

$k(x, x') = e^{-\beta \sin^2(\pi|x-x'|/p)}$
 כלל ש- σ נמוך יותר יש יותר מחזורים בתחום מסוים. למעשה זה בדיוק אורך המחזור.

משפט בניית קרנלים חוקיים

- $h^T(x) \cdot h(y)$
- סכום ופכל של קרנלים
- כפל בסקלאר חיובי
- $e^{k(x,y)}$
- אם PSDA אז $x^T A y$
- $f(x) \cdot k(x, y) \cdot f(y)$
- אם g פולינום עם מקדמים אי שיליים, אז $g(k(x, y))$

משפט קרנל PD יכול להתאים ב-100% לדוגמאות אימון (כלומר לקבל 0 שגיאה) אם מניחים שאין רעש.

משפט לכל קרנל המקיים את התכונות שלו קיימות (אולי אינסוף) פונקציות בסיס מתאימות

תהליכים גאוסיאנים

הגדרה: התפלגות GP של פונקציות:
 $f \sim GP(m(x), c(x, x'))$
 אם לכל אוסף משתנים $\{x_i\}_{i=1}^l$ מתקיים:
 $f(\{x_i\}) \sim N(m(\{x_i\}), K)$

משפט: בגדול הקטע של הפרדיקציה הוא כמו קרנלים, וקיבלנו את התוצאה הבאה. אם נתון דאטאסט $D = \{x_i, y_i\}$ כאשר $y_i = f(x_i) + \eta_i$ כאשר $\eta_i \sim N(0, \sigma^2)$ ונקודה חדשה x^* אז:

$$k^* = \begin{pmatrix} k(x_1, x^*) \\ \vdots \\ k(x_l, x^*) \end{pmatrix}$$

נסמן

$$\hat{f}(x^*) | D \sim N(k^{*T} K^{-1} y, \dots \\ \dots k(x^*, x^*) - k^{*T} K^{-1} k^*)$$

אם יש רעש:

$$\hat{f}(x^*) | D \sim N(k^{*T} (K + I\sigma^2)^{-1} y, \dots \\ \dots k(x^*, x^*) - k^{*T} (K + I\sigma^2)^{-1} k^*)$$

וגם: $\hat{f}(x^*)^{MMSE} = k^{*T} (K + I\sigma^2)^{-1} y$ (זה אותו פתרון כמו לקרנל).

משפט: ה-Evidence של מודל GP:
 אם θ_k הם ההיפר פרמטרים של המודל (פונקציית קרנל, רעש), ה-Evidence הוא:
 $N(y | 0, K + I\sigma^2)$

והלוג Evidence הוא:

$$-\frac{1}{2} y^T \alpha - \frac{1}{2} \log |C + I\sigma^2| - const$$

כאשר הביטוי הראשון מתאר את הפיט, והמודל השני מתאר את ה"סיבוכיות" של המודל.

סיווג בידיאני

גישה דיסקרימינטיבית – ממדלים את $p(y = c | x, \theta)$
 גישה גנרטיבית – ממדלים את $p(x | y = c, \theta)$ ואת $p(y = c | \theta)$.

Classification as regression

(דיסקרימינטיבי): מאמנים GP על הדאטא הדיסקרטי. איך? מניחים שפלוס זה 1, ומינוס זה -1.

ואז בפרדיקציה עושים פונקציית Sign על התוצאה.

הן, לחשב עם גאוסיאן רב מימדי. (וזוה נותן ישר את ה-Evidence).

