נושאים באנליזה סטטיסטית מרובת משתנים: תרגיל מס' 2

תאריך הגשה: 20.6.2022

1. שערוך פרמטרי של צפיפיות פילוג הסתברות רב מימדית באמצעות אלגוריתם ה-EM:

יהי $\mathbf{X}\in\mathbb{R}^p$ וקטור אקראי לא גאוסי עם צפיפות פילוג הסתברות הנתונה לפי מודל GMM:

$$f_{\mathbf{X}}\left(\mathbf{x};\boldsymbol{\theta}\right)=\sum_{m=1}^{M}w_m\phi\left(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\right)$$

כאשר $\phi\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ צפיפות פילוג גאוסית עם תוחלת $\boldsymbol{\mu}$ וקוואריאנס $\boldsymbol{\Sigma}$, $\left\{w_m\right\}_{m=1}^{M}$ הם מקדמי ערבוב שסכומם 1, $\left\{\boldsymbol{\mu}_m\right\}_{m=1}^{M}$ סדרת וקטורי תוחלות, $\left\{\boldsymbol{\Sigma}_m\right\}_{m=1}^{M}$ היא סדרת מטריצות קוואריאנס

ו- $\boldsymbol{\theta}\triangleq\left[w_1,\ldots,w_M,\boldsymbol{\mu}_1^T,\ldots,\boldsymbol{\mu}_M^T,\mathrm{vec}\left(\boldsymbol{\Sigma}_1\right),\ldots,\mathrm{vec}\left(\boldsymbol{\Sigma}_M\right)\right]^T$ מסמן את וקטור פרמטרי המודל.

א. פתחו אלגוריתם EM לשיערוך פרמטרי המודל בהינתן סדרת דגימות i.i.d $\mathbf{X}_1,\ldots,\mathbf{X}_N$ מהפילוג של $\mathbf{X}$ יש לרשום במפורש את צעדי האלגוריתם.

ב. הניחו כי פרמטרי המודל מקבלים את הערכים הבאים:

- סדר המודל $M=3$,
- $w_1=w_2=0.4$, $w_3=0.2$,
- $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\boldsymbol{\mu}_3=\left[0,0\right]^T$,
- $\boldsymbol{\Sigma}_1=\begin{bmatrix}1 & 0.9\\ 0.9 & 1\end{bmatrix}$, $\boldsymbol{\Sigma}_2=\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}$, $\boldsymbol{\Sigma}_3=\begin{bmatrix}0.05 & 0\\ 0 & 5\end{bmatrix}$

באמצעות האלגוריתם שפיתחתם בסעיף א', שערכו את פרמטרי המודל בהינתן סדרה של $N=1000$ דגימות i.i.d מהפילוג של $\mathbf{X}$.

- נסמן ב- $\hat{\boldsymbol{\theta}}_i$ את וקטור הפרמטרים המשוערך באיטרציה i של האלגוריתם. ציירו את פונקציית ה-log-likelihood $\log f_{\mathbf{X}_1,\ldots,\mathbf{X}_N}\left(\mathbf{X}_1,\ldots,\mathbf{X}_N;\hat{\boldsymbol{\theta}}_i\right)$ כפונקציה של אינדקס האיטרציה. האם הפונקציה מונוטונית עולה/יורדת? הסבירו את התוצאה.
- ציירו scatter-plot של הריאליזציות שהגרלתם. השתמשו בפונקציית EllipsPlot2D.m המצורפת כנספח לדף התרגיל על מנת לצייר על גבי ה-scatter-plot את התוחלות המשוערכות ואת ה-

concentration ellipses של מטריצות הקוריאנס ששיערכתם. כמו כן, ציירו את התוחלות האמיתיות ואת ה-concentration ellipses של מטריצות הקוריאנס האמיתיות.

