אוניברסיטת ת"א, ביה"ס להנדסת חשמל, למידת מכונה סטטיסטית

תרגיל בית 8

תרגיל בית זה עוסק במודלים מרקוביים חבויים.
התרגיל מורכב משני חלקים – חלק א' תיאורטי וחלק ב' שהינו תרגיל מחשב. יש להגיש כל חלק בנפרד, כל עבודה תיבדק בנפרד.

הגשה: עליכם להגיש קובץ **zip** שמכיל **PDF** עבור התרגיל התיאורטי, וקובץ **py/.m**. עבור תרגיל המחשב. חובה לציין מספר ת.ז. בקבצי ההגשה (בקוד – בהערה בתחילתו).
תזכורת: מי שמגיש בזוג, יש להגיש פעם אחת בלבד תרגיל תיאורטי ופעם אחת תרגיל מחשב (ניתן לערבב זוגות).

חלק א' - שאלות תיאורטיות:

שאלה 1

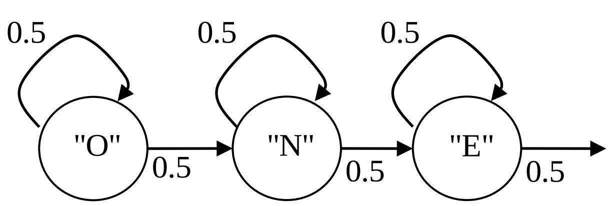
השאלה עוסקת במודל HMM עם פילוג מוצא בדיד, שהוא דומה למודל ה- HMM עם פילוג מוצא גאוסי רציף שלמדנו בכיתה, אלא שסדרת המוצא $\mathbf{X}=\left(X_1,X_2,\ldots,X_n\right)$ היא סדרה בינארית ופילוג המוצא הוא

$$P\left(X_t=r\middle|S_t=j\right)=\begin{cases} q_j & \text{if } r=1 \\ 1-q_j & \text{if } r=0 \end{cases}$$

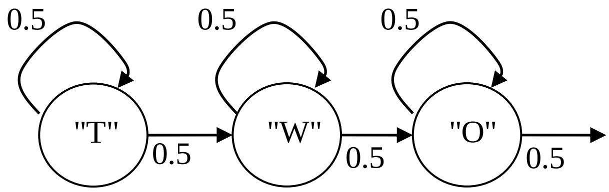
עבור $0\le q_j\le 1$. רשמו אלגוריתם EM לשערוך פרמטרי המודל מתוך L סדרות אימון, $\left\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L\right\}$. פרמטרי המודל כוללים את פילוג ההסתברות של המצב הראשוני, את מטריצת הסתברויות המעברים ממצב למצב, ואת הסתברויות המוצא.

שאלה 2

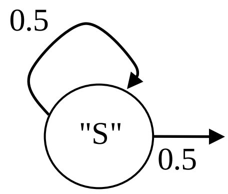
השאלה עוסקת באימון מערכת זיהוי לרצף ספרות. נניח לשם פשטות שמעוניינים לזהות הגיות של רצפים של הספרות "one" ו- "two" בלבד. עוד נניח שהספרה "one" מיוצגת ע"י left-to-right HMM בעל 3 מצבים כמתואר בציור:



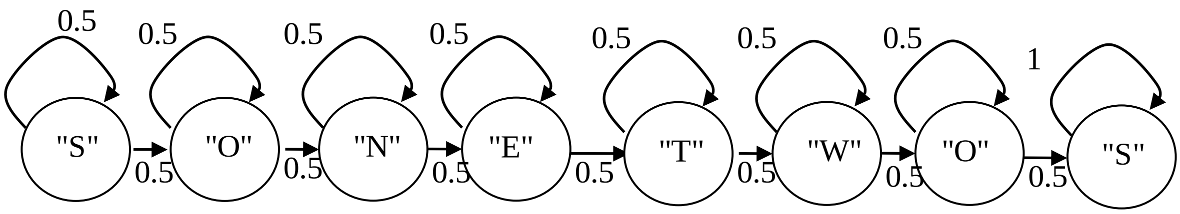
באופן דומה, הספרה "two" מיוצגת ע״י left-to-right HMM בעל 3 מצבים הבא:



נשתמש גם במודל הבא עבור שקט בתחילת מילה:



עבור שקט בסוף מילה משתמשים במצב דומה (אותה הסתברות מוצא) אבל ההסתברות להישאר במצב השקט בסוף מילה היא 1, ז"א שלא ניתן להשתחרר ממנו למצב אחר. כל רצף ספרות ממודל כדלקמן. המודל לרצף מתחיל ומסתיים במצב שקט (מצב "S"), ז"א אנחנו מניחים שכל הגיה של רצף ספרות מתחילה ומסתיימת בשקט. מצב השקט הראשוני צמוד למודל ה- HMM לספרה הראשונה ברצף, לאחריה הספרה השנייה וכן הלאה. המצב האחרון בספרה האחרונה ברצף צמוד למצב השקט המסיים את הרצף. לדוגמא, המודל שלנו לרצף "one two" הוא:



פילוג המוצא של כל מצב הוא גאוסי עם וקטור תוחלת ומטריצת שונות לא ידועים שיש לשערך מבסיס נתונים לאימון. האימון מתבצע תוך שימוש באלגוריתם EM לחישוב מקורב של משערך הסבירות המירבית. סה"כ יש לשערך ששה וקטורי תוחלת ושש מטריצות קוואריאנס עבור ששת המצבים "S","W","T","E","N","O" (בשאלה הזאת הסתברויות המצב ההתחלתי וכן הסתברויות המעבר ידועות ואין צורך לשערכן). הניחו שבסיס הנתונים לאימון כולל L_1 הגיות של הרצף "one two" ו- L_2 הגיות של הרצף "two one".

רשמו את נוסחת האיטרציה של אלגוריתם EM לאימון הפרמטרים.

רמז: אפשר להסתכל על בעיית השערוך כאילו לפנינו שתי מילים מחוברות: האחת היא "one two" שממנה יש לנו L_1 הגיות והשניה היא "two one" שממנה יש לנו L_2 הגיות. בניגוד למקרה הרגיל שלמדנו, עכשיו יש לנו אילוצי שיתוף (tying) בין הפרמטרים של שתי המילים המחוברות הללו.

<u>שאלה 3</u>

נתבונן על נוסחת השערוך של אלגוריתם ה- EM עבור הסתברויות המעברים,

$\hat{A}(k,j) \quad \left(k=1,2,\ldots,M; \quad j=1,2,\ldots,M\right)$, (M הוא מספר המצבים השונים במודל) במודל ה-

HMM שלמדנו בכיתה. לשם פשטות נניח סדרת אימון אחת ($L=1$). על פי מה שלמדנו:

$$\hat{A}(k,j)=\frac{\sum_{t=2}^{T}P\left(S_{t-1}=k,S_t=j\middle|\mathbf{x},\boldsymbol{\theta}_0\right)}{\sum_{l=1}^{M}\sum_{t=2}^{T}P\left(S_{t-1}=k,S_t=l\middle|\mathbf{x},\boldsymbol{\theta}_0\right)}=\frac{\sum_{t=2}^{T}\alpha_{t-1}(k)\beta_t(j)A_{\boldsymbol{\theta}_0}(k,j)f\left(x_t\middle|S_t=j,\boldsymbol{\theta}_0\right)}{\sum_{l=1}^{M}\sum_{t=2}^{T}\alpha_{t-1}(k)\beta_t(l)A_{\boldsymbol{\theta}_0}(k,l)f\left(x_t\middle|S_t=l,\boldsymbol{\theta}_0\right)}$$

כאשר את $\alpha_t(j)$ ואת $\beta_t(j)$ מחשבים באמצעות הרקורסיות הבאות,

$$\alpha_1(j)=\lambda_j f\left(x_1\middle|S_1=j,\boldsymbol{\theta}_0\right) \qquad j=1,\ldots,M$$
for $t=2,\ldots,T$
$$\alpha_t(j)=\sum_{k=1}^{M}\alpha_{t-1}(k)A_{\boldsymbol{\theta}_0}(k,j)f\left(x_t\middle|S_t=j,\boldsymbol{\theta}_0\right) \qquad j=1,\ldots,M$$
end

$$\beta_T(j)=1 \qquad j=1,\ldots,M$$
for $t=T-1,\ldots,1$
$$\beta_t(j)=\sum_{k=1}^{M}\beta_{t+1}(k)A_{\boldsymbol{\theta}_0}(j,k)f\left(x_{t+1}\middle|S_{t+1}=k,\boldsymbol{\theta}_0\right) \qquad j=1,\ldots,M$$
end

כאלטרנטיבה לחשבון לוגריתמי ניתן לחשב ערכים מנורמלים של $\alpha_t(j)$ ו- $\beta_t(j)$, שנסמן