משפט חישוב Evidence אלטרנטיבי - עבור מודל הרגרסיה הרגיל:

$$p(y) = p(\theta = \mu_{\theta | D}) \cdot \dots \\ p(D | \theta = \mu_{\theta | D}) \cdot | \Sigma_{\theta | D} |^{1/2} \cdot (2\pi)^{n/2}$$

כאשר n הוא מספר הפיצ'רים. יש גורם שאומר עד כמה הנתונים מתאימים לפרדיקציה, וגורם אחר שמתעדף מודלים פשוטים. אם השונות של הפריור ממש גבוהה ה-Evidence הולך ל-0. בנוסף גם:

$$p(y) = N(y | H\mu_{\theta}, H\Sigma_{\theta}H^T + I\sigma^2)$$

משפט צורת וודברי (לפתרון BMSE רגרסיה, מאפשרת להפוך מטריצה בגודל $I \times I$ במקום $N \times N$. (אבל עבור פרדיקציה עדיין צריך לכפול בוקטור באורך N . אם רוצים להתגבר גם על זה - קרנלים)

$$\theta | D \sim N(\mu_{\theta | D}, \Sigma_{\theta | D})$$

$$M = I\sigma^2 + H\Sigma_{\theta}H^T$$

$$\mu_{\theta | D} = \mu_{\theta} + \Sigma_{\theta}H^T M^{-1}(y - H\mu_{\theta})$$

$$\Sigma_{\theta | D} = \Sigma_{\theta} - \Sigma_{\theta}H^T M^{-1}H\Sigma_{\theta}$$

קרנלים

משפט (דואליות). עבור בעיית הרגרסיה המוגדרת ע"י:

$$f_{\theta}(x) = \sum_{i=1}^n \theta_n \cdot h_n(x)$$

$$\theta \sim N(0, \Sigma_{\theta})$$

$$y_i = f_{\theta}(x_i) + \eta_i \quad \eta \sim N(0, \sigma^2 I)$$

נגדיר בעיה דואלית:

$$f_{\alpha}(x) = \sum_{i=1}^l \alpha_i \cdot k(x, x_i)$$

$$\alpha \sim N(0, K^{-1})$$

$$y_i = f_{\alpha}(x_i) + \eta_i \quad \eta \sim N(0, \sigma^2 I)$$

כאשר $k(x, x_i) = h^T(x) \cdot \Sigma_{\theta} \cdot h(x_i)$

וגם $K_{i,j} = k(x_i, x_j)$ מקיימת

1. לכל x מתקיים:
 $f_{\hat{\theta}}^{MMSE}(x) = f_{\hat{\alpha}}^{MMSE}(x)$

2. מתקיים: $\hat{\alpha}^{MMSE} = (K + \sigma^2 I)^{-1} y$

מסקנה ממשפט הדואליות. אם קיים קרנל שמקיים את התנאי לעיל עבור פונקציות הבסיס, אפשר להפוך מטריצה בגודל מספר הדוגמאות במקום מספר הפיצ'רים. אפשר גם לחשב פתרונות בזמן I במקום N .

משפט קרנל חוקי שיכול לקיים דואליות $k(x, x_i) = h^T(x) \cdot \Sigma_{\theta} \cdot h(x_i)$

מקיים:

- סימטריות: $k(x_i, x_j) = k(x_j, x_i)$
- חיוביות: $k(x_i, x_i) > 0$ (כי Σ_{θ} היא PD).
- לכל D המטריצה K היא PD.

הגדרה קרנלים נפוצים

קרנל RBF (PD):

$$k(x, x') = e^{-\beta \|x - x'\|^2}$$

כלל ש- $\beta \rightarrow 0$ זה יהיה יותר חלק ופחות צמוד ל-0. כלל ש- $\beta \rightarrow \infty$ זה יהיה צמוד ל-0 חוץ מבנקודות אימון.

קרנל לפליסאן (PD):
 כמו RBF אבל נורמה 1. חד.

קרנל פולינומי (לא PD):
 $k(x, x') = (x^T x + c)^d$

מתאים לפונקציות בסיס שהן קומבינציות וזה נראה כמו פולינום.

קרנל פריאודי (לא PD):

הגדרה משעך MAP

$$\hat{\theta}^{MAP}(D) = \arg \max_{\theta} P(\theta | D)$$

משפט $\hat{\theta}^{MAP}$ הוא אופטימלי במובן Hit-or-Miss.

זה יהיה ממוצע ממושקל של $\hat{\theta}^{MLE}$ ושל הפריור.

