זיהוי תבניות

זיהוי תבניות

לדוגמא: זיהוי פונמה; בהינתן אות דיבור, צ"ל איזו פונמה זו.

נפריד (ככל האפשר) את בעיית הסיווג לשתי בעיות נפרדות:

- חילוץ מאפיינים (feature extraction):

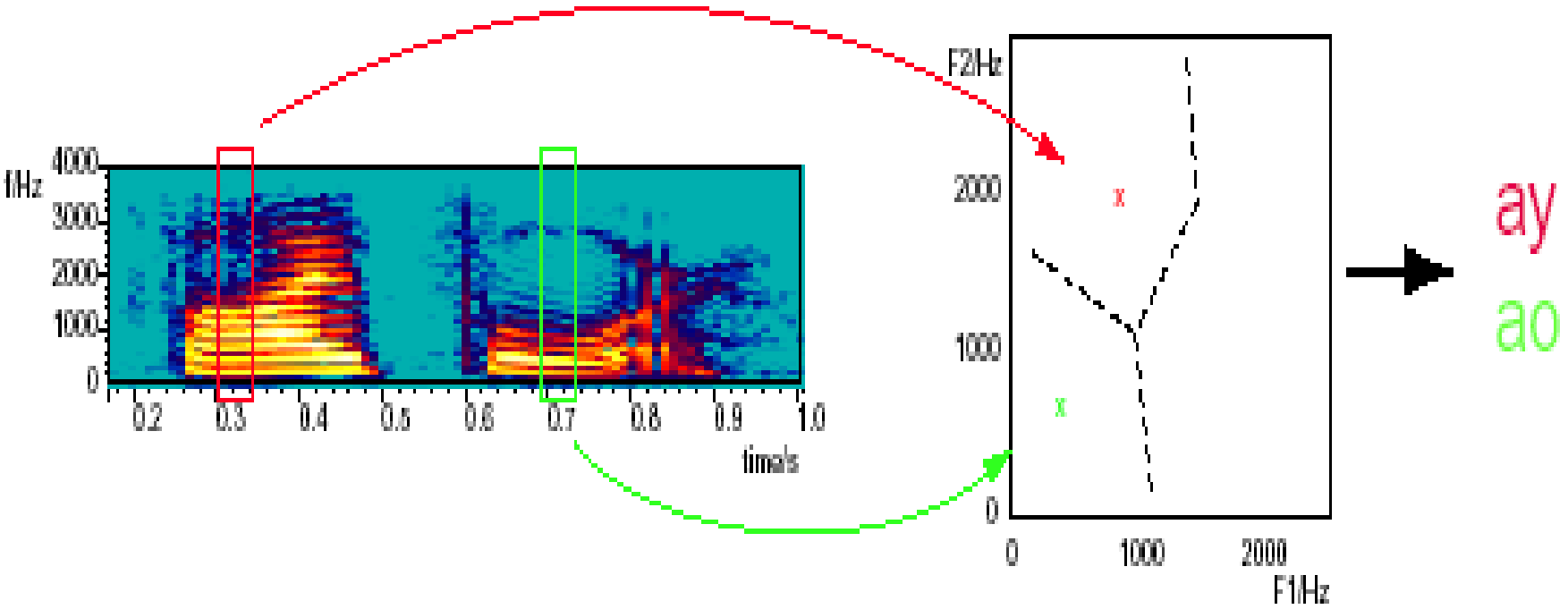
  פעולה הממירה כל תצפית לווקטור במרחב המאפיינים.

- סיווג (classification):

  פעולה הממירה כל ווקטור במרחב המאפיינים למחלקה (class)

  המתאימה ביותר.

סיווג: מציאת תיוג דיסקרטי לתצפיות רציפות



בניית מסווג

הגדרת מחלקות (classes) – פלט רצוי

לדוגמא: הפונמות שונות.

הגדרת מרחב המאפיינים (feature space) – קלט נתון

לדוגמא: התדרים של הפורמנט הראשון והשני.

