

Gaussian Bayes

שרון מלטר, אתגר 17

15 באוגוסט 2024

1 הקדמה

תחילה נזכיר התפלגות גאוסיינית (או 'נורמלית') שלמדו עלייה בקורס הסתברות.

1.1 התפלגות נורמלית

1.2 מימד אחד

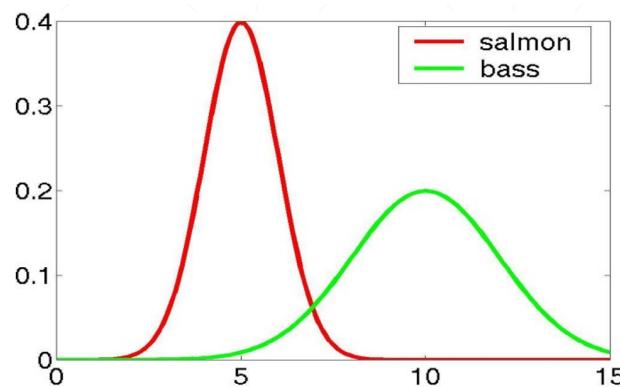
נתחיל ממשתנה רנדומלי נורמלי במימד אחד. פונקציית הצפיפות שלו היא $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ כאשר התוחלת היא μ וסטיית התקן היא σ . ואם אנחנו ממש לא זוכרים הסתברות, תוחלת היא $E[X] = \mu = \int_{-\infty}^{\infty} x \cdot p(x)dx$ כמובן, תוחלת היא הערך שאנו מוצפים לקבל עבור דוגמיה רנדומלית ממנה אנחנו למדים אחרים. איןסוף מספרים. כמו כן, שונות היא $\sigma^2 = Var(X) = E[(X - E[X])^2]$

דוגמה!
נרצה לשווג דג שאנו מקבלים לאחד מחלקות בעזרת נתון- האורך שלו. נניח שיש לנו שתי התפלגות,

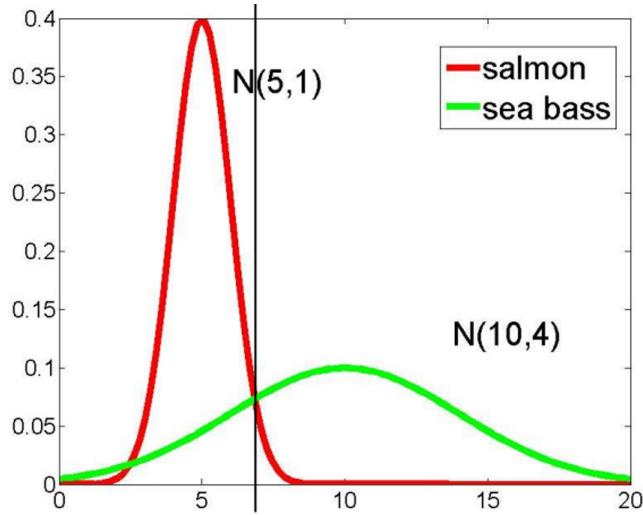
$$p(\text{length}|\text{salmon}) \sim N(5, 1)$$

$$p(\text{length}|\text{sea-bass}) \sim N(10, 4)$$

כלומר;



בדיל באופן עדין יותר בין אותן התפלגויות באמצעות מציאת נקודת חיתוך בין אותן פונקציות צפיפות;



ועלシו עליית שלבי התפלגות נורמלית ב- $2D$ ($O: 0$)

1.3 שני ממדים

בහינתן $X = [X_1, X_2]$ אנחנו רוצים לתאר את ההסתפלות של X (עבור שני הפיתרים ביחד) כמובן, אנחנו רוצים התפלגות משותפת של שני מ"מ נורמליים היא נורמלית, וכך גם התפלגות שולית והסתפלות מותנית.

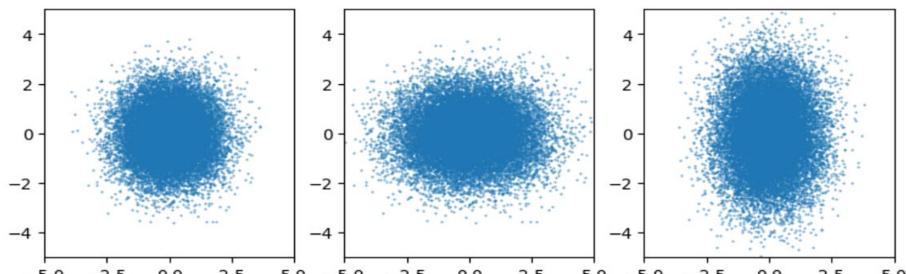
$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Cov(X_2, X_2) \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

התוחלת של התפלגות זו היא המטריצה

$$(f_X(x)) = \frac{1}{\sqrt{2\pi|\Sigma|}} \cdot e^{\frac{-1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

אם $Cov(X_1, X_2)$ היא שונות משותפת, $Cov(X_1, X_2) = E[(X - \mu_X)(Y - \mu_Y)]$. ואם גודלים ביחד, מתקיים אם יש ביניהם קורלציה שלילית אז $Cov(X_1, X_2) < 0$ ואחרת $Cov(X_1, X_2) > 0$.

