

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

### תרגיל בית 7

תרגיל בית זה עוסק באלגוריתם EM ובמודל עירובים. 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 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

בפיתוח אלגוריתם ה-EM לשערוך מודל עירובים גאוסים רב ממדים שבצענו בכיתה קיבלנו

$$Q(\theta|\theta') = \sum_{j=1}^N \sum_{l=1}^M \alpha_l^j \left[ \log c_l - \frac{1}{2} \log |\Sigma_l| - \frac{1}{2} (\mathbf{x}_j - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_l) \right]$$

עבור  $\theta' = \{c_l', \boldsymbol{\mu}_l', \Sigma_l'\}_{l=1}^M$ ,  $\theta = \{c_l, \boldsymbol{\mu}_l, \Sigma_l\}_{l=1}^M$  הראינו ש-

$$\hat{c}_l = \frac{\alpha_l}{\sum_{k=1}^M \alpha_k} = \frac{\alpha_l}{N} \quad l=1,2,\dots,M$$
 ממקסמים את  $Q(\theta|\theta')$  ביחס ל- $\{c_l\}_{l=1}^M$ . הראו ש-

$$\hat{\boldsymbol{\mu}}_l = \frac{\sum_{j=1}^N \mathbf{x}_j \alpha_l^j}{\alpha_l}, \quad \hat{\Sigma}_l = \frac{1}{\alpha_l} \sum_{j=1}^N \alpha_l^j \mathbf{x}_j \mathbf{x}_j^T - \hat{\boldsymbol{\mu}}_l \hat{\boldsymbol{\mu}}_l^T, \quad l=1,2,\dots,M$$
 ממקסמים את  $Q(\theta|\theta')$  ביחס ל-

$\{\boldsymbol{\mu}_l, \Sigma_l\}_{l=1}^M$ . השתמשו בתוצאה שקיבלנו בגיליון תרגילים מספר 2, שאלה 3.

#### שאלה 2

רשמו את נוסחת האיטרציה של אלגוריתם EM לשערוך מודל עירובים גאוסים רב ממדים, עבור המקרה של פילוגים עם מטריצות שונות אלכסוניות.

#### שאלה 3

השאלה דנה במודל העירובים הבא. הוקטור הנצפה הוא  $\mathbf{x} = (x_1, x_2, \dots, x_d)$  כאשר  $x_i \in \{0,1\}$ . וקטור הפרמטרים  $\theta$  נתון ע"י

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{c}), \quad \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M), \quad \boldsymbol{\mu}_l = (\mu_{l,1}, \mu_{l,2}, \dots, \mu_{l,d}) \text{ for } 1 \leq l \leq M, \quad \mathbf{c} = (c_1, c_2, \dots, c_M)$$

כאשר מקדמי הערוב  $c_l$  מקיימים  $0 \leq c_l \leq 1$  ו-  $\sum_{l=1}^M c_l = 1$ . תהליך היצירה של תצפית מהתהליך מתחיל

על ידי בחירת העירוב  $l$  בהסתברות  $c_l$ . לאחר מכן פולטים את וקטור התצפית  $\mathbf{x}$  בהסתברות

$$\prod_{i=1}^d \mu_{l,i}^{x_i} (1 - \mu_{l,i})^{1-x_i} \text{ . לכן}$$

$$p(x|\boldsymbol{\theta}) = \sum_{l=1}^M c_l \cdot \prod_{i=1}^d \mu_{l,i}^{x_i} (1 - \mu_{l,i})^{1-x_i}$$

נתונים  $N$  וקטורי תצפית בלתי תלויים סטטיסטית,  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ , מהתהליך. רשמו את נוסחת

האיטרציה של אלגוריתם EM לשערוך הפרמטרים  $\boldsymbol{\theta}$  מתוך וקטורי התצפית הללו.

## חלק ב' – תרגיל מחשב

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

בתרגיל זה נכתוב סימולציה לאלגוריתם ה-EM. יש להגיש קובץ אחד אשר כולל את הפתרון עבור שני הסעיפים, כאשר תופיע הערה בתחילת כל פתרון שמציין את מספר הסעיף.

נניח פילוג עירובים גאוסיים עם מטריצות שונות אלכסוניות בעל הפרמטרים הבאים:

$$c_1 = 0.5, c_2 = 0.5, \mu_1 = (1,1)^T, \mu_2 = (3,3)^T, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}$$

א. תכנתו את האלגוריתם שקיבלתם בשאלה 2 עבור  $N = 2000$  דגימות מוגרלות מתוך התהליך.

אתחלו את הפרמטרים  $c_1, c_2$  ע"י הגרלה אחידה בתחום  $(0,1)$ . כ"א מיתר הפרמטרים יאותחל

ע"י פילוג אחיד בתחום  $(0,5)$ . רשמו בטבלה את הפרמטרים המשוערכים לאחר 2, 10 ו-100

איטרציות EM.

ב. עכשיו, חזרו על השאלה עבור אלגוריתם  $K$  ממוצעים עבור  $N = 2000$  דגימות מוגרלות כאשר

$K = 2$  והמטריקה היא מטריקה אוקלידית. רשמו בטבלה את וקטורי הממוצעים שהאלגוריתם

שערך לאחר 2, 10 ו-100 איטרציות.