שאלה 1 - סעיף 1

סדרת משתנים מקריים $Y_1, Y_2, \ldots, Y_n$ נקראת **שרשרת מרקוב**
אם לכל $n \ge t \ge 1$ מתקיים $Pr[Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_1] = Pr[Y_t | Y_{t-1}]$
כלומר בזמן t האיבר הנוכחי בסדרה, ההסתברות a האיבר
הנוכחי תלוי-תלוי במשתנים הקודמים. במילים אחרות, ההסתברות
שהאיבר הנוכחי יהיה שווה לאות כלשהי, בזמן t האיברים
הקודמים בסדרה, **שווה** להסתברות שהאיבר הנוכחי יהיה
שווה רק אם הנמצא בזמן t-1 האיבר הקודם בסדרה ככה.

הנחת **ההומוגניות בזמן** לגבי שרשרת מרקוב היא
שההסתברות *הנחה חזקה מאוד.. $Pr[Y_t | Y_{t-1}]$ אינה תלויה באינדקס t.
כלומר ההסתברות למעבר כלשהו $y \to y'$ אינה תלויה
במיקום a המעבר בסדרה (האם לכיוון ההתחלה/לסוף וכו').

כאשר אנו מתמקדים רק בשימוש HMM, אנחנו מניחים שההנחה
הזו, שמס' הפרמטרים שצריך ללמוד עבור מעברים מ-P_oS אחד
לאחר הוא כמו מס' הלוגית $|P_oS|^2$, אבל אם היינו
מוותרים על ההנחה הזו היינו צריכים ללמוד את ההסתברות
למעבר בנפרד לכל מיקום בסדרה (ונוצר עוד דברים היו
נדרשים), כמו למשל את ב-inference היינו מבצעים סדרה ארוכה
יותר מכל ההסתברויות שראינו במהלך האימון, לא היינו יודעים את
ההסתברויות למעבר באינדקסים הגבוהים), ...

הערות הבודק

ומה אנחנו מפסידים בגלל ההנחה? -3

שאלה 1 - סעיף 1 [המשך...]

מה שאנחנו מפספסים מהמודל הזה זה שכל המילים הינן תלויות במיקום, והסתברות למעבר תלויה באמצע משפט או בסוף אינה זהה.

וכן, אנחנו מפספסים את המודל ב-HMM כמו למשל של POS היה שינוי מרקוב (ראינו בכיתה שגיא לא תמיד כמו, לא שינוי לאיזה סדר מספיק את המודל - 2-ה קודמים, 3, וכו') אבל לא לגבי אינה המודל היחיד שאנחנו מניחים שאינו תלוי במיקום.

* הערה: בסעיף שאלה אם יותר הרגלים המודל הסטציונרי במובן הבא:
לכל i,j ולכל t,t' מתקיים

$$P_r[Y_t = i \mid Y_{t-1} = j] = P_r[Y_{t'} = i \mid Y_{t'-1} = j]$$

(כאשר i,j הם אינדקסים של משתנה Y עבור לייבלים, ו-t,t' הם זמנים...)

שאלה 1 - סעיף ב

בהינתן משפט $X_1,...,X_n$ (מילות קלט) אנו רוצים למצוא את ה- PoS המתאימים ביותר למשפט.

נניח שקבוצת ה- PoS היא S, אז אנו מחפשים את

הקבוע כי $P_r[X_1,...,X_n]$ לא תלוי ב- $y_1,...,y_n$

$$\underset{y_1,...,y_n\in S^t}{\operatorname{argmax}} P_r[y_1,...,y_n \mid X_1,...,X_n] = \underset{y_1,...,y_n\in S^n}{\operatorname{argmax}} P_r[y_1,...,y_n,X_1,...,X_n]$$

אנו נגדיר לכל $t\in\{1,...,n\}$ ולכל $i\in S$ את

$$\pi(t,i)=\max_{\substack{y_1,...,y_t\in S^t\\ y_t=i}} P_r[y_1,...,y_t,X_1,...,X_t]$$

~~[illegible]~~ הניסיון (כלומר זה חיפוש בגודל אקספוננציאלי) בגודלו
ולכן אנו משתמשים בהגדרה רקורסיבית

$$\pi(t,i)=\max_{j\in S}\left(\pi(t-1,j)\cdot\tau(j,i)\cdot\varepsilon(X_t,i)\right)$$

$$\pi(1,i)=\tau(START,i)\cdot\varepsilon(X_1,i)$$

כאשר τ היא הסתברות המעבר $\tau(j,i)=P_r[Y_t=i \mid Y_{t-1}=j]$

ו- ε היא הסתברות הפליטה שלו $\varepsilon(X,i)=P_r[X_t=x \mid Y_t=i]$

(הוספנו PoS ראשוני y_0 - START בהסתברות 1).