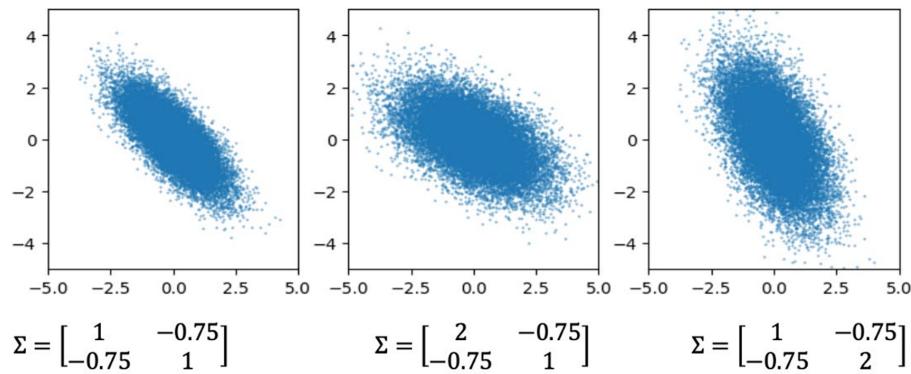
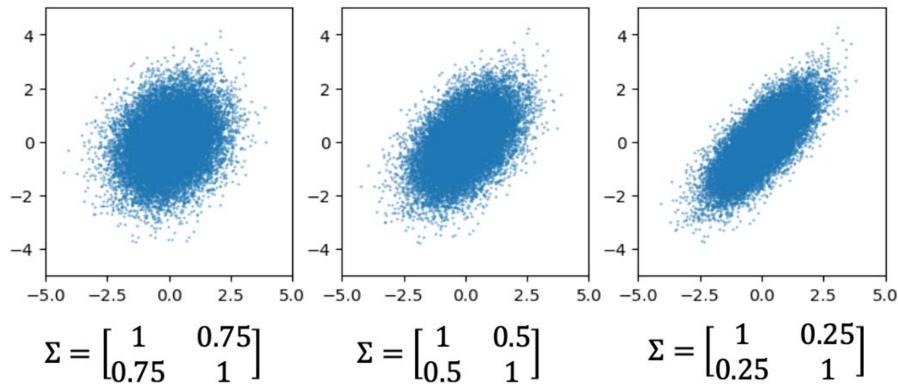
זמן טריויה!
התאימו בין הנתונים לבין מטריצות השוניות בכל שורה;



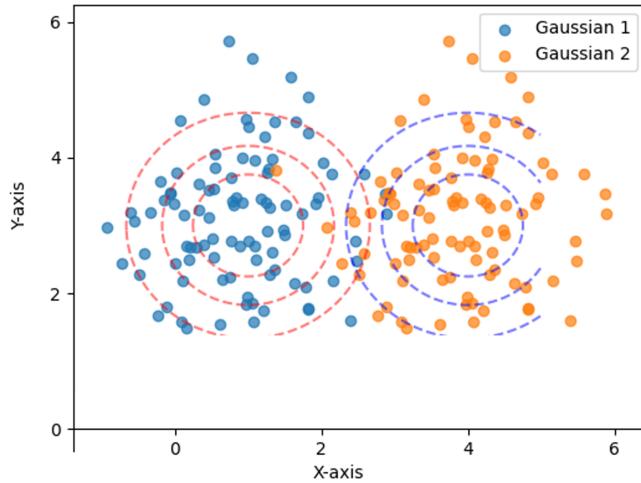
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$



שאלה שנייה: יהיו שתי מחלקות $c = 1, 2 \sim$ כך שפונקציית הצפיפות של ההתפלגות שלן היא $p(x|c=1) \sim N(\mu_1, \Sigma)$ ו- $p(x|c=2) \sim N(\mu_2, \Sigma)$ ונסמך σ .
נרצה למצוא את הישיר המפריד הטוב ביותר בינויהם.



תשובות בדף הבא.

פתרונות

1. • בגרף השמאלי ביותר הקוארדינטות האופקיות והאנכיות רוחקות ממרכז הגרף (התוחלת) באופן זהה, لكن השינויות של X_1, X_2 חייבות להיות זהות. אזי המטריצה האמצעית היא הנכונה.
- בגרף האמצעי ישנה שנותר הרבה יותר בקוארדינטות האופקיות, כך שהמטריצה הנכונה היא המטריצה הימנית ביותר.
- בגרף הימני ישנה שנותר הרבה יותר בקוארדינטות האנכיות, כך שהמטריצה הנכונה היא המטריצה השמאלית ביותר.

2. • בגרף השמאלי הקורלציה בין הנתונים היא הקטנה ביותר (הקוארדינטות האנכיות לא גדלות כל שהקוארדינטות האופקיות גדולות) כך שהמטריצה המתאימה היא הימנית ביותר.
- בגרף האמצעי מתבל החץ של הקורלציה בין הנתונים. אזי המטריצה המתאימה היא האמצעית.
- בגרף הימני הקורלציה היא הגודלה ביותר, כך שהמטריצה המתאימה היא השמאלית.

3. אותו ההיגיון של השורה הקודמת, עם קורלציה שלילית.