בתור $\hat{\alpha}_t(j)$ ו- $\hat{\beta}_t(j)$ כדלקמן,

$$\tilde{\alpha}_1(j)=\lambda_j f\left(x_1\middle|S_1=j,\boldsymbol{\theta}_0\right) \qquad j=1,\ldots,M$$
$$\hat{\alpha}_1(j)=\tilde{\alpha}_1(j)/\sum_{l=1}^{M}\tilde{\alpha}_1(l) \qquad j=1,\ldots,M$$
for $t=2,\ldots,T$
$$\tilde{\alpha}_t(j)=\sum_{k=1}^{M}\hat{\alpha}_{t-1}(k)A_{\boldsymbol{\theta}_0}(k,j)f\left(x_t\middle|S_t=j,\boldsymbol{\theta}_0\right) \qquad j=1,\ldots,M$$
$$\hat{\alpha}_t(j)=\tilde{\alpha}_t(j)/\sum_{l=1}^{M}\tilde{\alpha}_t(l) \qquad j=1,\ldots,M$$
end

את $\hat{\beta}_t(j)$ מחשבים באופן דומה, אבל משתמשים באותו גורם הכיול $1/\sum_{l=1}^{M}\tilde{\alpha}_t(l)$ שבו השתמשנו

בחישוב $\hat{\alpha}_t(j)$:

$$\hat{\beta}_T(j) = 1 / \sum_{l=1}^{M} \tilde{\alpha}_T(l) \qquad j = 1, \ldots, M$$

for $t = T-1, \ldots, 1$

$$\hat{\beta}_t(j) = \sum_{k=1}^{M} \hat{\beta}_{t+1}(k) A_{\boldsymbol{\theta}_0}(j,k) f\left(x_{t+1} \middle| S_{t+1} = k, \boldsymbol{\theta}_0\right) / \sum_{l=1}^{M} \tilde{\alpha}_t(l) \qquad j = 1, \ldots, M$$

end

בטאו את נוסחת השערוך $\hat{A}(k,j)$ באמצעות $\hat{\alpha}_t(j)$ ו- $\hat{\beta}_t(j)$ (במקום באמצעות $\alpha_t(j)$ ו- $\beta_t(j)$).

<u>חלק ב' – תרגיל מחשב</u>

יש לציין בשורה הראשונה בקובץ הקוד את ת.ז ושמות המגישים, בצורת הערה.

בהמשך לשאלה 1 בחלק התיאורטי, נתון מודל HMM בעל פילוג מוצא בדיד: המודל הוא בעל שני מצבים, 0 ו- 1, יש לו מוצא בינארי, גם כן 0 או 1, ומתקיים:

$$P\left(X_t = 1 \middle| S_t = 0\right) = 0.2$$
$$P\left(X_t = 1 \middle| S_t = 1\right) = 1$$
$$P\left(S_{t+1} = 1 \middle| S_t = 0\right) = 0.2$$
$$P\left(S_{t+1} = 0 \middle| S_t = 1\right) = 0.1$$

כמו כן, המצב הראשוני של המודל הוא תמיד 0.

ראשית השתמשו במודל על מנת ליצור בסיס נתונים לאימון הפרמטרים ע"י הגרלת $L = 200$ סדרות אימון, כ"א באורך $T = 10$. ניתן להשתמש בפונקצית matlab, hmmgenerate(), או לחילופין בספריית הפייתון https://github.com/guyz/HMM *שמבצעות* זאת.

עכשיו השתמשו בבסיס הנתונים ע"מ לשערך את פרמטרי המודל: הסתברות המצב ההתחלתי, הסתברויות המעבר והסתברויות המוצא. אתם מתבקשים כמובן לכתוב בעצמכם את תוכנית השערוך ולא להשתמש בפונקציית ספריה מוכנה. רישמו את הפרמטרים המשוערכים בטבלה לאחר 3 איטרציות EM ולאחר 20 איטרציות EM.

אתחול יתבצע כדלקמן:

- הסתברויות המצב ההתחלתי: $\begin{pmatrix} 1/2 & 1/2 \end{pmatrix}$
- מטריצת המעבר ההתחלתית: $\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$
- הסתברויות המוצא ההתחלתיות: $\begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}$ (במצב 0 פולטים 1 בהסתברות $1/4$, במצב 1 פולטים 1 בהסתברות $3/4$).