משפט שערוך פוסטריר תוחלת של גאוסיאן חד מימדי. כאשר $\mu \sim (\mu_0, \sigma_0^2)$ ו-

$$p(\mu | D) \sim N(\mu_N, \sigma_N^2)$$

$$\mu_N = \sigma_N^2 \left(\frac{1}{\sigma^2} \hat{\mu}_{ML} + \frac{1}{\sigma_0^2} \mu_0 \right)$$

$$\sigma_N^2 = \frac{1}{N/\sigma^2 + 1/\sigma_0^2} = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

משפט שערוך פוסטריר תוחלת של גאוסיאן רב מימדי. כאשר $\mu \sim (\mu_0, \Sigma_0)$ ו-

$$p(\mu | D) \sim N(\mu_N, \Sigma_N)$$

$$\mu_N = \Sigma_N (N\Sigma^{-1} \hat{\mu}_{ML} + \Sigma_0^{-1} \mu_0)$$

$$\Sigma_N = (N\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

גרסיה

משפט התפלגות של פוסטריר של גרסיה עם פריור:

$$\theta | D \sim N(\mu_{\theta | D}, \Sigma_{\theta | D})$$

נסמן $H_{I \times N}$ המטריצה ששורותיה $h(x_i)$

$$\mu_{\theta | D} = \Sigma_{\theta | D} \cdot \left(\frac{1}{\sigma^2} H^T y + \Sigma_{\theta} \mu_{\theta} \right)$$

$$\Sigma_{\theta | D} = \left(\frac{1}{\sigma^2} H^T H + \Sigma_{\theta}^{-1} \right)^{-1}$$

זה מוגדר היטב תמיד כי מה שבשונות הוא PD.

ומתקיים:

$$\hat{\theta}^{MAP} = \hat{\theta}^{MMSE} = \mu_{\theta | D}$$

משפט פרדיקציה אופטימלית במובן BMSE של גרסיה ליניארית:

$$\hat{y} = \mu_{\theta | D}^T \cdot h(x)$$

וניתן לאמוד את השגיאה ע"י:

$$Var(\hat{y} | D) = Var(h^T(x) \cdot \hat{\theta})$$

$$= h^T(x) \cdot \Sigma_{\theta | D} \cdot h(x)$$

כלומר לכל $\hat{\theta}$ מתקיים:

$$E_{\theta, D}((f_{\hat{\theta}}^{MMSE} - y)^2) \leq E_{\theta, D}((f_{\hat{\theta}} - y)^2)$$

משפט משעך MLE לרגרסיה:

$$\hat{\theta}^{MLE}(D) = (H^T H)^{-1} H^T y$$

וזוה מוגדר היטב רק אם מספר הדוגמאות גדול ממספר הפיצ'רים.

משפט קשר לרגרסיית רידג': למצוא את $\hat{\theta}^{MMSE}$ עם פריור $\theta \sim N(0, I\alpha^2)$ ורעש σ^2 זה

$$\lambda = \frac{\sigma^2}{\alpha}$$

כמו לפתור רגרסיית רידג' עם λ .

הפתרון הקלאסי הוא:

$$\hat{\theta}^{RIDGE}(D) = (H^T H + I\lambda)^{-1} H^T y$$

הגדרה Evidence של מודל z (עבור בחירת מודל):

$$p(D | z) = \int_{\theta} p(\theta, D | z) \cdot d\theta$$

$$= \int_{\theta} p(\theta | z) \cdot p(D | \theta, z) \cdot d\theta$$

בדוגמא בשיעור עבור רגרסיה השיטה הייתה למצוא לפי הפריור על θ את ההתפלגות

$$y \sim N(H\theta, Hcov(\theta)H^T)$$

ולחצוב בה את

רקע

הגדרה מטריצה A היא PD אם היא ריבועית וסימטרית ולכל ווקטור $x \neq 0$ מתקיים:

$$x^T A x > 0$$

תכונות:

- קיים פירוק: $A = R^T R$
- ערכים עצמיים חיוביים.
- הפיכה (אם PD). כי $\lambda_i > 0$. $|A| = \prod_i \lambda_i$

הגדרה התפלגות גאוסיאנית:

$$p = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$p = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \cdot \dots$$

$$\exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

משפט טריק הנגזרת. אם יודעים שמשוואה גאוסיאן (למשל על ידי רק שהוא

אקספוננט של פונקציה ריבועית). מסתכלים על מה שבאקספוננט וגוזרים את זה (בלי -, עם 1/2 !!!) מביאים לצורה:

$$\frac{\partial \Delta}{\partial x} = \Sigma^{-1} (x - \mu)$$

ואז יודעים כי ההתפלגות היא $N(\mu, \Sigma)$.

משפט נגזרת נפוצה:

$$\frac{\partial}{\partial x} (ax + b)^T A (ax + b) = 2aA(x + b)$$

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

שערוך

משפט חוק בייס:

$$p(x | D) = \frac{p(x) \cdot p(D | x)}{p(D)}$$

$$posterior = \frac{prior \cdot likelihood}{evidence}$$

הגדרה שגיאה ריבועית קלאסית:

$$MSE_{\theta^{GT}}(\hat{\theta}) = E_D(\|\hat{\theta}(D) - \theta^{GT}\|^2)$$

$$= bias^2(\hat{\theta}) + var(\hat{\theta})$$

$$= (E_D[\hat{\theta}] - \theta^{GT})^2 + E_D[(\hat{\theta} - E_D[\hat{\theta}])^2]$$

הגדרה שגיאה ריבועית בייזיאנית:

$$BMSE(\hat{\theta}) = E_{\theta, D}(\|\hat{\theta}(D) - \theta\|^2)$$

$$= \int_{\theta} p(\theta) \cdot MSE_{\theta}(\hat{\theta}) \cdot d\theta$$

נפוץ לעשות את התוחלת רק על D או רק על תטא, זה נכון בכל מקרה כי הם קשורים.

הכוונה ב-D זה דאטאסטים שונים, ולא ריצה על התוכן של דאטאסט יחיד.

הגדרה משעך MMSE

$$\hat{\theta}^{MMSE}(D) = E(\theta | D)$$

$$= \int_{\theta} \theta \cdot P(\theta | D) \cdot d\theta$$

משפט $\hat{\theta}^{MMSE}$ הוא אופטימלי במובן BMSE.

הגדרה שגיאת Hit-or-Miss

$$L_e(\hat{\theta}) = 0 \text{ if } \|x - y\| < \epsilon, \\ 1 \text{ otherwise}$$

הערה: זו היוריסטיקה וזה לא בהכרח אופטימלי!

LDA (גנרטיבי):

פתרון MLE: נניח כי הדאטא מגיע מ-GMMs ובנוסף נניח כי הקווריאנס של כל הגאוסיאנים זהה. כלומר הפרמטרים שלנו הם:

$$\theta = \{\Sigma, \{\mu_i, \pi_i\}_{i=1}^C\}$$

נמצא אותם לפי משעריך MLE. לאחר מכן נחשב את:

$$b_c = \log(\pi_c) - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c$$
$$w_c = \mu_c^T \Sigma^{-1}$$

ונסווג על פי:

$$\hat{y}(x) = \operatorname{argmax}_c \{b_c + w_c^T x\}$$

הערה: גבולות הסיווג ליניארים.

QDA (גנרטיבי):

פתרון MLE:

במקרה זה גם נניח כי הדאטא מגיע מ-GMM אבל לכל קלאסטר יכול להיות קווריאנס שונה. באופן דומה נמצא את הפרמטרים $\{\mu_i, \Sigma_i, \pi_i\}_{i=1}^C$ עם משעריך MLE. ואז נחשב:

$$b_c = \log(\pi_c) - \frac{1}{2} \mu_c^T \Sigma_c^{-1} \mu_c - \text{const}$$
$$w_c = \mu_c^T \Sigma_c^{-1}$$

ונסווג על פי:

$$\hat{y}(x) = \operatorname{argmax}_c \{b_c + w_c^T x - \frac{1}{2} x^T \Sigma_c^{-1} x\}$$

הערה: גבולות הסיווג ריבועיים.

