מבוא למערכות לומדות

הוקלד בידי אייל צוכר

מבוא למערכות לומדות 67577 תשפ"ב סמסטר ב'

תוכן העניינים

1 כלים מתמטיים 2

1 הרצאה - תורת האמידה 2

2 תרגול - קצת לינארית 13

שבוע 1

כלים מתמטיים

1 הרצאה - תורת האמידה

מטרות הקורס

- היכרות עם בעיות למידה ואלגוריתמי למידה (חלק תיאורטי)
- הבנה של העקרונות מאחורי האלגוריתמים (חלק מתמטי)
- יכולת לממש אלגוריתמים של בעיות למידה (חלק תכנותי)

שאלה

נניח שאנחנו רוצים לדעת מה השעה, אבל אין לנו שום דרך לדעת.
אז נשאל מישהו מה השעה. נניח שהוא ענה $13:15$ והשעה היא באמת $13:13$.

- איך נדע מה השעה האמיתית?
- איך נדע כמה כל תשובה מדוייקת?

ניתן לשאול הרבה אנשים ולחשב ממוצע של התשובות שלהם.
התשובות שקיבלנו מהאנשים נקראות <u>תצפיות</u> (observation). נסמנן ב-$x_1,\dots,x_m$.
הזמן המשוער שלנו הוא

$$\bar{x}=\frac{1}{m}\sum_i x_i$$

סט התצפיות שלנו נקרא <u>המדגם</u> (sample).
נחשוב על $x_1,\dots,x_n$ כערכים שהתקבלו (realized) במשתנים מקריים $X_1,\dots,X_n$.

הגדרה

נאמר כי $x_1,\dots,x_m$ הם מתפלגים באופן זהה (identically distributed) אם כולם ערכים שנתקבלו מערכים משתנים המקריים המתפלגים עם אותה התפלגות, כלומר, $X_1,\dots,X_m\sim\mathcal{P}$ ו-$X_i=x_i$ לכל i.

הגדרה

נאמר שדגימות $x_1,\dots,x_m$ הן בעלות אותה התפלגות ובלתי תלויים (independently identically distributed or i.i.d) אם הם מתפלגים באופן זהה ובלתי תלויים. במקרה זה נסמן

$$X_1,\dots,X_m \text{i.i.d.} \mathcal{P},\quad \forall i\ X_i=x_i$$

מהי $\mathcal{P}$?

נניח כי $\mathcal{P}$ היא התפלגות פרמטרית כלשהי שמוגדרת על ידי קבוצת פרמטרים $\theta\in\Theta$.

- $\vec{\theta}$ היא וקטור של פרמטרים.
- Θ היא קבוצת כל הערכים האפשריים לפרמטר.
- עבור $\mathcal{P}:=\text{Pois}\left(\lambda\right)$ אזי $\vec{\theta}:=\{\lambda\}$ ו-$\Theta:=\mathbb{R}_+$
- עבור $\mathcal{P}:=\text{N}\left(\mu,\sigma^2\right)$ אזי $\vec{\theta}:=\left\{\mu,\sigma^2\right\}$ ו-$\Theta:=\mathbb{R}\times\mathbb{R}_+$

בהינתן דגימות $x_1,\dots,x_m$ שנדגמו .i.i.d לפי $\mathcal{P}\left(\vec{\theta}\right)$ <u>כאשר אנחנו לא יודעים את $\vec{\theta}$</u>, נרצה לדעת מהו $\vec{\theta}^*$ שנותן את ההתאמה הכי טובה ל-$\vec{\theta}$ האמיתי.
(למשל, נניח ש-$\mathcal{P}$ היא פואסונית ונחפש את הפרמטר λ הכי טוב.)

פורמליזציה

- נגדיר פונקציית החלטה/כלל החלטה $\delta : \mathbb{R}^m \to \Theta$.
  - עבור $\mathcal{P} := \text{Pois}(\lambda)$,

$$(x_1, \ldots, x_m) \overset{\delta}{\mapsto} \lambda$$

  - עבור $\mathcal{P} := \text{N}\left(\mu, \sigma^2\right)$,

$$(x_1, \ldots, x_m) \overset{\delta}{\mapsto} \left(\mu, \sigma^2\right)$$

- את קבוצת כל פונקציות ההחלטה נסמן ב-$\Delta := \{\delta : \mathbb{R}^m \to \Theta\}$. היא מכונה מרחב ההשערות.

הגדרה

תהי $\delta \in \Delta$ פונקציית החלטה. אזי $\delta(X_1, \ldots, X_m)$ נקראת אומד (estimator) של הפרמטר $\vec{\theta}$, או לפעמים רק אומד. $\vec{\theta}$ המתקבל נקרא אומדן.

אומדים של התפלגות גאוסיאנית

נניח שיש לנו דגימות $\mathcal{P} \overset{\text{i.i.d.}}{\sim} x_1, \ldots, x_m$ עבור $\mathcal{P} := \text{N}\left(\mu, \sigma^2\right)$ ואנחנו רוצים להעריך את μ, σ^2. איך נוכל לעשות את זה? איזה אומד נבחר עבור μ, σ^2?

- האומד ממוצע המדגם (sample mean) מוגדר על ידי

$$\hat{\mu}_X := \frac{1}{m} \sum x_i$$

הערה: הסימון $\hat{\mu}$ הוא האומד לפרמטר μ.

- האומד שונות המדגם (sample variance) מוגדר על ידי

$$\hat{\sigma}_X^2 := \frac{1}{m-1} \sum \left(x_i - \hat{\mu}_X\right)^2$$

תכונות של אומדים

נתבונן למשל בשונות המדגם. יכולנו לבחור בהרבה אפשרויות אחרות לפונקציות:

- $\hat{\sigma}_1 := \frac{1}{m-1} \sum \left(x_i - \hat{\mu}\right)^2$
- $\hat{\sigma}_2 := \frac{1}{m} \sum \left(x_i - \hat{\mu}\right)^2$
- $\hat{\sigma}_3 := \frac{1}{m} \sum \left|x_i - \hat{\mu}\right|$

מה מהם "הכי טוב"? מה זה בכלל אומר "להיות הכי טוב"?