שאלה שנייה: יש למצוא ישר המקביל לציר האנכי כך שקוארדינטת ציר x שלו מקיימת ($p(x|c=1) = p(x|c=2)$) כך שהוא היישר המפריד.
הסיבה לכך היא שיש לה מזאה את נקודות החיתוך בין המחלקות, בתנאי שהן מתפלגות נורמלית (עבור הפיטרים שלהן)
כך שיטה זו נcona בתנאי שורצים ישר המקביל לציר האנכי, מה לגבי מקרים יותר מורכבים?

תזכורות:

- ההתפלגות של כל הפיטרים במחלקות נורמלית וכמו כן גם ההתפלגות המותנית של כל אחד מהם נורמלית.
- אם הפיטרים בלתי תלויים, אזי $Cov(X_1, X_2) = 0$ (אස"ס)
- אם הפיטרים בת"ל איזי $p((x_1, x_2)) = p(x_1) \cdot p(x_2)$
- משפט הגבול המרכזיאן:

$$\sum_{i=1}^n X_i \sim N(n\mu, n\Sigma)$$

עבור כת עדר על מספר מימדים כללי.

1.4 מספר מימדים

נסמן דוגמיה של d מימדים כך $X = [x_1, x_2, \dots, x_d]$; כלומר, פשוט יש לה d פיטרים; זיכור, מתקיים;

$$p(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} \cdot e^{\frac{-1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}$$

כך ש-

$$\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Var(X_n) \end{bmatrix}$$

באו נחקרו את המקירה, ונתהיל מהמקירה הקל, בו כל הפיטרים בלתי תלויים ($Cov(X_i, X_j) = 0$) כך ש- Σ מטריצה מסדר $d \times d$ אלכסונית ומתקיים ($p(x) = p(x_1)p(x_2)\dots p(x_d)$, קלומר;

$$p(x) = \prod_{i=1}^d \frac{1}{\sigma_i \sqrt{2\pi}} \cdot e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$$

ניתן למצוא את מטריצת השוניות גם כך- $E[(X - \mu)(X - \mu)^t] = \Sigma$. لكن בדרך זו ניתן למצוא את השוניות הפרטיות של כל פיטר ואת סוג ה תלות (קורלציה, אנטיקורלציה, או אין תלות) של זוגות הפיטרים. כמו מכון מיפוי פשט מפה נוכל לקבל ש- $\mu\mu^t + \Sigma = E[XX^t]$.

אבל אנחנו חוגגים יותר מדי. פונקציית הצפיפות שלנו תליה בכך ש- Σ הפיכה, אך למה יהיה המצב? אנחנו תלויים בכך ש- $\Sigma = 0$ קלומר בהאם יש שורה או טור של אפסים או האם ישם שני שורות או טורים שהווים תלויים לינארית (או שורה / טור ייחד שתלוי לינארית בשאר) או מאחר ש- $E[XX^t] = \Sigma + \mu\mu^t$ לא הפיכה אם $\mu = 0$.

אז מה עושים כשהמטריצות לא הפוכות? - מוחקים פיטרים תלויים (כחול מהתכנה מראש) הם בכל מקרה לא מסיימים לסייע הנתונים, מכיוון שיש להם קורלציה עם פיטר אחר. ככל שהקורלציה ביןיהם גדולה בערךה המוחלט, כך הם שוקלים יותר.

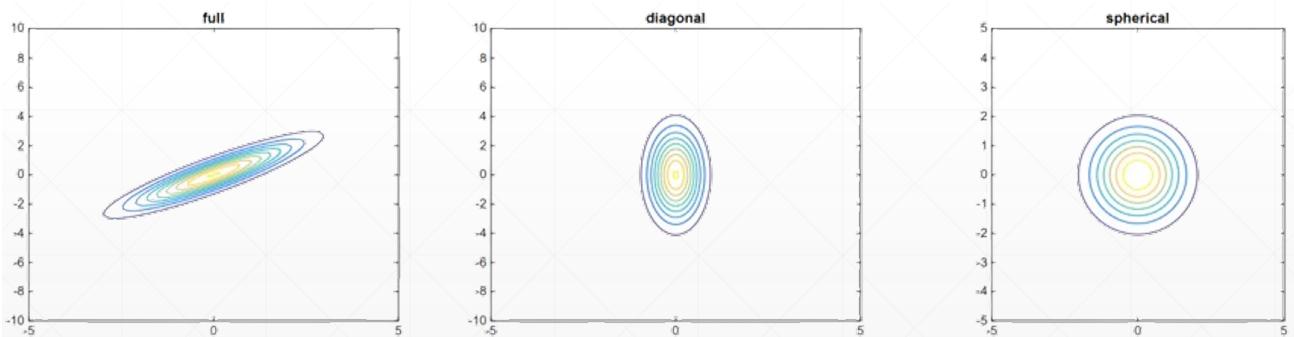
1.5 שונות וקורלציה

הגדרה המתמטית של קורלציה היא;

$$\rho(X_j, X_i) = \frac{Cov(X_i, X_j)}{\sigma_i \cdot \sigma_j} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_i \Sigma_j}}$$

כלומר, היא זהה ל- $\rho(X_i, X_j)$ חוץ מכך שהיא נמצאת בתחום $[-1, 1]$. אם $\rho(X_i, X_j) = 0$ אז הפיטרים בת"ל. ככל שיש יותר קורלציה בין פיטרים כך מתקבל $X_i = \alpha X_j + \beta$ ו- Σ מתקרבת להיות סינגולרית (מטריצה שהDETרמיננטה שלה היא 0)

ובודק האם הבנתם את המשוג' מה ניתן להגיד על מטריצות השונות והקורלציה של הנתונים בגרפים הבאים?