פתרון בייזיאני (GDA):

נוסיף פריורים:

$$\mu_c \sim N(\mu_0^c, \Sigma_0^c)$$
$$\pi \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_c)$$

אז מתקיים:

$$p(\pi|y) \propto p(\pi) \cdot p(y|\pi)$$
$$= \text{Dirichlet}(\pi|\alpha_c + N_c)$$

על כן:

$$\hat{\pi}_c^{MMSE} = \frac{\alpha_c + N_c}{\sum_{c'} (\alpha_{c'} + N_{c'} - 1)}$$
$$\hat{\pi}_c^{MAP} = \frac{\alpha_c + N_c - 1}{\sum_{c'} (\alpha_{c'} + N_{c'} - 1)}$$

מה לגבי השערכים ל- $\hat{\mu}, \hat{\Sigma}$? שערך בייזיאני עם הנוסחא הרגילה (לכל קלאס בנפרד!).

הגדרה Bayesian Classification Error (BCE):

נסמן f_θ הפונקציה האמיתית

$$BCE(\hat{f}) = E_{\theta,D} \left(\begin{cases} 0 & \hat{f}(x) = f_\theta(x) \\ 1 & \text{otherwise} \end{cases} \right)$$

משפט: נסמן f_θ הפונקציה האמיתית. אז

המשעריך האופטימלי $\hat{f}(x)$ במובן BCE הוא:

$$\hat{f}(x) = \operatorname{argmax}_c \{p(f_\theta(x) = c|D)\}$$

נשים לב כי זה משעריך MAP וראינו שהוא אופטימלי במובן Hit-Or-Miss.

רגרסיה ללא הנחת גאוסיאניות

משפט – התפלגות פוסטריור: בהינתן

סטאפ הרגרסיה הרגיל, אבל הפריור הוא

GMM כלומר

$$\theta \sim GMM(\{\mu_i\}, \{\Sigma_i\}, \{\pi_i\}_{i=1}^K)$$

(הרעש כן מתפלג גאוסיאני) אז הפוסטריור

מתפלג כך:

$$\theta|D \sim GMM(\{\mu_i^{\theta|D}\}, \{\Sigma_i^{\theta|D}\}, \{q_i\})$$

כאשר:

$$q_i = \frac{\pi_k \cdot p(D|k)}{\sum_c \pi_c \cdot p(D|c)}$$

פתרון MMSE:

$$\hat{\theta}_{MMSE} = \sum_k q_k \cdot \mu_i^{\theta|D}$$

אבל זה לא בהכרח פתרון טוב כי יכול להיות שזה ינחת בעמק. על כן עדיף פתרון ה-MAP אבל אי אפשר לחשב אותו בקלות בלי אינפורמציה נוספת (במקרה של הקלספיקציה יודעים את הקלאסים). כאן נכנס EM שמוצא מקסימום לוקאלי ב-GMM.

דוגמא – רגרסיה דלילה

נניח כי אנו עושים רגרסיה עם הפרמטרים θ_1, θ_2 וכי הפריור הוא שרק פרמטר אחד שונה מ-0 (וכי כולם סביב 0). אז בעצם הפריור מתפלג כך:

$$\theta \sim GMM(\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0.0001 \end{bmatrix}, \begin{bmatrix} 0.0001 & 0 \\ 0 & 1 \end{bmatrix} \right\}, \left\{ \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \right\})$$

באופן כללי, אם k פרמטרים מתוך n שונים מ-0, זה יהיה GMM שעשוי מ- $\binom{n}{k}$ גאוסיאנים (ולכל אחד הסתברות שווה).