הערה

$X_1, \ldots, X_m$ הם משתנים מקריים, ולכן, כיוון ש-δ היא פונקציה של $X_1, \ldots, X_m$ אזי δ עצמו הוא משתנה מקרי. לכן, אנחנו יכולים לשאול מהי התוחלת של δ, מה השונות שלו, וכו'.

רעיון

כאשר נגדיר אומד, תכונה שנרצה שתהיה לו היא שהוא יהיה בלתי מוטה (unbiased), כלומר, בממוצע הערך שהוא מחזיר שווה לערך האמיתי.

הגדרה

יהי δ אומד עבור פרמטר $\vec{\theta}$. ההפרש $d := \delta\left(X_1, \ldots, X_n\right) - \theta$ נקרא <u>השגיאה</u> של δ.

הגדרה

יהי δ אומד עבור פרמטר $\vec{\theta}$. הגודל

$$\begin{aligned}\text{Bias}_{\vec{\theta}}\left[\delta\left(X_1, \ldots, X_m\right)\right] &:= \mathbb{E}_{X_1, \ldots, X_m \mid \vec{\theta}}[d] \\ &= \mathbb{E}_{X_1, \ldots, X_m \mid \vec{\theta}}\left[\delta\left(X_1, \ldots, X_m\right) - \vec{\theta}\right]\end{aligned}$$

נקראת ההטייה (bias או systemic error) של δ.
(התוחלת נלקחת על הרבה דגימות של $X_1, \ldots, X_m$ מההתפלגות עם פרמטר $\vec{\theta}$ מסויים)

הגדרה

יהי δ אומד עבור פרמטר $\vec{\theta}$.
נאמר כי δ הוא <u>בלתי מוטה</u> אם לכל $\vec{\theta} \in \Theta$ מתקיים

$$\text{Bias}_{\vec{\theta}}\left[\delta\left(X_1, \ldots, X_m\right)\right] = 0$$

טענה

האומד ממוצע המדגם $\hat{\mu}$ הוא בלתי מוטה עבור ההתפלגות $\mathcal{P} := \mathrm{N}\left(\mu, \sigma^2\right)$

הוכחה

לכן $\hat{\mu}$ הוא אומד בלתי מוטה.

טענה

האומד שונות המדגם $\hat{\sigma}^2$ הוא בלתי מוטה עבור ההתפלגות $\mathcal{P} := \mathrm{N}\left(\mu, \sigma^2\right)$.

הוכחה

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{1}{m-1}\sum_i \mathbb{E}\left[\left(X_i-\hat{\mu}\right)^2\right] = \frac{1}{m-1}\mathbb{E}\left[\left(X_i - \frac{\sum_j X_j}{m}\right)^2\right]$$

$$= \frac{1}{m-1}\mathbb{E}\left[X_i^2 - 2X_i\cdot\underbrace{\left(\frac{1}{m}\sum_j X_i\right)}_{\hat{\mu}} + \underbrace{\frac{1}{m^2}\sum_{j,k}X_jX_k}_{\hat{\mu}^2}\right] =$$

$$= \frac{1}{m-1}\mathbb{E}\left[\sum_i\left[X_i^2\right] - \frac{2}{m}\sum_{i,j}\mathbb{E}\left[X_iX_j\right] + \frac{1}{m}\sum_{j,k}\mathbb{E}\left[X_jX_k\right]\right] =$$

$$= \frac{1}{m-1}\left(\left(1-\frac{1}{m}\right)\sum_i\mathbb{E}\left[X_i^2\right] - \frac{1}{m}\sum_{i\neq j}\mathbb{E}\left[X_iX_j\right]\right) =$$

מכיוון שהמדידות הן i.i.d. אזי נגדיר $X\sim \mathrm{N}\left(\mu,\sigma^2\right)$ ויתקיים:

- לכל i, $\mathbb{E}\left[X_i\right]=\mathbb{E}\left[X\right]$ ו-$\mathbb{E}\left[X_i^2\right]=\mathbb{E}\left[X^2\right]$
- לכל $i\neq j$, $\mathbb{E}\left[X_iX_j\right]=\mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]=\left(\mathbb{E}\left[X\right]\right)^2$ (כי המ"מ ב"ת)

לכן

$$= \frac{1}{m-1}\left(\frac{m-1}{m}\cdot m\mathbb{E}\left[X^2\right] - \frac{m\left(m-1\right)}{m}\left(\mathbb{E}\left[X\right]\right)^2\right) =$$
$$= \frac{1}{m-1}\left(\left(m-1\right)\mathbb{E}\left[X^2\right] - \left(m-1\right)\left(\mathbb{E}\left[X\right]\right)^2\right) =$$
$$= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \mathrm{Var}\left(X\right) = \sigma^2$$

כנדרש.

הגדרה

יהי δ אומד עבור פרמטר $\vec{\theta}$. <u>השונות</u> של δ היא

$$\mathrm{Var}\left(\delta\right) := \mathbb{E}_{X_1,\ldots,X_m|\vec{\theta}}\left[\left(\delta\left(X_1,\ldots,X_m\right) - \mathbb{E}_{X_1,\ldots,X_m|\vec{\theta}}\left[\delta\left(X_1,\ldots,X_m\right)\right]\right)^2\right]$$

דוגמא

תהי $\mathcal{P}$ התפלגות עם שונות של σ^2.
יהיו $X_1,\ldots,X_m \overset{\text{i.i.d.}}{\sim} \mathcal{P}$. מהו $\mathrm{Var}\left(\hat{\mu}\right)$?

$$\mathrm{Var}\left(\hat{\mu}\right) = \mathrm{Var}\left(\frac{1}{m}\sum X_i\right) = \frac{1}{m^2}\mathrm{Var}\left(\sum X_i\right) =$$
$$\underset{\substack{\uparrow\\ \text{i.i.d.}}}{=} \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}\left(X_i\right) = \frac{1}{m^2}\cdot m\cdot\sigma^2 = \frac{\sigma^2}{m}$$

משמעות

השונות של האומד נותן לנו מדד לכמה טובים הביצועים של האומד שלנו.
אם השונות גדולה, האומד לא מאוד טוב.