תשובות בעמוד הבא.

פתרון

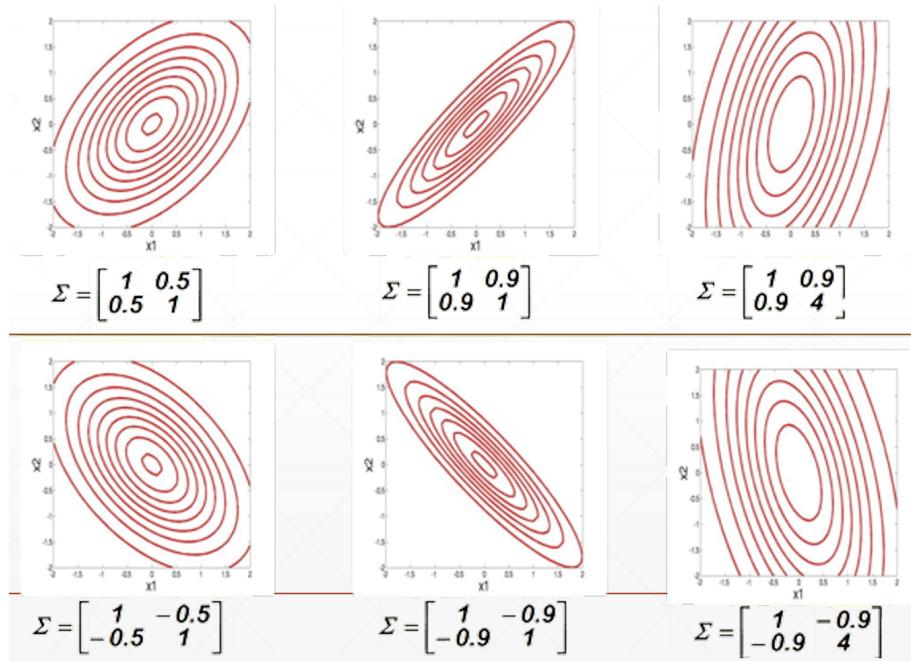
בגרף השמאלי הקוארדינטות האנכיות גדולות כל שעולים בקוארדינטות האופקיות, לכן הקורלציה בין הנתונים קרובת ל-1. נסיק גם כי השונות ביןיהם קטנה מכיוון שהקוארדינטות האופקיות מתרחקות ממרכז הגרף יותר מהאנכיות. באוטה התבוננות ניתן לקבל כי הקורלציה שואפת ל-0 בגרף האמצעי והשונות גדולה, ובגרף הימני הקורלציה גם שואפת ל-0 והשונות שואפת היא 0 (הקוארדינטות האופקיות והאנכיות באוטה מרחק מן המרכז)

1.6 משפט על גאוסיאן עם מספר משתנים

קיימת מטריצה M כך ש-

$$(x - \mu)^t \cdot \Sigma^{-1} \cdot (x - \mu) = |M(x - \mu)|^2$$

כל הנקודות x עם אותו ערך של הביטוי $|M(x - \mu)|^2$ נמצאות על אותה אליפסה.
להלן מספר דוגמאות לכך;



עד כאן ההקדמה על התפלגות נורמלית :

MLE – Likelihood Maximum
 נניח שיש לנו דוגמיות x_1, \dots, x_n ולכל אחת d פיטרים. מישו אמר לנו שההתפלגות של הדאטה היא נורמלית (!) איך נוכל לשערך את התוחלת וסטיית התקן? (O: $\hat{\mu}$: $\hat{\Sigma}$)
 בעזרת $(\text{maximum likelihood})$!MLE נחשב את הערך עם MLE של התוחלת וסטיית התקן.

$$L(\mu) = \prod_{i=1}^n -\frac{1}{2}(x_i\mu)^t \cdot \Sigma^{-1} \cdot (x_i\mu) - \frac{1}{2}n \cdot \ln|\Sigma|$$

$$\frac{dL}{d\mu} = 2\Sigma_{i=1}^n \Sigma^{-1} (x_i - \mu) = 0$$

$$\Sigma^{-1} \sum_n^{i=1} x_i = n\Sigma^{-1}\mu$$

↓

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^t$$

אך ניתן לבצע חישוב מהיר יותר אם מחשבים את הביטוי השקול
 $\hat{\Sigma} = \frac{1}{n}(X - \mu)^t(X - \mu)$ וכך סיימנו את הנושא של התפלגות נורמלית.

או שלא!

(אתם בחצי הדרך)

2 שיעורן פרטמרים

בහינתן דוגמית x , נרצה להחליט מהי המחלקה אליה היא שייכת מבחן $\{w_1, w_2, \dots, w_c\}$ כאשר c הוא מספר המחלקות. נרצה לבחור את המחלקה w_i שמקסמת את $P(w_i|x)$, כלומר, לפי חוק ביס (bayes). הערך $P(x|w_i) \cdot P(w_i) / P(x)$ זהה לכל מחלקה, שכן אנו מעוניינים במינימום של $P(x|w_i) \cdot P(w_i)$. עם זאת, כדי להשתמש בחוק ביס, אנחנו צריכים לדעת את הסתברות הכללית לכל מחלקה, $P(w_i)$, ואת הסתברות המותנית $P(x|w_i)$. אבל, מה אם אנחנו לא יודעים את הסתברויות האלה?