דוגמא – רגרסיה עם מדידות חריגות

$$y_i = \begin{cases} y_\theta(x_i) + \eta & w.p. 0.9 \\ N(0, 100^2) & w.p. 0.1 \end{cases}$$

אז אם יש N דוגמאות, $y|\theta$ זה בעצם GMM עם 2^N גאוסיאנים. נניח שיש שתי דוגמאות, אז הגאוסיאנים הם:

$$\pi = 0.9 \cdot 0.9 \cdot N\left(\theta, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right)$$
$$\pi = 0.1 \cdot 0.9 \cdot N\left(\begin{bmatrix} 0 \\ y_\theta(x_2) \end{bmatrix}, \begin{bmatrix} 100^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right)$$
$$\pi = 0.9 \cdot 0.1 \cdot N\left(\begin{bmatrix} y_\theta(x_1) \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 100^2 \end{bmatrix}\right)$$
$$\pi = 0.1 \cdot 0.1 \cdot N\left(0, \begin{bmatrix} 100^2 & 0 \\ 0 & 100^2 \end{bmatrix}\right)$$

אלגוריתם EM – מוצא מקסימום לוקאלי

בהתפלגות שרירותית (בפרט גם GMM). האלגוריתם לא בהכרח מגיע למקסימום הגלובלי, אבל בהכרח לא מגיע לתוצאה פחות טובה מהשלב הקודם בכל שלב.

נניח שרוצים למצוא θ שממקסם את $p(x|\theta)$ אבל זה קשה, אבל כן קל למקסם את $p(x, z|\theta)$. אז האלגוריתם הוא:

$$Q(\theta|\theta^{(t)}) = E_{z|x, \theta^{(t)}} [\log p(x, z|\theta)]$$
$$\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta|\theta^{(t)})$$

אפשר גם למקסם את $p(\theta|x)$, בשלב 1:

$$Q(\theta|\theta^{(t)}) = E_{z|x, \theta^{(t)}} [\log p(\theta, z|x)]$$

ושלב 2 אותו דבר!

EM לשערוך של GMM (MLE) – ללא

(פריור):

לכל $t = 1, \dots, T$:

שלב E: לכל i ולכל k חשב את

$$r_{i,k}^{(t)} = \frac{\pi_k^{(i)} \cdot N(x|\mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_c \pi_c^{(i)} \cdot N(x|\mu_c^{(i)}, \Sigma_c^{(i)})}$$

שלב M: לכל k חשב את

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1, \dots, N} r_{i,k}^{(t)}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1, \dots, N} r_{i,k}^{(t)} x_i}{\sum_{i=1, \dots, N} r_{i,k}^{(t)}}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1, \dots, N} r_{i,k}^{(t)} (x_i - \mu_k^{(t+1)}) (x_i - \mu_k^{(t+1)})^T}{\sum_{i=1, \dots, N} r_{i,k}^{(t)}}$$

החזר: $\theta_k^{(T)} = (\pi_k^{(T)}, \mu_k^{(T)}, \Sigma_k^{(T)})$ (לכל k).

הערה: $r_{i,k}$ זה ה-Responsibility, מייצג כמה הגאוסיאן k אחראי לדוגמא x_i . עבור הדוגמא מתקיים:

$$p(x|\theta^{(t)}) \geq p(x|\theta^{(t-1)})$$

EM למציאת פתרון MAP לרגרסיה עם פריור GMM:

נגדיר

$$Q(\theta|\theta^{(t)}) = E_{z|x, \theta^{(t-1)}} [\log p(\theta, z|x)]$$

ראינו כי הפוסטריור מתפלג

$$\theta|D \sim GMM(\{\mu_i^{\theta|D}\}, \{\Sigma_i^{\theta|D}\}, \{q_i\})$$

לכל $t = 1, \dots, T$:

שלב E: לכל k חשב את

$$r_k^{(t)} = q_k \cdot N(\theta_k|\mu_k^{\theta|D}, \Sigma_k^{\theta|D})$$

שלב M: חשב את

$$\theta_{t+1} = \left(\sum_k r_k^{(t)} \Sigma_i^{\theta|D-1} \right)^{-1} \left(\sum_k r_k^{(t)} \Sigma_i^{\theta|D-1} \mu_k^{\theta|D} \right)$$

החזר את θ_T .