מחשבים את כל $\alpha(\cdot,\cdot)$ מ- $t=1$ ועד $t=n$, ולכן [illegible] [illegible]
בסוף - מחזירים [illegible] back-pointers [illegible] argmax [illegible]

הערות הבודק

מצוין!

נא לא לכתוב בשוליים

שאלה 1 - סעיף 3

נשים לב שמתקיים:

$$Pr[y_n=j \mid X_1,\dots,X_n] = \frac{Pr[y_n=j, X_1,\dots,X_n]}{Pr[X_1,\dots,X_n]}$$ ✓

ראינו בהרצאה איך לחשב את $Pr[X_1,\dots,X_n]$ בזמן פולינומי, ונזכיר:

[illegible]

$$Pr[X_1,\dots,X_n] = \sum_{y_1,\dots,y_n} Pr[X_1,\dots,X_n,y_1,\dots,y_n]$$

$$= \sum_{y_1,\dots,y_n} \prod_{t=1}^{n} Pr[Y_t=y_t \mid Y_{t-1}=y_{t-1}] \cdot Pr[X_t=x_t \mid Y_t=y_t]$$

נגדיר לכל $t\in[n]$ ולכל $i\in S$

$$\alpha(t,i) = \sum_{\substack{y_1,\dots,y_t \\ y_t=i}} Pr[X_1,\dots,X_t,y_1,\dots,y_t]$$ ✓

וניתן לחשב בצורה רקורסיבית ע"י:

$$\alpha(t,i) = \sum_{j\in S} Pr[Y_t=i \mid Y_{t-1}=j] \cdot Pr[X_t \mid y_t=i] \cdot \alpha(t-1,j)$$

$$\alpha(1,i) = Pr[Y_1=i \mid Y_0=START] \cdot Pr[X_1=x_1 \mid Y_1=i]$$ ✓

~~[illegible]~~ וכך, [illegible] בתכנון דינאמי [illegible] [illegible] עבור כל

~~[illegible]~~ (החל מ- $t=1$ עד n). כעת נשים לב

$\sum_{i\in S}\alpha(n,i)$ [illegible] [illegible] ~~[illegible]~~ [illegible] [illegible]

[illegible] [illegible] על-ידי [illegible] [illegible] כלל.

~~[illegible]~~ [illegible] [illegible] [illegible] ~~[illegible]~~ [illegible] $Pr[y_n=j \mid X_1,\dots,X_n]$

~~$\sum_{y_1,\dots,y_{n-1}} Pr[$~~ ~~[illegible]~~

שאלה 1 - סעיף 3 [המשך...]

נרצה לחשב את $\Pr[Y_n=j, X_1,\dots,X_n]$

ובגלל שאין ניתן לסכום על-פני S גורמים באורך $n-1$ של y-ים שונים אקספוננציאלי -

$$\sum_{y_1,\dots,y_{n-1}} \Pr[Y_1=y_1,\dots,Y_{n-1}=y_{n-1},Y_n=j, X_1,\dots,X_n]$$

ונשתמש בחישוב את כל כבר בכלל אלגוריתם קודם, לא

בקוד זה שווה ל- $\alpha(n,j)$

ולכן התשובה בסוף היא
$$\Pr[Y_n=j|X_1,\dots,X_n]=\frac{\alpha(n,j)}{\sum_{i\in S}\alpha(n,i)}$$

מצוין!

הערות הבודק

שאלה 2 - סעיף 1

שכבת Dropout מאפסת אחוז מסוים של נוירונים בקלט שלה.
היא מאופיינת ע"י היפר פרמטר $p \in (0,1]$ שמסמל את
הסיכוי של כל נוירון להתאפס. לכן נסמן $Dropout_p$

עבור וקטור $X \in \mathbb{R}^n$, התוצאה של הפעלת $Dropout_p(x)$
בשלב האימון היא וקטור $\tilde{X}$ כאשר לכל $1 \le i \le n$
מתקיים שהקואורדינטה $\tilde{X}_i$ שווה ל-0 בהסתברות p ו-X_i בהסתברות
$(1-p)$.