שיטות *parametric density* מציאות להניח שאנו יודעים את צורת ההתפלגות (זאת אומרת, יוניפורמיית או נורמלית), אבל לא את **פרמטרים** שלה. כדי לשער אוטם, יש שתי שיטות עיקריות;

MLE.

Bayesian Estimation.

בדוק כפי שעשינו בקורס הסתברות!
נתחיל מהמקרה בו המחלקות בת"ל.

2.1 סידור הנתונים

בשתי השיטות נדרש קודם לסדר את הדאטה. תחילת, נסמן ב- $x_{i,k}^j$ את הפית'ר ה- k של הדוגמית ה- i של המחלקה w_i . כמו כן נסמן שיש לנו n דוגמיות אימון בסך הכל, ו- n_i דוגמיות למחלקה w_i . כך נלמד על ההתפלגות של כל מחלקה בנפרד.

2.2 שיעורן פרטמרים עם *MLE*

נניח שאנו יודעים את הצורה של ההתפלגות $P(x|w_i)$ ושיהיא נורמלית עם תוחלת μ וסטיית תקן σ . נציג כיצד משערכים את μ לפי השיטה (שיעורון סטיית התקן נעשה באופן דומה) בהינתן דוגמיות מהמחלקה, נמצאת μ שמקסם את הסתברותות של אותן דוגמיות בה. הסתברות זו היא;

$$P(D|\mu) = P(x_1|\mu) \cdot p(x_2|\mu) \cdots P(x_n|\mu)$$

נמצא את μ המתאים בעזרת *MLE*:

$$\ln(L(\mu)) = \sum_{k=1}^n \ln(P(x_k|\mu))$$

$$\frac{d(\ln(L))}{d\mu} = \sum_{k=1}^n \left(-\ln(\sqrt{2\pi\sigma}) - \frac{(x_k - \mu)^2}{2\sigma^2} \right)$$

$$\frac{d\ln(L)}{d\mu} = \sum_{k=1}^n \left(-\ln\sqrt{2\pi\sigma} - \frac{(x_k - \mu)^2}{2\sigma^2} \right)$$

$$\frac{d\ln(L)}{d\mu} = \sum_{k=1}^n \frac{1}{\sigma^2} (x_k - \mu) = 0$$

↓

$$\sum_{k=1}^n x_k - n\mu = 0$$

↓

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k$$

אוCut אנו יודעים כיצד לשערך את התפלגות המחלקות. אך מה לגבי השיעורן השני שיש לעשות, להסתברות $P(w_i)$?
נשערך את $(P(w_i), MLE)$, באופן מפתיע, עם

$.P(w_i) = \frac{\alpha}{100}$ אם מספר הדוגמויות ששייכות ל- w_i הוא α אחוזים מכלל הדוגמויות, אוⁱ ניתן אכן להוכיח שמדובר בשיעורן הטוב ביותר ביותר לפי MLE .

המחשבה הראשונה היא, אם מספר הדוגמויות ששייכות ל- w_i נסמן $= \theta$ ו- $z_k = \begin{cases} 1 & x_k \in w_i \\ 0 & x_k \notin w_i \end{cases}$ נטען אכן $P(z_1, \dots, z_n | \theta)$ שאהה ערך שמיקסם אותה. הפונקציה $P(z_1, \dots, z_n | \theta) = \prod_{k=1}^n P(z_k | \theta)$

$$P(z_1, \dots, z_n | \theta) = \prod_{k=1}^n \theta^{z_k} (1 - \theta)^{1-z_k}$$

נשים את הפונקציה בלבד כדי להקל על הגירה.

$$\ln(P(z_1, \dots, z_n | \theta)) = \sum_{k=1}^n z_k \cdot \ln(\theta) + (1 - z_k) \cdot \ln(1 - \theta)$$

$$\frac{d\ln(P(z_1, \dots, z_n | \theta))}{d\theta} = \sum_{k=1}^n \frac{z_k}{\theta} - \frac{1 - z_k}{1 - \theta} = 0$$

$$\frac{1}{\theta} \Sigma - \frac{n}{1 - \theta} + \frac{1}{1 - \theta} \Sigma = 0$$

$$\hat{\theta} = \frac{\sum_{k=1}^n z_k}{n}$$

¶

$$P(w_i) \approx \frac{n_i}{n}$$

כנדרש :

הערה: הדוגמה שראינו מניחה שצורת התפלגות המחלקות נורמלית, אך ניתן באופן דומה להניח כי צורת ההתפלגות יוניפורמת ולהשתמש באויה שיטה.

Cut, לאחר שייערכנו את הפרמטרים, נסוויג כל דוגמיה x למחלקה w_i שמיקסמת את $P(w_i | x) = \frac{P(x|w_i) \cdot P(w_i)}{P(X)}$ אבל חפרנו על הגישה זו מספיק, מה לגבי שיעורן פרמטרים לפי ביס?