דוגמא

נוכל לומר שהאומד הכי טוב הוא $\delta \in \Delta$ כך ש-δ בלתי פוטה וגם $\mathrm{Var}\left(\delta\right)$ **מינימלי**.

נראות מקסימלית

הגדרה

יהי $X \sim \mathcal{P}\left(\vec{\theta}\right)$ ו-f פונקצית הצפיפות של $\mathcal{P}$.
פונקציית הנראות היא

$$\mathcal{L}\left(\vec{\theta} \mid x\right) := f_{\vec{\theta}}\left(x\right)$$

כאשר x הוא realization של X.

דוגמא

נתבונן בהתפלגות גאוסיאנית עם $\mathrm{N}\left(\mu, \sigma^2\right) \overset{\text{i.i.d.}}{\sim} x_1, \ldots, x_m$ ונסמן $\vec{\theta} = \left\{\mu, \sigma^2\right\}$.
פונקציית הנראות היא

$$\mathcal{L}\left(\vec{\theta} \mid x_i\right) = f_{\vec{\theta}}\left(x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}}$$

לכל $i \in [m]$.
מכיוון ש-$x_1, \ldots, x_m$ הן i.i.d. אזי

$$\begin{aligned}\mathcal{L}\left(\vec{\theta} \mid x_1, \ldots, x_m\right) &= f_{\vec{\theta}}\left(x_1, \ldots, x_m\right) = \\ &= \prod_{i=1}^{m} f_{\vec{\theta}}\left(x_i\right) = \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum \left(x_i - \mu\right)^2\right)\end{aligned}$$

הערה

נשתמש ב-$\mathcal{L}$ על מנת להשוות בין דברים באופן יחסי! אם יהיו לנו הרבה דגימות, $\mathcal{L}$ ישאף ל-0 בכל מקרה.
אבל נוכל לשאול מי יותר סביר מבין $\vec{\theta}_1$ או $\vec{\theta}_2$ על ידי כך שנשווה בין $\mathcal{L}\left(\vec{\theta}_1 \mid x_1, \ldots, x_m\right)$ ו-$\mathcal{L}\left(\vec{\theta}_2 \mid x_1, \ldots, x_m\right)$.

הגדרה

תהי $\mathcal{L}$ פונקציית הנראות עבור התפלגות $\mathcal{P}$ כלשהי, המתואר על ידי $\vec{\theta} \in \Theta$.
יהי $X \sim \mathcal{P}\left(\vec{\theta}\right)$ משתנה מקרי ויהי x ערך שנדגם ממנו.

אומד הנראות המרכזית (Maximum Likelihood Estimator or MLE) עבור $\vec{\theta}$ הוא

$$\hat{\theta}^{\mathrm{MLE}} := \underset{\vec{\theta} \in \Theta}{\operatorname{argmax}}\, \mathcal{L}\left(\vec{\theta} \mid x\right)$$

דוגמא

נחשב את ה-MLE של הממוצע של ההתפלגות הגאוסיאנית $\hat{\mu}^{\text{MLE}}$, כאשר נתון לנו σ^2.
יהיו $x_1, \ldots, x_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\mu, \sigma^2\right)$.
נרצה למצוא את

$$\hat{\mu}^{\text{MLE}} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \mathcal{L}\left(\mu \mid x_1, \ldots, x_m, \sigma^2\right)$$

מתקיים

$$\begin{aligned}
\hat{\mu}^{\text{MLE}} &= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \mathcal{L}\left(\mu \mid x_1, \ldots, x_m, \sigma^2\right) = \\
&\underset{\text{i.i.d.}}{\overset{\uparrow}{=}} \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right) = \\
&\underset{(*)}{\overset{\uparrow}{=}} \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \prod_i \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right) = \\
&= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \exp\left(-\frac{1}{2\sigma^2} \sum_i \left(x_i - \mu\right)^2\right) = \\
&\underset{\text{מונוטונית עולה } \log}{\overset{\uparrow}{=}} \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \log\left(\exp\left(-\frac{1}{2\sigma^2} \sum_i \left(x_i - \mu\right)^2\right)\right) = \\
&= \operatorname{argmax}\left(-\frac{1}{2\sigma^2} \sum_i \left(x_i - \mu\right)^2\right) = \\
&= \operatorname{argmax}\left(-\sum_i \left(x_i - \mu\right)^2\right)
\end{aligned}$$

$(*)$ נבחין כי אנחנו מחפשים את מי שממקסם את הפונקציה ולא את הערך המקסימלי של הפונקציה! הקורס $\frac{1}{\sqrt{2\pi\sigma^2}}$ אולי משפיע על ערך הפונציה אך הוא אינו משפיע על הערך שממקסם אותה

על מנת למקסם את הפונקציה, נגזור ונשווה ל-0:

$$\frac{\partial}{\partial \mu}\left(-\sum_i \left(x_i - \mu\right)^2\right) = -\sum_{i=1}^m \frac{\partial\left(x_i - \mu\right)^2}{\partial \mu} = \sum_{i=1}^m 2\left(x_i - \mu\right) = 0$$

ולכן

$$\boxed{\hat{\mu}^{\text{MLE}} = \frac{1}{m} \sum_{i=1}^m x_i}$$

מסקנה

אומד ממוצע המדגם (sample mean), שראינו שהוא אומד בלתי מוטה, הוא גם האומד maximum likelihood!