בשלב ה-test התוצאה של $Dropout_p(x)$ היא $p \cdot x$

* הערה: כלל יותר מקובל הוא Inverted Dropout
ואז בשלב האימון הפלט הוא $1/p \cdot \tilde{X}$
(כאשר $\tilde{X}$ מוגדר כמו מקודם) ובשלב ה-test הפלט
הוא פשוט X (כלומר אין צורך לעשות scaling)

✓

שאלה 2 - סעיף ב

ה- loss נקרא cross-entropy loss

עבור דוגמא (=תמונה) x ו-y מוגדר ע"י

$$f(x,y) = -\log\left(\frac{e^{f(x)_y}}{\sum_c e^{f(x)_c}}\right)$$

כאשר f(x) מסמן את ~~מצב~~ הפלט של הרשת על התמונה x
שלה גבוה ונכון בסוף |C| (כאשר C זו קבוצת הקלאסים האפשריים)
ואז יש יותר ביטוי ה- log בגלל המנה בהסתברויות שאנו
נותנים לתיוג הנכון בתמונה הפלט שאנו קיבלנו
(המנה המקוצרת ועל יותר להסתברות כלשהו הלה נקראת softmax)

~~[illegible]~~ מה שלמדנו בהתחלת אימון לקורס,
האם הרשת של מתנהגת נכון שהוא לא מסתבכת ויוצר המ"ב נבה
בהסתברות אחידה.

לכן, גם אנחנו מוסיפים למשק את ה-cross-entropy loss לא למלא אותה,
אנו נקבל ערך יותר רק הנותן בהסתברות הנכונה שאנו
ההסתברות שיש את הפלט של ~~זה~~ הקלאס הנכון.
לכן, ההסתברות שהאלגוריתם יתכנס אליו ~~[illegible]~~ היא ההסתברות
האחידה של כל הקלאסים, כלומר p(y|x) תהיה
ההסתברות האחידה על פני הקבוצה הקלאסים C.

X -11

הערות הבודק

ה cross entropy ממקסם כאשר ההסתברות על התיוג הנכון היא 100% ולא תלוי כמה מתפלגות בין שאר התיוגים.

שאלה 6 - סעיף 3

לקרוא לפונקציה tf.add(...) ~~בלולאה~~ בלולאה יוצר אופרטורים
חדשים בגרף, שגורם שיש קודקודים ומורכבות שאינה הכרחית בגרף.

בגלל אי יעילות session ניתן לראות fetch, ולבצע
את גרף של האופרטור הללו.

הבעיה כשקוראים לפונקציה tf.add בתוך לולאה (שנוצר)
למשל בתוך session בגרף, למשל sess.run(tf.add(x,y),...)
כאשר x ו- y שני קודקודים בגרף האופרטורים,
היא שבגרף נוצרים אופרטורים חדשים יוצר n פעמים
כאשר n לא מס׳ בגבולות הזיכרון.
אם n גדול מאוד, למשל מס׳ גדול בזיכרון האימון,
הגרף "יתפוצץ" ויכול לגרום בזבוז זיכרון.

הפתרון הוא קודם להגדיר את האופרטור כמשתנה מחוץ (לפני)
הלולאה (כמו add_op = tf.add(...)) ובתוך הלולאה
להריץ את הערך של add_op ע"י sess.run(add_op,...)

שאלה 3 - סעיף 1

אלגוריתם ה- kmeans:

```
kmeans(X, k):
    μ_1, ..., μ_k ← init-centroids(X, k)
    while True:
        μ'_1, ..., μ'_k ← copy(μ_1, ..., μ_k)
        c⃗ ← assign(X, μ_1, ..., μ_k)
        μ_1, ..., μ_k ← re-calculate-centroids(X, c⃗)
        if ∀ 1≤i≤k  μ_i = μ'_i :
            break
    return c⃗ (or μ_1, ..., μ_k, or both - why not...)
```

נסמן את פונקציית המחיר ב- C, והיא מוגדרת ע"י

$$C(X, \mu_1, \dots, \mu_k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|_2^2$$

שימו לב סכום המרחקים (בריבוע) a 6 נק' [illegible]

[illegible]

שאלה 3 - סעיף 1 [המשך...]

נראה שפונקציית המטרה יורדת ממש בכל איטרציה של האלגוריתם (אך היא לא יורדת ממש לכל אחד מהקלאסטרים לא בהכרח וכפולות (שאלה 3 ג).

נשים לב שהאלגוריתם עוצר אחרי הפעם הראשונה שלא משתנו (או לא משתנים הקלאסטרים ולא משתנה) ולכן בכל איטרציה של האלג' הקלאסטרים השתנו בהכרח.

וכן, בשלב הראשון מסויים של כל קלאסטר ~~[illegible]~~ של קלאסטר שנעבור על בפעם הראשונה לקבל את ~~[illegible]~~ ביותר,

כלומר הפונקציה $assign(X, \mu_1, \ldots, \mu_k)$ משייכת ולא בדיוק מס' נק' לקלאסטר, כאשר כל נקודה יש את האינדקס לקלאסטר המשויך לה הנ"ל הקלאסטר משויך לה

$$i = \arg\min_{1 \le i \le k} \|X_j - \mu_i\|$$

ולכן אם כל נקודה בסיום של X_j

בהגדרה מפספסת ... נקבל μ לכל אחד, ש- $\|X_j - \mu_i\| \le \|X_j - \mu_l\|$

ולכן התרומה של X_j לפונקציה רק קטנה.