2.3 שיעורץ פרמטרים עם Naive Bayes

נתחיל מוגירה הנאיית. ומה נאיבית? בקרוב. נניח שיש לנו דוגמיה עם m פיטרים, $(x_1, \dots, x_m) = X$, ואנו רוצים לדעת אם היא שייכת למחלקה a או b (סיווג ביןאי). נסמן את המחלקה ב- \hat{Y} ואנו מניחים שככל הפיטרים הם בת"ל בהסתברות מותנית עברו שייכות לכל מחלקה (בלתי תלויים אחד בשני עבור כל מחלקה) מדובר בගירסה הנאיית בדיק בಗל הנחה זו. נסמן ב- \hat{Y} את שיעורץ המחלקה שלנו. אנחנו רוצים לבחור מחלקה שמקסמת את $P(X, Y)$ (הסתברות המקסימלית של הדוגמיה והסיווג הנבחר לה) ככלمر למקסם את $P(Y|X) \cdot P(X)$, מאחר שכמצון $P(X)$ קבוע לכל מחלקה. נתחיל מלחקר את $P(X|Y)$, ולהנות מהנוחות שספקת את

$$P(X|Y) = P(x_1, x_2, \dots, x_m|Y) = \prod_{i=1}^m P(x_i|Y)$$

כלומר, יש למצוא את הסתברות למחלקה כלשהי לפי ערך של פיטר. חזרנו למידה של המחלקות בנפרד: איך עושים זאת? - כפי שמצאנו את $P(w_i)$ בשיטה הקודמת. נמצא את אחוז הפיטרים שמשווים לכל מחלקה.

אבל לא תמיד הפתרון כל כך פשוט. מה אם ישנו ערכי פיטרים שלא נמצא בנתוני האימון? הסתברות שלהם עבור כל מחלקה תהיה 0 אוטומטית ובעצם נתעלם מהם לחילוץ. **נוסיף α (ערך רילקסציה) למודל הבסיס;**

$$\hat{P}(x_i|w_j) = \frac{N_{x_i, w_j} + \alpha}{N_{w_j} + \alpha d}, \quad (i = (1, \dots, d))$$

כasher;

Lidstone smoothing ו- **$\alpha < 1$ בגירסת Laplace smoothing** •

N_{x_i, w_j} זהו מספר הפעם שהפיטר x_i מופיע בדוגמאות שבמחלקה w_j . •

N_{w_j} זהו המספר הטוטאלי של פיטרים במחלקה w_j . •

d הוא המינד של וקטורי הפיטרים $X = [x_1, \dots, x_d]$. •

בקורס שלנו השתמש ב- *Laplace smoothing*

עבור כתע למודל האחרון והמסובך ביותר, *Gaussian Bayes*,

3 פונקציות דיסקרימיננטה

לכל מודל החלטות קיימות פונקציות דיסקרימיננטה, כך שהמחלקה שנבחרת לסיווג היא המחלקה w_i עבורה $i \neq j$, $g_i(x) > g_j(x)$. למשל;

- חוק ההחלטה של ML : $g_i(x) = P(x|w_i)i$
- חוק ההחלטה של MAP : $g_i(x) = P(w_i|x)$
- חוק ההחלטה של $Bayes$: $g_i(x) = -R(c_i|x)$ מדבר בפונקציית $Risk$, שלא נכללות בחומר הקורס (:

שימוש לב שלפי הגדרת חוק הסיווג של MAP , זהו המסוג עבוריו מתקבלת הסתברות השגיאה הקטנה ביותר.

כפי שראינו מוקדם, פונקציית הדיסקרימיננטה של בייס נאיבי היא $P(x|w_i) \cdot P(w_i)$ או $P(x|w_i) \cdot lnP(x|w_i) + ln(P(w_i))$ כתוב, נתרגם גישה זו ל- $gaussian bayes$ (הגירסה הלא נאיבית)

Gaussian Bayes 4

נתחיל מהמקרה של *naive gaussian bayes*. כמו בהקדמה, נניח ש- Σ , מטריצת השוניות של הנתונים, ניתנת להפיכת. כפי שראינו, ניתן ממטריצה זו למצוא את סוג התרומות שבין הפיטרים. נניח **שהתרומות כל הפיטרים היא גaussiana**. אזי אין בין הפיטרים j, i קורלציה אס"ם מתקיים $0 = \Sigma_{i,j}$. בסך הכל נקבל שאם $I = \Sigma_c$, כאשר Σ_c היא מטריצת השוניות של הדוגמויות ששתייכות למחלקה c , אזי אין בין הפיטרים $d, 1, \dots, n$ קורלציה. ולכן;

$$P(x|c) = \prod_{i=1}^d \frac{1}{\sigma_i \sqrt{2\pi}} \cdot e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$$

נקרא לכך **naive gaussian bayes** כאשר ההבדל בין *naive bayes* היא שכן אנו יודעים כי **הפיטרים מתרוגים נורמלית (לא רק המחלקות)**. כמו כן, אנו מניחים כי $P(x|c_i) \approx N(\mu_i, \Sigma_i)$ (שהתרוגות הדוגמויות במחלקה זהה להתרוגות המחלקה עצמה) את הפרמטרים של התרוגויות המחלקות נבחר באמצעות MLE , לפי הנוסחאות שפיתחנו. כמו כן נבחר $P(w_i) = \frac{1}{\sqrt{2\pi|\Sigma_j|}} \cdot e^{-\frac{1}{2}(x - \mu_j)^t \cdot \Sigma_j^{-1} \cdot (x - \mu_j)}$. מאחר ש- $\frac{n_i}{n}$ ומתקיים ונקבל שפונקציית הדיסקרימיננטה לכל מחלקה w_i היא;

$$g_i(x) = lnP(x|c_i) + lnP(c_i)$$

↓

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \cdot \Sigma_i^{-1} \cdot (x - \mu_i) - \frac{1}{2}ln|\Sigma| + lnP(c_i)$$

וזהו המקרה הכללי. בקרוב נלמד עליו עוד קצת, בהקשר של המשפט שלמדו על התרוגות נורמלית, אך קודם נלמד שני טריקים שניתן לביצוע בשני מקרים מסוימים.