הערה ממוצע המדגם ממזער את ריבועי המרחקים ממנו.

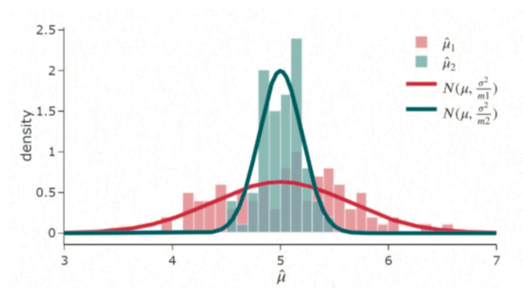
כמה טוב השיערוך?

מכיוון שהאומד הוא משתנה מקרי, יש לו התפלגות.
למשל, במקרה של ה- sample mean, אנחנו יודעים ש-$x_1,\dots,x_m \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\mu,\sigma^2\right)$.
לכן,

$$\mathbb{E}\left[\hat{\mu}\right]=\mu,\quad \mathrm{Var}\left(\hat{\mu}\right)=\frac{\sigma^2}{m}$$

ניתן גם להראות (אבל לא נראה) שמתקיים $\hat{\mu}\sim\mathrm{N}\left(\mu,\frac{\sigma^2}{m}\right)$.
בפרט:

- האומד נותן התפלגות גאוסיאנית סביב הערך האמיתי μ (וזה הגיוני כי הוא unbiased)
- השונות פרופורציונלית לשונות של המידע המקורי, והיא דועכת באופן לינארי ביחס למספר הדגימות.



הגרף מתאר מצב בו הגרלנו הרבה פעמים m_1 ו-m_2 ערכים מהתפלגות גאוסיאנית עם μ,σ^2, וראינו אילו ערכי $\hat{\mu}_1,\hat{\mu}_2$ התקבלו לנו לפי ממוצע המדגם.
לכן נוכל לשאול שאלות כגון:

- מה הסיכוי שנשערך ערך מסויים?
- כמה אנחנו בטוחים בתוצאה שהתקבלה לנו?
- מה הסיכוי לסטייה מהערך האמיתי, וכיצד זה תלוי במספר הדגימות?

דוגמא

נניח שאנחנו רוצים להעריך את ההטייה של מטבע לא הוגן כלשהו.
כלומר, נתון מטבע שיש לו סיכוי של p לתת $Heads$ ואנחנו רוצים לדעת מהו p.
נחשוב על הטלת מטבע בתור משתנה מקרי ברנולי

$$\mathcal{D}_p\left(X\right):=\begin{cases}p & X=1\\ 1-p & X=0\end{cases},\quad p\in\left[0,1\right]$$

בהינתן דגימה של m הטלות $S:=\left\{x_1,\dots,x_m\right\}$, נסמן את התפלגות של m הדגימות על ידי $\mathcal{D}_p^m$ ונסמן את ההסתברות לקבל את הדגימות S ב-$\mathcal{D}_p^m\left(S\right)$.
בהינתן מדגם כזה, נרצה להפיק אלגוריתם למידה $\mathcal{A}$ שיעריך את p ונשאל כמה האלגוריתם שלנו מדוייק.
האלגוריתם מקבל את S כקלט, שנבחר i.i.d. לפי $\mathcal{D}$ ומוציא כפלט את ההערכה ל-p.
נסמן אותה ב-$\mathcal{A}\left(S\right)$ או $\hat{p}\left(S\right)$ או בקיצור נמרץ $\hat{p}$.
אלגוריתם הלמידה בו נשתמש יהיה פשוט השכיחות האמפירית (כלומר, כמה פעמים מופיע $Heads$ במדגם), שזה בדיוק ממוצע המדגם.

$$\hat{p}\left(S\right)=\frac{1}{m}\sum_i x_i$$

אנחנו יודעים שהאומד הזה הוא בלתי מוטה. לכן, בתוחלת נקבל $\mathbb{E}_S\left(\hat{p}\right)=p$.
בעיות:

- המדגם שלנו סופי, ולכן $\hat{p}$ שלנו לא יכול מדויק. לכן נסתפק ב-$\mathcal{A}$ שיהיה מדויק מספיק עבורנו, כלומר, $|\hat{p}-p| \leq \varepsilon$ עבור $\varepsilon \in (0,1)$ כלשהו.
- אלא אם כן $p \in \{0,1\}$, תמיד יש סיכוי כלשהו לקבל S שלא מייצג את $\mathcal{D}$. לכן, יהיה עלינו לשאול מה הסיכוי ש-$\mathcal{A}$ יחזיר תוצאה טובה מספיק, כלומר, נחשב $\Pr\left(|\hat{p}-p| \leq \varepsilon\right)$.

תזכורת (א"ש מרקוב)

עבור מ"מ X אי שלילי עם תוחלת סופית מתקיים:

$$\Pr\left(X \geq a\right) \leq \frac{\mathbb{E}\left(X\right)}{a}$$

במקרה שלנו, $|\hat{p}-p|$הוא מ"מ אי שלילי.
לכל רמת דיוק $\varepsilon \in (0,1)$ נוכל לחסום את הסיכוי לסטות מ-p ביותר מ-ε על ידי

$$\Pr\left[|\hat{p}-p| \geq \varepsilon\right] \leq \frac{\mathbb{E}\left(|\hat{p}-p|\right)}{\varepsilon}$$

נותר לנו לחסום מלמעלה את $\mathbb{E}\left(|\ \hat{p}-p\right)$.
ניזכר כי $\mathrm{Var}\left(A\right) = \mathbb{E}\left(A\right) - \left(\mathbb{E}\left(A\right)\right)^2$ ולכן

$$\mathrm{Var}\left(|\hat{p}-p|\right) = \mathbb{E}\left(|\hat{p}-p|\right) - \mathbb{E}^2\left(|\hat{p}-p|\right)$$