כלומר, בשלב ה-assign בכל שלב רק יכולה לרדת.

כעת, בשלב השני מעדכנים את הקלאסטרים, מאחר שאנו מחשבים $\mu_i = \sum_{x_j \in C_i} X_j$

μ_i הוא בדיוק הממוצע של הנק' שבקלאסטר הזה

וה-μ_i הזה הוא ממזער את סכום המרחקים

עבור הנקודות של הקלאסטר ...

(ניתן להראות זאת בגזירה - ~~[illegible]~~ μ_i והשוואה לאפס - $\mu_i = \sum_{x_j \in C_i} X_j$ אם אותו זמן..)

-7

הערות הבודק

שאלה 3 - סעיף 1 [המשך...]

ולכן הפונ' הכוללת בהכרח לא רק רק יחידה עבורה.
בסך הכל קיבלנו שלכל אי-זוגי n הצלחנו הפונ' מתאימה
מילה.

~~[illegible]~~

נשים לב אמנם הקלטים באורך זוגי הם n^k
(כאשר n לא מתח' כל (כאשר) בכלל של קלט שאורכו באורך
הגדלה לוגיקה ב- $\{1,\ldots,k\}^n$ שמגדיר את הקבוצה של נ'.

כעת, נניח (בשלילה) שהפונקציה איננה מלאה. לכן אומר שיש
קלטים שאלו לא עצמן כמעט (עיקרון שובך היונים)
ובאורך-זוגי המתאימה יוגדר הפונק' היא לכל אורך מספר
שמכיל את באי-זוגי המאותר יוגדר, ולוגי' כמובן ~~סתירה~~
כי בפונק' הגדלה כגודל קלטים לכל n.

שאלה 3 - סעיף ב

בשלב האחרון מגיעים של K-means ויוצג הראשוני K גו״ע משמעותיים יותר

הקיים בימד. לכן לאחר הטווח a כל הדבר מורכב מ-K גו״ע בלבד.

σ ע״ע מגיעים לרכיב קשירות כגודל שמדובר ע״י הנקודות בדאטה.

אין לנו דרך לקשירות האמיתי הוא K, אך פרמטרים גם K-1, ..., מימד

הו״ע יגאיוני לרכיבי-קשירות שווים, [illegible]

1	0	1
1	0	1
1	0	1
1	1	0
1	1	0
1	1	0
1	0	0
1	0	0
1	0	0

למשל אם יש 9 נק׳ ובגרף יש 3 רכיבי-קשירות, אלגוריתם (לפי נרמול) הוקטורים העצמיים

ואז אם מגיעים לקלסטרינג ל-k=2 אז הנקודות של 2 הקבוצות באותה, ובא

שם ניתן לקלסטר ל-2 באותו מקום (שורות 4,5,8) בכדי אחד ומתרכז בהן)

לכן הנקודות לא הגיעו מתמטית אין ל-3 כמו בגרף ה-elbow method

ולכן ה-elbow כאן לא כל-כך אינפורמטיבי.

1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	1	0
1	0	1	0
1	1	0	0
1	1	0	0

ניתן לראות גם על דוגמא עם 8 נק׳ ו-4 רכיבי קשירות

ובו ראה קלסטר ל-2 או 3 קלסטרים יניב אותו 0

משמעות ולכן לא יגיע ״ג׳אמפ״ במעבר מ-3 ל-4...

שאלה 3 - סעיף 3

כיוון שאם רכיבי הקשירות שווים לריבוי a הע״ע 0 לפחות, ככה לכן

ההגדרה שבו הגודל כאמור, רכיבי קשירות נפרדים (הוכחה)

אם הריבוי a הע״ע 0 הוא n (סומן σ הע״ע גם σ)

אז לכן אומר שיש n רכיבי קשירות נפרדים, סומן

σ נק׳ מסוגל קלסטר בהינתן עצמה. לכן הגרף אומר שפרמטר σ

בגרף הגאוסיאני היה קטן מידי, ולכן אין נק׳ דומות לה

לה.

כמו כן, שגדול אך הם לא יגיעו רכיבי קשירות נפרדים לגמרי כי לגבי גרף

גאוסיאני שניתן ערכים חיוביים תמיד, אבל גם נורא קטנים ולכן המבדל נשאר

קטנה - גרף σ a סגת הנקודות הגאוסיאני קטן מידי.

הערות הבודק

8

1	11 11
2	11 0 11
3	10 — 11