4.1 מקרה מיוחד ראשון

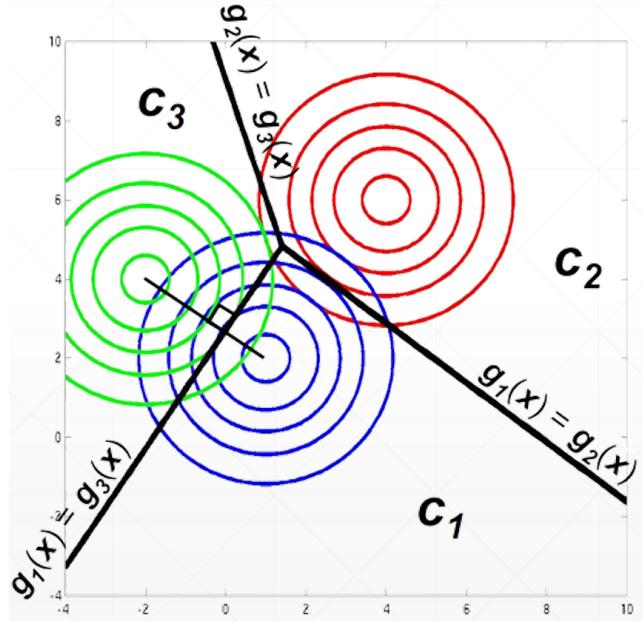
במקרה בו $I = \Sigma_1 = \dots = \Sigma_c = \sigma^2 I$ כולם כשל הפיטרים בלתי תלויים ולמחלקות יש את אותה השונות, מתקיים;

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \cdot \Sigma_i^{-1} \cdot (x - \mu_i) - \frac{1}{2} + ln|\Sigma| + lnP(c_i)$$

$$g_i(x) = \frac{\mu_i^t}{\sigma^2} x - \frac{\mu_i^t \mu_i}{2\sigma^2} + lnP(c_i)$$

$$g_i(x) = \frac{\mu_i^t}{\sigma^2}x + \left(-\frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(c_i)\right)$$

נקבל שפונקציות הדיסקרימיננטה לינאריות, $g_i(x) = w_i^t + w_0$, כך שנית לסוג דוגמאות בעזרת מיקום יחסית אליהן. לדוגמה;



4.2 מקרה מיוחד שני

כאשר מותקים $\Sigma = \sum_i$, כלומר מטריצות השונות של כל מחלקה זהות (עם הדוגמאות שישיכות לאוותן מחלקה) אז, ניתןפשט את הfonקציה

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \cdot \Sigma_i^{-1} \cdot (x - \mu_i) - \frac{1}{2}\ln|\Sigma| + \ln P(c_i)$$

$$g_i(x) = \mu_i^t \Sigma_i^{-1} x + \left(\ln P(c_i) - \frac{1}{2}\mu_i^t \Sigma_i^{-1} \mu_i\right)$$

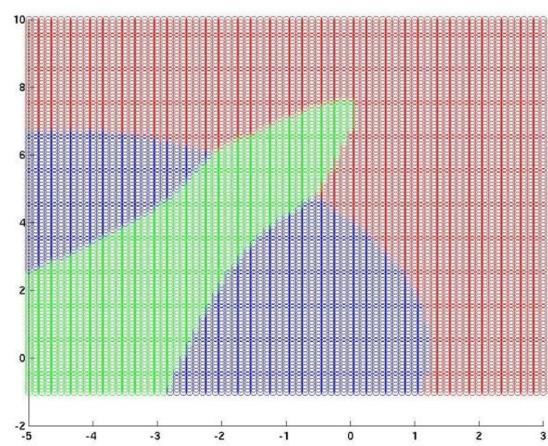
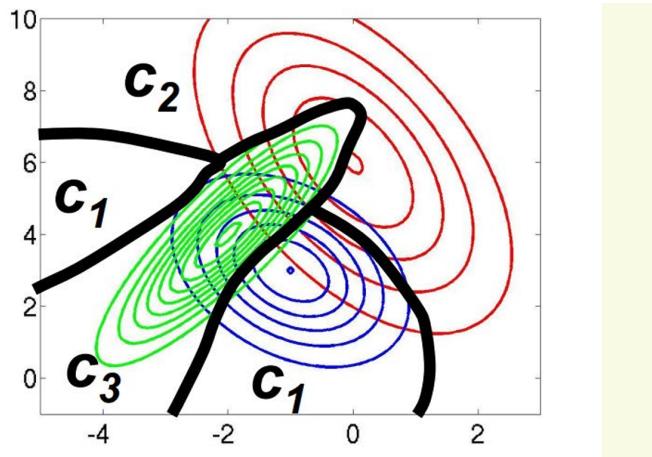
וקיבלנו שוב פונקציה לינארית :