2. קונסיסטנטיות במובן MSE של Multivariate kernel density estimator:

נניח כי פרמטר רוחב החלון h הוא פונקציה של מספר המדידות N, כלומר $h \equiv h(N)$. בהנחה כי פונקציית צפיפות הפילוג $\hat{f}_{\mathbf{X}}(\mathbf{x})$ ופונקציית הגרעין $K(\mathbf{x})$ הן חסומות מעל $\mathbb{R}^p$, מצאו תנאי מספיק על $h(N)$ שעבורו ה-multivariate kernel density estimator:

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{Nh^p(N)} \sum_{n=1}^{N} K\left(\frac{1}{h(N)} (\mathbf{x} - \mathbf{X}_n) \right)$$

הוא קונסיסטנטי במובן של MSE, כלומר מתקיים:

$$\mathrm{E}\left[\left(\hat{f}_{\mathbf{X}}(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x}) \right)^2 \right] \xrightarrow[N \to \infty]{} 0 \quad \forall \mathbf{x} \in \mathbb{R}^p$$

3. הערכת גודל מדגם עבור דיוק נקוב של Multivariate kernel density estimator:

בשאלה זו, נרצה לקבל הערכה לגבי גודל המדגם הנדרש עבור מימדים שונים כדי לקבל דיוק נקוב. נניח כי רוצים לשערך פונקציית צפיפות פילוג גאוסית סטנדרטית (תוחלת אפס וקוריאנס יחידה) באמצעות kernel density estimator עם פונקציית גרעין גאוסית.

א. כתבו ביטוי מפורש עבור ה-relative mean squared error בנקודה $\mathbf{x} = \mathbf{0}$ כפונקציה של המימד p, גודל המדגם N ופרמטר רוחב הגרעין h:

$$\mathrm{MSER}(p, N, h) \triangleq \frac{\mathrm{E}\left[\left(\hat{f}_{\mathbf{X}}(\mathbf{0}) - f_{\mathbf{X}}(\mathbf{0}) \right)^2 \right]}{f_{\mathbf{X}}^2(\mathbf{0})}$$

ב. בצעו את השלבים הבאים:

- לכל ערך של $p \in \{1, 2, \ldots, 20\}$ ולכל ערך של $N \in \{2^1, 2^2, \ldots, 2^{50}\}$ מצאו באופן נומרי את ה- optimal relative mean squared error:

$$\mathrm{MSER\_OPT}(p, N) \triangleq \min_{h \in [0.01, 10]} \mathrm{MSER}(p, N, \mathrm{h})$$

- לכל ערך של המימד p מצאו באמצעות תוצאת סעיף ב' את הערך המינימלי של N שעבורו $\mathrm{MSER\_OPT}(p,N) \le 0.1$
- ציירו את הערכים המינימליים של N כפונקציה של המימד. רצוי להשתמש בסקלה לוגריתמית על ציר גודל המדגם.

ג. מהגרף שקיבלתם בסעיף ב' הסיקו לגבי מהירות השינוי בגודל המדגם הנדרש עם הגידול במימד.

4. פרמטר רוחב חלון אופטימאלי במובן AMISE:

יהי $\mathbf{X} \in \mathbb{R}^2$ וקטור אקראי לא גאוסי עם צפיפות פילוג הסתברות הנתונה לפי מודל GMM:

$$f_{\mathbf{X}}(\mathbf{x}) = 0.5\phi(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + 0.5\phi(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$$

כאשר $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 5/9 & -4/9 \\ -4/9 & 5/9 \end{bmatrix}$, $\boldsymbol{\mu}_1 = [2/3, 2/3]^T$ ו- $\boldsymbol{\mu}_2 = [-2/3, -2/3]^T$. נשים לב כי תחת המודל המדובר, ל-$\mathbf{X}$ יש תוחלת $\mathbf{0}$ וקוואריאנס יחידה. מעוניינים לשערך את $f_{\mathbf{X}}(\mathbf{x})$ באמצעות סדרה של $N = 1000$ דגימות i.i.d. מהפילוג של $\mathbf{X}$ ע"י שימוש ב-kernel density estimator עם גרעין גאוסי.

א. חשבו את פרמטר רוחב החלון האופטימלי במובן AMISE המתקבל ע"י הנחת פילוג רפרנס גאוסי סטנדרטי. חשבו את ה-AMISE המתקבל עבור רוחב החלון שמצאתם.

ב. חשבו את פרמטר רוחב החלון האופטימלי במובן AMISE המתקבל ע"י הנחת פילוג רפרנס הזהה לפילוג האמיתי. חשבו את ה-AMISE המתקבל עבור רוחב החלון שמצאתם.

ג. השוו את ה-AMISEs שהתקבלו בסעיפים הקודמים. גבו את מסקנתכם באמצעות ציור של הפילוג האמיתי ושל השיערוכים שקיבלתם.