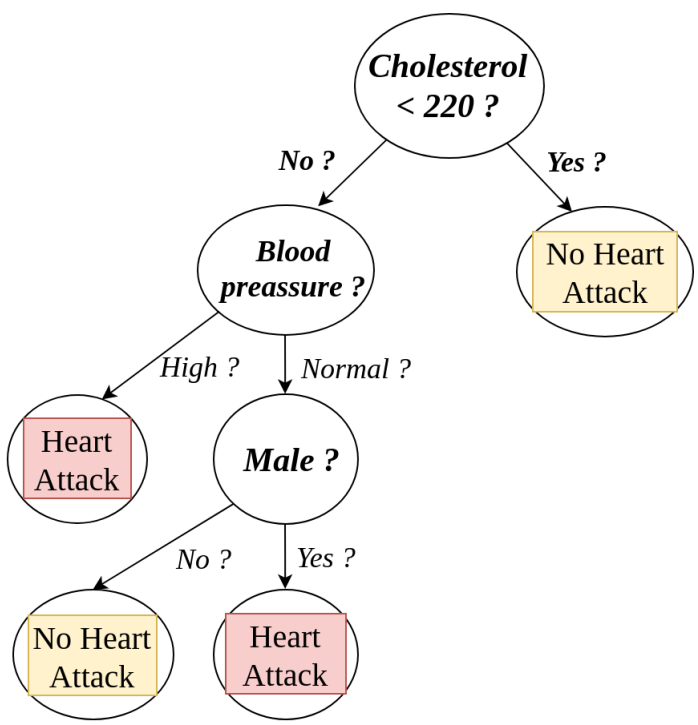
#	History		Σ_1	Male		Σ_2	Blood Pres.		Σ_4
Yes	2	1	3	2	0	2	2	0	2
No	1	1	2	1	2	3	1	2	3
	P(j)								
#	**History**		**Σ/N_j**	**Male**		**Σ/N_j**	**Blood Pres.**		**Σ/N_j**
Yes	0.6667	0.3333	0.6	1	0	0.4	1	0	0.4
No	0.5	0.5	0.4	0.3333	0.6667	0.6	0.3333	0.6667	0.6
#	**History**		**ΣH_1j**	**Male**		**ΣH_2j**	**Blood Pres.**		**ΣH_4j**
Yes	0.222	0.222	0.444	0.000	0.000	**0.000**	0.000	0.000	**0.000**
No	0.250	0.250	0.500	0.222	0.222	0.444	0.222	0.222	0.444
Q(S_j)	0.4667			**0.2667**			**0.2667**		

Once again we get an optimal **Gini Index** value for 2 criteria (*Male* & *Blood Pressure*). We'll focus this time only on Blood Pressure, so after division and sorting :

ID	Family Heart attack events	Male	Result (Heart Attack)
7	Yes	No	No
3	No	No	No
2	Yes	Yes	Yes

#	History		Σ_1	Male		Σ_2
Yes	1	1	2	1	0	1
No	0	1	1	0	2	2
	P(j)					
#	**History**		**Σ/N_j**	**Male**		**Σ/N_j**
Yes	0.5	0.5	0.6667	1	0	0.3333
No	0	1	0.2	0	1	0.6667
#	**History**		**ΣH_1j**	**Male**		**ΣH_2j**
Yes	0.250	0.250	0.500	0.000	0.000	**0.000**
No	0.000	0.000	0.000	0.000	0.000	**0.000**
Q(S_j)	0.3333			**0.0000**		

We can see that the *History* criterion is irrelevant since *Male* condition obtains optimallity :



שאלה 3

א. ציין האם הקביעות הבאות נכונות או לא והסבר בקצרה מדוע: שני עצים שונים שמשרים תיוג זהה ובעל שגיאה אפס על מדגם הלימוד ייסוגו כל קלט בצורה זהה

Not True : Similar classification might be caused incidentally due to the **given data**. Further data (beyond our dataset) may result in a different calculation and go through different nodes, resulting in a different classification.

ב. הנח בעיית למידה עבורה נתונים מאפיינים רועשים רבים (כלומר מאפיינים בעלי קורלציה נמוכה לתיוג). איזה אלגוריתם עדיף במקרה זה: עץ החלטה או אלגוריתם שכן קרוב (NN-1)?

Decision Tree : As seen in last question, a tree may be robust to noisy data once an optimal result is obtained, the branch is "cut" and no further splits / calculation will be needed. Contrarily, the 1-NN algorithm calculates each item in the dataset, resulting in lower accuracy and vulnerability to bias.

ג. הנח בעיית סיווג עם סט אימון מעל R^{100}. סטודנט א' אימן עץ החלטה (כפי שנלמד בכיתה). סטודנט ב' קודם נירמל את קבוצת האימון כך שהממוצע של כל קורדינטה הוא 0 וסטיית התקן היא 1, ולאחר מכן אימן עץ בדיוק כמו סטודנט א'. מי מהמשפטים הבאים נכון:

1. שני הסטודנטים יקבלו בהכרח את אותו העץ
2. שני הסטודנטים יקבלו בהכרח עץ שונה
3. שני הסטודנטים יקבלו לעיתים עצים זהים ולעיתים עצים שונים

Same Tree : Let $S = \{x_i,\ y_i\}$ be a set of classified samples, and probability function be : $\hat{p}_j = \frac{1}{N}\sum_{k=1}^{N} I\{d_k\ =\ c_j\}$. One can tell that the indicator function is insensitive to a normalization.

$-fin-$