עבור הדוגמא מתקיים:

$$p(\theta^{(t)}|x) \geq p(\theta^{(t-1)}|x)$$

אלגוריתמי דגימה

הגדרה – אלגוריתם דגימה ל- $\theta|D$ הוא אלגוריתם סטוכסטי שבהינתן D מחזיר ווקטור θ כך שההסתברות לקבל θ כפלט של האלגוריתם היא $p(\theta|D)$

דוגמא – אלגוריתם דגימה מפוסטריור

ברגרסיה עם פריור GMM

- שערך את התפלגות הפוסטריור כפי שראינו לעיל
 - דגום $k \sim \text{multinomial}(q)$
 - דגום $\theta \sim N(\mu_k^{\theta|D}, \Sigma_k^{\theta|D})$
- נשים לב שזה בעייתי מבחינת סיבוכיות, כי למשל ברגרסיה דלילה במימד d , נצטרך לשמור ווקטור בגודל 2^d .

אלגוריתם דגימת גיבס

יהא ווקטור z , נניח כי רוצים לדגום מ- $p(z)$ אבל לא יכולים, אבל כן יכולים לדגום מ- $p(z_i|z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_d)$

- אתחל $z^{(0)}$ כלשהו.
- לכל $t = 1, \dots, T$:

א. בחר קואורדינטה i

ב. קבע $z_i^{(t)}|z_{-i}^{(t)}$

ג. קבע $z_{-i}^{(t+1)} = z_{-i}^{(t)}$

טענה: כאשר $T \rightarrow \infty$ ו- $p(z)$ מקיים תנאי שקשור לארגודיות של שרשאות מרקוב, דגימת גיבס זה אלגוריתם דגימה תקין (על פי ההגדרה).

דוגמא – דגימת גיבס מרגרסיה דלילה

הטריק: לדגום מ- $p(\theta, k|D)$ כאשר k הוא ווקטור המציין אילו קואורדינטות דלילות. זה מוצא את מה שסביר שדליל.

דוגמא – דגימת גיבס מגאוסיאן דו מימדי

מתבסס על נוסחאות ההתפלגות המותנית של גאוסיאן דו מימדי. אם

$$p\left(\begin{matrix} x \\ y \end{matrix} \middle| \begin{matrix} \mu_x \\ \mu_y \end{matrix} \right) \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}\right)$$

אז:

$$p(x|y) \sim N\left(\mu_x - \frac{\sigma_{xy}^2}{\sigma_y^2} (y - \mu_y), \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}\right)$$

(וסימטרי הפוך עבור $p(y|x)$)

נרדומלי

הגדרה: $AIC = 2d - 2\log p(y|\theta^{ML}, x)$

משפט אם הפריור מתפלג ממשפחה P והלייקליהוד מתפלג ממשפחה Q , יש כמה זוגות של P ו- Q לא דווקא שווים כך שהפוסטריור מתפלג P .

θD	$D \theta$	θ
N	N	N
דיריכלה	מולטינום	מולטינום
GMM	N	GMM
W^{-1}	N, μ ידוע	W^{-1}

התפלגות דיריכלה יוצרת ווקטור שסכום אחד והקואורדינטות ממושקלות על פי α .

משפט הרחבה על דיריכלה בתור

Conjugate prior. אם הלייקליהוד קטגורי כאשר נצפו c_i מכל קטגוריה (מספר לא הסתברות) והפריור דיריכלה עם פרמטר α אז:

$$\theta \sim \text{Dirichlet}(\alpha + c)$$
$$\hat{\theta}_i^{MMSE} = \frac{\alpha_i + c_i}{\sum_j (\alpha_j + c_j)}$$
$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + c_i - 1}{\sum_j (\alpha_j + c_j - 1)}$$

כאשר θ הוא ווקטור ההסתברויות להיות בכל קטגוריה.