ולכן, מהיות $\mathrm{Var}\left(|\hat{p}-p|\right) \geq 0$,

$$\mathbb{E}^2\left(|\hat{p}-p|\right) \leq \mathbb{E}\left(|\hat{p}-p|^2\right)$$

לכן

$$\begin{aligned}\mathbb{E}^2\left(|\hat{p}-p|\right) &\leq \mathbb{E}\left(|\hat{p}-p|^2\right) = \mathbb{E}\left((\hat{p}-p)^2\right) = \\ &\underset{(*)}{=} \mathbb{E}\left(\left(\hat{p}-\mathbb{E}\left(\hat{p}\right)\right)^2\right) = \mathrm{Var}\left(\hat{p}\right) =\end{aligned}$$

$(*)$ כיוון ש-$\hat{p}$ הוא אומד בלתי מוטה, $\mathbb{E}\left(\hat{p}\right) = p$.
נזכור כי $\hat{p}$ הוא ממוצע של m הטלות מטבע, ולכן

$$= \mathrm{Var}\left(\frac{1}{m}\sum X_i\right) = \frac{1}{m^2}\mathrm{Var}\left(\sum x_i\right) \underset{\text{i.i.d.}}{=} \frac{p\left(1-p\right)}{m} \underset{(*)}{\leq} \frac{1}{4m}$$

$(*)$ שכן הפונקציה $f\left(p\right) := p\left(1-p\right)$ מקבל את הערך המקסימלי שלה בקטע $[0,1]$ עבור $p = \frac{1}{2}$ ו-$f\left(p\right) = \frac{1}{4}$.

לכן

$$\mathbb{E}\left(|\hat{p}-p|\right) \leq \frac{1}{\sqrt{4m}}$$

אם כן

$$\Pr\left[\left|\hat{p}-p\right|\geq\varepsilon\right]\leq\frac{\mathbb{E}\left(\left|\hat{p}-p\right|\right)}{\varepsilon}\leq\frac{1}{\sqrt{4m\varepsilon^{2}}}$$

כלומר

$$\Pr\left[\left|\hat{p}-p\right|\leq\varepsilon\right]\geq1-\frac{1}{\sqrt{4m\varepsilon^{2}}}$$

בתור שלב אחרות, נסמן $\delta=\frac{1}{\sqrt{4m\varepsilon^{2}}}$. נוכל לבודד את m ונקבל $m=\frac{1}{4\varepsilon^{2}\delta^{2}}$.
לכן, לכל $\varepsilon,\delta\in(0,1)$, אם נקבל $m\geq\frac{1}{4\varepsilon^{2}\delta^{2}}$ דגימות, אזי

$$\Pr\left(\left|\hat{p}-p\right|\leq\varepsilon\right)\geq1-\delta$$

אפשר לשפר את התוצאה הזו!
למשל, אם נשתמש באי-שוויון צ'בישב.

תזכורת (א"ש צ'בישב)

יהי X מ"מ בעל תוחלת ושונות. אזי לכל $\varepsilon>0$ מתקיים

$$\Pr\left[\left|X-\mathbb{E}\left[X\right]\geq\varepsilon\right|\right]\leq\frac{\mathrm{Var}\left(X\right)}{\varepsilon^{2}}$$

לכן, בהינתן $\mathrm{Ber}\left(p\right)\overset{\text{i.i.d.}}{\sim}x_{1},\ldots,x_{m}$ ו-$\hat{p}$ כמו שהגדרנו נקבל

$$\begin{aligned}\Pr\left[\left|\hat{p}-p\right|\geq\varepsilon\right]&=\Pr\left[\left|\hat{p}-\mathbb{E}\left(\hat{p}\right)\right|\geq\varepsilon\right]\leq\frac{\mathrm{Var}\left(\hat{p}\right)}{\varepsilon^{2}}=\\&=\frac{1}{\varepsilon^{2}}\mathrm{Var}\left(\frac{1}{m}\sum x_{i}\right)\underset{\substack{\uparrow\\\text{i.i.d.}}}{=}\frac{1}{m^{2}\varepsilon^{2}}\sum\mathrm{Var}\left(x_{i}\right)=\\&=\frac{p\left(1-p\right)}{m\varepsilon^{2}}\leq\frac{1}{4m\varepsilon^{2}}\end{aligned}$$

החסם יתנהג כמו $\frac{1}{m}$, וזה יותר טוב מהחסם של מרקוב, שנותן $\frac{1}{\sqrt{m}}$.
לכן, נגדיר $\delta=\frac{1}{4m\varepsilon^{2}}$ ונקבל ש-$m=\frac{1}{4\delta\varepsilon^{2}}$.
אם כן, הסקנו כי לכל $\varepsilon,\delta\in(0,1)$, אם ניתן $m>\frac{1}{4\delta\varepsilon^{2}}$ דגימות אזי

$$\Pr\left(\left|\hat{p}-p\right|\leq\varepsilon\right)\geq1-\delta$$

התפלגויות רבות משתנים

עד עכשיו התמודדנו עם התפלגויות של משתנה אחד.
אבל מה אם יש כמה דגימות של משתנים שתלויים זה בזה?
למשל, כיצד נמדל גובה ומשקל של אנשים?
אפשר להחליט למשל ש-$\mathrm{N}\left(75,3\right)\overset{\text{i.i.d.}}{\sim}w_{1},\ldots,w_{m}$ ו-$\mathrm{N}\left(170,5\right)\overset{\text{i.i.d.}}{\sim}h_{1},\ldots,h_{m}$.
אבל הגובה והמשקל תלויים זה בזה!
לכן נשתמש בוקטורים מקריים.