4.3 בחזקה למקירה הכללי

במקרה זה, כפי שראינו, מתקיים:

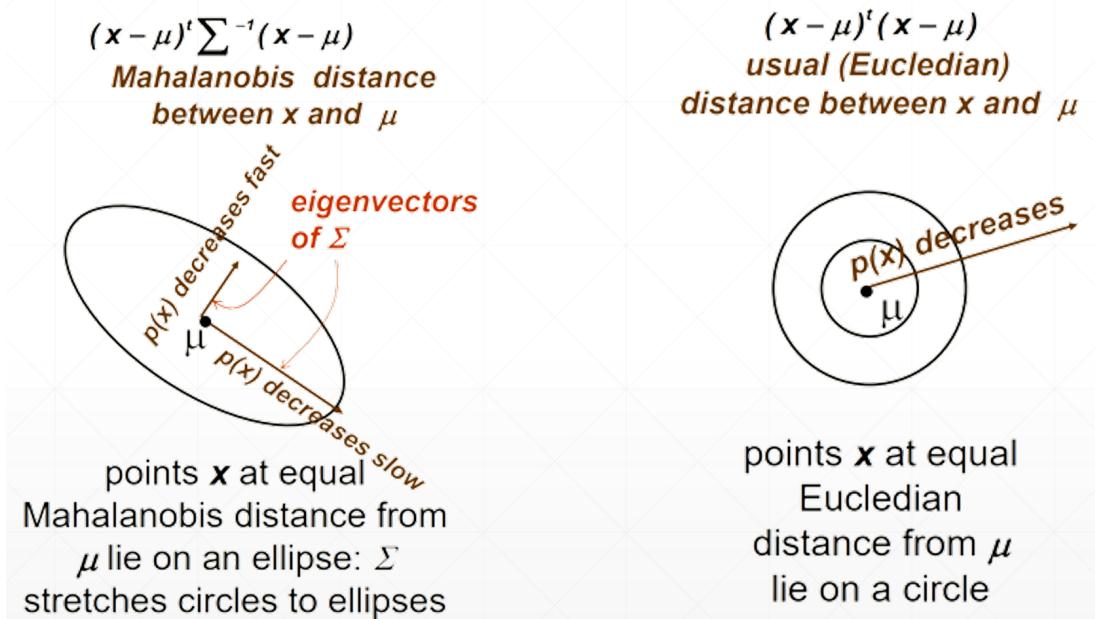
$$g_i(x) = -\frac{1}{2}(x - \mu_i)^t \cdot \Sigma_i^{-1} \cdot (x - \mu_i) - \frac{1}{2}\ln|\Sigma| + \ln P(c_i)$$

כלומר $x^t W x + w^t x + g_i(x) = x^t W x + w^t x + \text{คง} \cdot \text{שגבולות ההחלטה (decision boundaries)}$ (כלומר הן אליפסות ופרבולואידים) לדוג', כמו כן, למדנו את המשפט האומר כי קיימת מטריצה M כך ש-



$$(x - \mu)^t \cdot \Sigma^{-1} \cdot (x - \mu) = |M(x - \mu)|^2$$

וכל הנקודות x שהערך $|M(x - \mu)|^2$ שלහן זהה, נמצאות על אותה אליפסה. לכן, במקרה זה עדיף להשתמש במרחב Mahalanobis, אשר מותאם לאפליפסות. להלן תרשימים המשווה בין שימוש במרחק אוקלידי ושימוש במרחק מהלבונייס; בסך הכל קיבלנו שילוב גבולות ההחלטה הן ריבועיות, ובמקרים מיוחדים לינאריות (:



Scailing 4.4

אלגוריתמי סיווג שבוססים על מרחק (כגון KNN , SVM , ...) מושפעים מאוד מהתוחום (*range*) של הפיטרים. כך גם אלגוריתמים המבוססים על משקל כגון בייס גאוסיאני ורגסיה לוגית.

אך זה לא המצב הרצוי. נרצה שהאלגוריתמים יתиיחסו באופן זהה לתחומים שונים של משקלים או מרחקים בין פיט'רים. כלומר, אם למשל ההבדל בין שני פיט'רים מסוימים הוא 100 יח' וההבדל בין שניים אחרים מסוימים הוא 1 יח', נרצה ששני הבדלים יהיו שקולים עבור האלגוריתם אם הם הבדלים ממוצעים עבור אותן פיט'רים. שתי טכניקות מוכנות לכך – *standartization* ו- *scaling min – max*. נלמד בעת על הגישה הבסיסית יותר, *standartization*.

Standartization 4.4.1

סטנדרטיזציה היא שיטה בה מרכזים את הערכים כך שהتوزולת שלהם תהיה 0 וסטיית התקן 1. שיטה זו יעילה עבור אלגוריתמי למידה ממוכנת כגון רגסיה לינארית ורגסיה לוגית שמניחות כי הדאטה מתפלגת נורמלית.

נינתן בפועל *standartization* כך;

$$x_{scaled} = \frac{x - \mu}{\sigma}$$