הגדרה

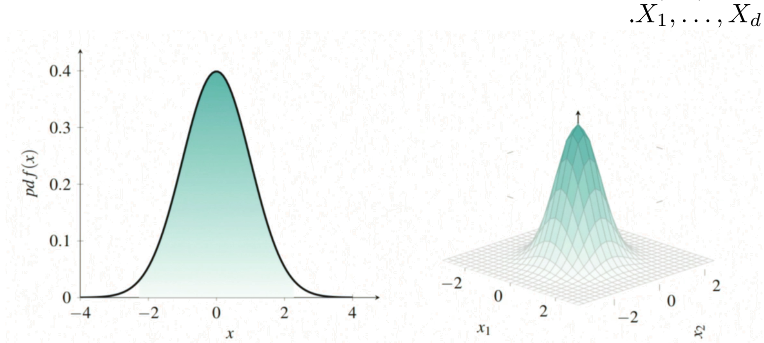
יהיו $X_1, \dots, X_d$ קבוצה של מ"מ המוגדרים מעל אותו מרחב הסתברות.

אזי $X := \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ נקרא <u>וקטור מקרי</u>.

וקטור מקרי הוא פונקציה ממרחב המדגם ל-$\mathbb{R}^d$.

הגדרה

יהי $X := \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ וקטור מקרי. <u>ההתפלגות המשותפת</u> של $X_1, \dots, X_d$ הוא פונקצית ההתפלגות על כל הערכים האפשריים של $X_1, \dots, X_d$.



***(a)** Univariate Gaussian* $\mathcal{N}(0,1)$ ***(b)** Bivariate Gaussian* $\mathcal{N}(0,I_2)$

נשים לב כי קווי הגובה הם אליפסות יפות.

הגדרה

יהי $X := \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ וקטור מקרי. <u>מטריצת השונות המשותפת</u> (covariance matrix) Σ היא מטריצה $d \times d$ שהערך ה-(i,j) שלה הוא השונות המשותפת:

$$\Sigma_{ij} := \mathrm{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[\left(X_i - \mathbb{E}\left(X_i\right)\right)\left(X_j - \mathbb{E}\left(X_j\right)\right)\right]$$

הערה המטריצה היא ריבועית וסימטרית, והיא גם מטריצת psd.

הגדרה

יהי $X := \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ וקטור מקרי עם התפלגות נורמלית של כמה משתנים עם תוחלת $\mu \in \mathbb{R}^d$ ומטריצת שונות משותפת $\Sigma \in \mathbb{R}^{d \times d}$.

אזי פונקציית ההתפלגות המשותפת היא

$$f\left(X\right) = \frac{1}{\sqrt{\left(2\pi\right)^d \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(X-\mu\right)^T \Sigma^{-1}\left(X-\mu\right)\right)$$

(כאשר $\left|\Sigma\right| = \det\left(\Sigma\right)$)

במקרה זה נסמן $X \sim \mathrm{N}\left(\mu, \Sigma\right)$.

הערה אם $d=1$ נקבל את ההתפלגות הגאוסית.

דוגמא

יהי $X\sim \mathrm{N}\left(\mu,\Sigma\right)$ וקטור מקרי גאוסיאני דו מימדי, עם מטריצת שונויות משותפות אלכסונית, כלומר,

$$X\sim \mathrm{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right)$$

אזי ההתפלגויות השוליות הן

$$f\left(X_i\right)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{1}{2}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2\right)$$

בחזרה לדוגמא של גובה ומשקל.
נוכל למדל כל דגימה כוקטור $\vec{x}\in\mathbb{R}^d$, שהוא realization של הוקטור $X:=\left(X_{weight},X_{height}\right)^T$.
לפיכך, המדגם שלנו הוא $S:=\{\vec{x}_1,\ldots,\vec{x}_n\}$, כאשר כל $\vec{x}_i$ הוא realization של X.
הקואורדינאטה ה-j של הדגימה ה-i היא $\vec{x}_i\left(j\right)$ או x_{ij}.
הקואורדינאטה ה-j היא realization של הקואורדינאטה ה-j של X_i, שהיא $X_i^{(j)}$.

נרצה לשערך את ההתפלגות של המשתנים

$$\vec{x}_1,\ldots,\vec{x}_m\overset{\text{i.i.d.}}{\sim}\mathrm{N}\left(\begin{pmatrix}75\\ 170\end{pmatrix},\begin{pmatrix}3 & 1\\ 1 & 5\end{pmatrix}\right)$$

האומד ממוצע המדגם רב המשתנים שלנו הוא פשוט האומד ממוצע המדגם לכל משתנה בנפרד:

$$\hat{\vec{\mu}}=\begin{pmatrix}\vdots\\ \hat{\mu}_j\\ \vdots\end{pmatrix}=\begin{pmatrix}\vdots\\ \frac{1}{m}\sum_i x_{i,j}\\ \vdots\end{pmatrix}$$

האומדן של ה-Cov מעט יותר מורכב, שכן יש תלויות בין המשתנים.
האומד שונות המדגם המשותפת הבלתי מוטה בין X_i ל-X_j נתון על ידי

$$\hat{\sigma}\left(X_i,X_j\right)=\frac{1}{m-1}\sum_k\left(x_{ki}-\hat{\mu}_i\right)\left(x_{kj}-\hat{\mu}_j\right)$$

לבסוף, האומד מטריצת השונות המשותפת של המדגם היא מטריצה $\hat{\Sigma}$ מסדר $d\times d$ כך ש-

$$\hat{\Sigma}_{ij}:=\hat{\sigma}\left(X_i,X_j\right)$$

בכתיב וקטורי, עבור $\boldsymbol{X}\in\mathbb{R}^{m\times d}$ מטריצה ששורותיה הן הדגימות $\vec{x}_1,\ldots,\vec{x}_m$,

- מטריצת שונות המדגם המוטה נתונה על ידי

$$\hat{\Sigma}:=\frac{1}{m}\sum_{i=1}^m\left(\vec{x}_i-\hat{\mu}\right)\left(\vec{x}_i-\hat{\mu}\right)^T=\frac{1}{m}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

כאשר $\tilde{\boldsymbol{X}}$ היא המטריצה הממורכזת: $\tilde{\boldsymbol{X}}_{\cdot,i}=\boldsymbol{X}_{\cdot,i}-\hat{\mu}$

2 תרגול - קצת לינארית

הגדרה

פונקציה $d:X\times X\to\mathbb{R}$ (עבור קבוצה X כלשהי) תקרא מטריקה אם

1. $d\left(x,y\right)=0$ אם"ם $x=y$
2. $d\left(x,y\right)=d\left(y,x\right)$
3. $d\left(x,y\right)\leq d\left(x,z\right)+d\left(z+y\right)$

דוגמא

$d\left(u,v\right)=\sum\limits_{i}\left|u_{i}-v_{i}\right|$ היא מטריקה.

הוכחה

1. $\sum\limits_{i}\left|u_{i}-v_{i}\right|=0$ אם"ם כל אחד מהאיברים בסכום הוא 0 אם"ם $u_{i}=v_{i}$ לכל i אם"ם $u=v$
2. $d\left(u,v\right)=\sum\limits_{i}\left|u_{i}-v_{i}\right|=\sum\limits_{i}\left|v_{i}-u_{i}\right|=d\left(v,u\right)$
3.

$$\begin{aligned}d\left(u,v\right)=\sum_{i}\left|u_{i}-v_{i}\right| & =\sum_{i}\left|\left(u_{i}-w_{i}\right)+\left(w_{i}-v_{i}\right)\right|\leq\\ & \leq\sum_{i}\left|u_{i}-w_{i}\right|+\sum_{i}\left|w_{i}-v_{i}\right|=d\left(u,w\right)+d\left(w,u\right)\end{aligned}$$

הגדרה

פונקציה $\left\Vert \cdot\right\Vert :\mathbb{R}^{d}\to\mathbb{R}$ תיקרא נורמה אם היא מקיימת את התכונות הבאות:

1. $\left\Vert v\right\Vert \geq0$ לכל v, ו-$\left\Vert v\right\Vert =0$ אם"ם $v=0$
2. $\left\Vert \alpha v\right\Vert =\left|a\right|\left\Vert v\right\Vert$
3. $\left\Vert v+u\right\Vert \leq\left\Vert v\right\Vert +\left\Vert u\right\Vert$

דוגמאות

1. נורמת ℓ_{1} המוגדרת על ידי $\left\Vert v\right\Vert =\sum\limits_{i}\left|v_{i}\right|$
2. נורמת ℓ_{2} המוגדרת על ידי $\left\Vert v\right\Vert _{2}=\left(\sum\limits_{i}\left|v_{i}\right|^{2}\right)^{\frac{1}{2}}$
3. נורמת ℓ_{∞} המוגדרת על ידי $\left\Vert v\right\Vert _{\infty}=\max\limits_{i}\left\Vert v_{i}\right\Vert$

הכללה

הנורמת הללו הן חלק ממשפחת נורמות ℓ_{p}, המוגדרת על ידי

$$\left\Vert v\right\Vert _{p}=\left(\sum_{i}\left\Vert v_{i}\right\Vert ^{p}\right)^{\frac{1}{p}}$$

הגדרה

כדור היחידה לפי הנורמה $\|\cdot\|$ מוגדר על ידי

$$B_{\|\cdot\|} = \{x \in V \mid \|x\| \leq 1\}$$

הגדרה

פונקציה $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ נקראת <u>מכפלה פנימית</u> אם היא מקיימת את התכונות הבאות:

1. <u>סימטריות</u>: $\langle u, v \rangle = \langle v, u \rangle$
2. <u>לינאריות</u>: $\langle \alpha v, u + w \rangle = \alpha \langle v, w \rangle + \alpha \langle v, u \rangle$
3. <u>אי-ניוון</u>: $\langle u, u \rangle \geq 0$ ו-$\langle u, u \rangle = 0$ אם"ם $u = 0$.

הגדרה

נגדיר נורמה על ידי $\|v\| = \langle v, v \rangle^{\frac{1}{2}}$

הגדרה

$$\cos \theta = \frac{\langle v, u \rangle}{\|v\| \|u\|}$$

הגדרה

נאמר כי <u>v מאונך ל-u</u> ונסמן $v \perp u$ אם $\langle u, v \rangle = 0$.

הגדרה

<u>הטלה אורתוגונית של u על v</u> הוא

$$\vec{P} = p\hat{u}$$

כאשר

$$p = \langle v, \hat{u} \rangle$$
$$\hat{u} = \frac{u}{\|u\|}$$

כלומר,

$$\vec{P} = \langle v, \hat{u} \rangle \hat{u}$$

הגדרה

אם $v \in \mathbb{R}^n$ ו-$u \in \mathbb{R}^m$, אזי המכפלה החיצונית של v, u היא מטריצה מסדר $n \times n$ כך ש-

$$(v \cdot u)_{ij} = v_i u_j$$

הדרגה של מטריצה זו היא 1.

הגדרה

יהי $V\subseteq\mathbb{R}^d$, $\dim\left(V\right)=k$, $v_1,\ldots,v_k$ בסיס אורתונורמלי ל-v. אזי מטריצת ההטלה האורתוגונלית על המרחב V היא

$$P=\sum_i v_i\otimes v_i=\sum v_iv_i^t$$

(P היא מטריצה מגודל $d\times d$ מדרגה k)

מסקנה

אם $x\in\mathbb{R}^d$, אזי יתקיים

$$Px\in V$$

למשל, אם $k=1$ נקבל

$$Px=\left(v_1v_1^t\right)x=v_1v_1^tx=v_1\left\langle v_1,x\right\rangle$$

שזו בדיוק ההטלה של x על $\mathrm{span}\left\{v_1\right\}$ (כי $\left\|v_1\right\|=1$).

דוגמא לאלגוריתם למידה

נניח כי יש d משוואות בת"ל ב-d נעלמים מהצורה $y_i=x_i^tw=\sum\limits_j x_{ij}w_j$ לכל $i\in[d]$.
נתונים x,y. נרצה למצוא את w.
כלומר, נתונים $y=\begin{pmatrix}y_1\\ \vdots\\ y_d\end{pmatrix}$ ו-$x=\begin{pmatrix}\cdots & x_1 & \cdots\\ & \vdots & \\ \cdots & x_n & \cdots\end{pmatrix}$, ומטרתנו היא למצוא w כך ש-$y=xw$.
נרצה לבנות את הפונקציה הבאה:

```
def lig_reg)x,y(:
        ...
        return w
```

המשוואות בת"ל, ולכן דרגתה של x מלאה ולכן קיימת לה מטריצה הופכית ולכן נוכל להשתמש באלגוריתם המחזיר את

$$w=x^{-1}xw=x^{-1}y$$

הגדרה

בהינתן מטריצה A, <u>ללכסן את A</u> משמעותו למצוא מטריצה אלכסונית D ומטריצה הפיכה P כך ש-

$$D=PAP^{-1}$$
$$P^{-1}DP=A$$

פירוק לערכים עצמיים (Eigen Value Decomposition)

תהי $A\in\mathbb{R}^{d\times d}$ סימטרית. אזי קיימות $U,D\in\mathbb{R}^{d\times d}$ כך ש-$A=UDU^{t}$ המקיימות

1. U אורתוגונלית שעמודותיה הן ו"ע של A
2. D אלכסונית שערכי האלכסון שלהם הם הע"ע של A

דוגמא

נתבונן במטריצה $A=\begin{pmatrix}2&1\\1&2\end{pmatrix}$. נבצע לה פירוק EVD.
נמצא ע"ע על ידי מציאת שורשים לפולינום:

$$\det\left(A-\lambda i\right)=0$$

השורשים הם $\lambda_{1}=3,\lambda_{2}=1$, ולכן $D=\begin{pmatrix}3&0\\0&1\end{pmatrix}$.
כעת נותר למצוא את הוקטורים העצמיים, ונקבל

$$U=\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}\end{pmatrix}$$

אם יש לנו וקטור $x\in\mathbb{R}^{d}$ המקיים $\|x\|=1$, אזי

$$Ax=\left(UDU^{t}\right)x=UD\begin{pmatrix}x^{t}u_{1}\\\vdots\\x^{t}u_{d}\end{pmatrix}=U\begin{pmatrix}\lambda_{1}x^{t}u_{1}\\\vdots\\\lambda_{d}x^{t}u_{d}\end{pmatrix}=\sum\lambda_{i}\left\langle x,u_{i}\right\rangle u_{i}$$

פירוק לערכים סינגולריים (Singular Value Decomposition)

הגדרה

v,u הם וקטורים סינגולריים של A המתאימים לערך סינגולרי σ אם"ם $Av=\sigma u$.
במקרה זה v יקרא וקטור סינגולרי ימני ו-u יקרא וקטור סינגולרי שמאלי.

משפט ה- SVD

תהי $A\in\mathbb{R}^{m\times d}$.
אזי $A=U\Sigma V^{T}$ כאשר

- $U\in\mathbb{R}^{m\times m}$ אורתוגונלית שעמודותיה $u_{1},\ldots,u_{m}$ ו"ס שמאליים של A.
- $V\in\mathbb{R}^{d\times d}$ אורתוגונלית שעמודותיה $v_{1},\ldots,v_{d}$ ו"ס ימניים של A
- $\Sigma\in\mathbb{R}^{m\times d}$ "אלכסונית" שערכיה ע"ס של A

הערה

המשמעות של מטריצה אלכסונית היא ש-Σ היא מהצורה $\left(\begin{array}{c|c}D&0\end{array}\right)$ כאשר D אלכסונית.

משמעות

יהי $x \in \mathbb{R}^d$. נזכור כי $x = \sum\limits_i \langle x_i, v_i \rangle v_i$.

$$Ax = A\sum_i \langle x, v_i \rangle v_i = \sum_i \langle x, v_i \rangle U\Sigma V^T v_i =$$

נבחין כי

$$\left[V^T v_i\right]_j = \delta_{ij}$$

ולכן

$$V^T v_i = e_i$$

(כאשר e_i הוקטור ה-i בבסיס הסטנדרטי)
ולכן

$$= \sum_i \langle x_i, v_i \rangle U\Sigma e_i = \sum_i \sigma_i \langle x_i, v_i \rangle U e_i = \sum_i \sigma_i \langle x_i, v_i \rangle u_i$$

שאלה

כיצד נמצא את V, U?

תשובה

נניח שבחרנו בסיס אורתונורמלי כלשהו $v_1, \ldots, v_d$. נבחין כי אם נפעיל את A על הבסיס נקבל

$$Av_1, \ldots, A_{v_d}$$

אבל אין שום סיבה ש-$Av_1, \ldots, A_{v_d}$ יהיו אורתוגונליים!
נבחין כי $A^T A$ היא מטריצה סימטרית, ולכן קיים לה פירוק EVD.
נניח שיש לנו פירוק SVD.

$$\begin{aligned} A^T A &= \left(U\Sigma V^T\right)^T U\Sigma V^T = V\Sigma^T \underbrace{U^T U}_{I} \Sigma V^T = \\ &= V \underbrace{D}_{\Sigma^T\Sigma} V^T \end{aligned}$$

לכן, בהינתן A נוכל לבצע EVD על $A^T A$, ונקבל $v_1, \ldots, v_d$ בסיס אורתונורמלי ו-$\lambda_1, \ldots, \lambda_d$.
נוכל לבצע תהליך דומה על AA^T ונקבל $u_1, \ldots, u_m$ בסיס אורתונורמלי ו-$\lambda_1, \ldots, \lambda_n$.
ויתקיים

$$Av_i = \sqrt{\lambda_i} u_i$$