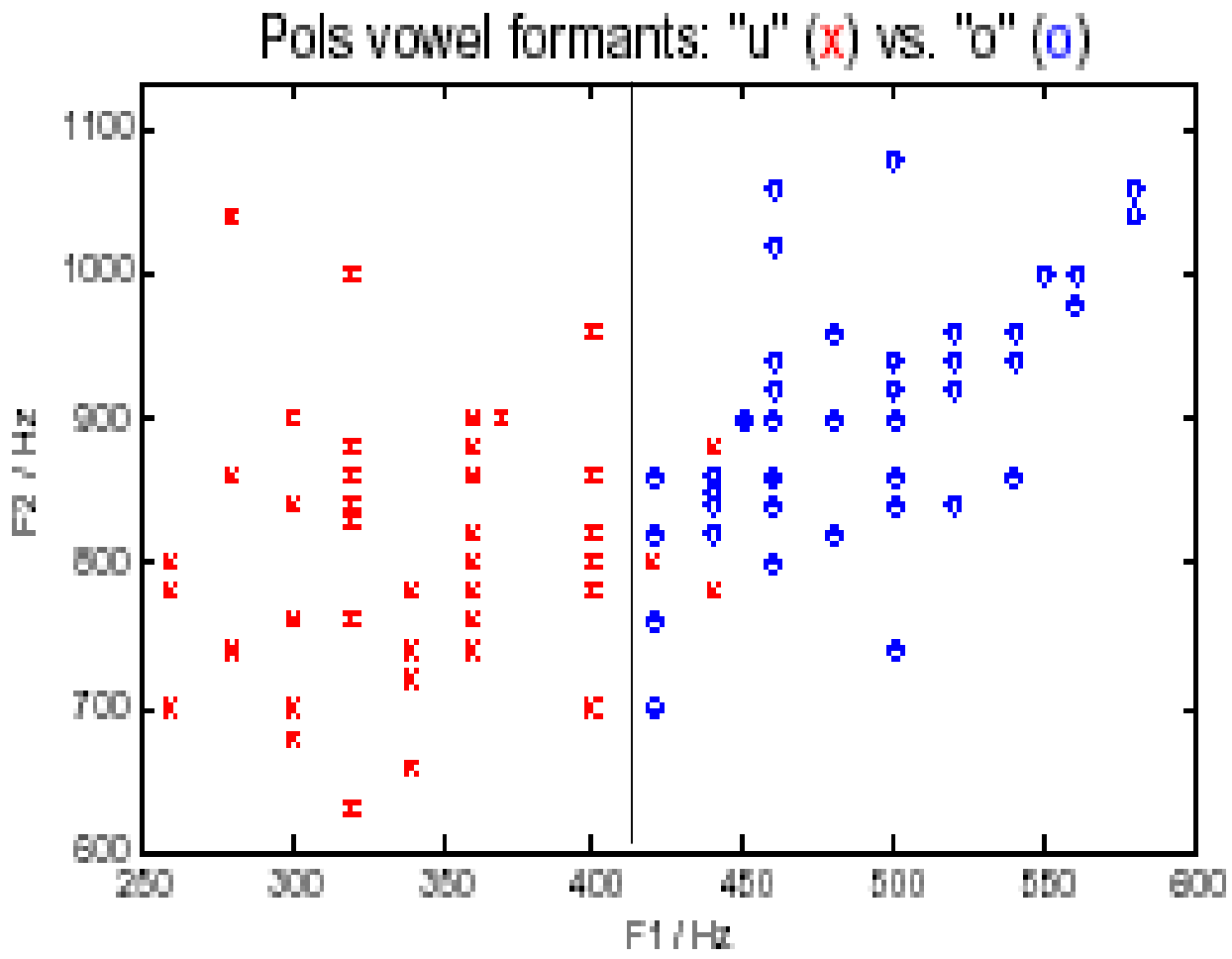
הגדרת אלגוריתם ההחלטה – הפונקציה המתאימה לכל פלט קלט

- אימון (עפ"י קבוצת אימון).

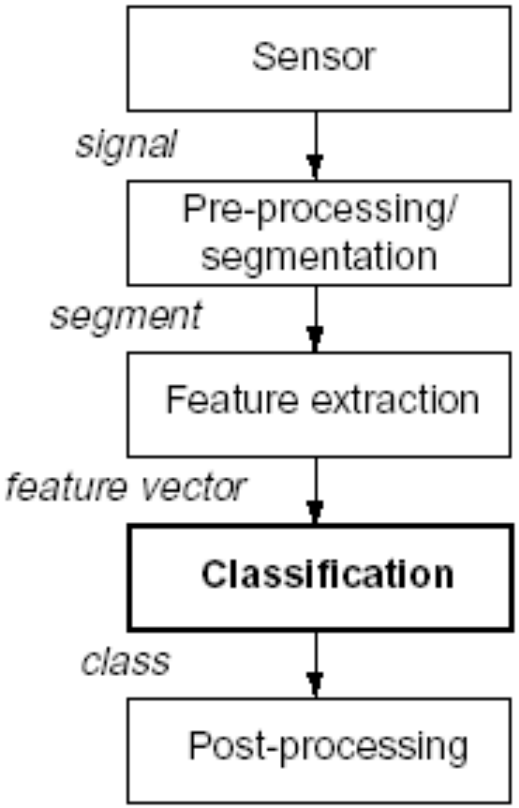
- סיווג.

מדידת ביצועים

מדידת אחוז שגיאה עבור קבוצת מבחן.


Pols vowel formants: "u" (x) vs. "o" (o)
F2 / Hz
F1 / Hz
1100
1000
900
800
700
600
250
300
350
400
450
500
550
600

Classification systems



- STFT
- Locate vowels

- Formant extraction

- Context constraints
- Costs/risk

חילוץ מאפיינים – Feature Extraction

בחירת המאפיינים היא קריטית להצלחת הסיווג.

מה מגדיר מאפיין טוב ?

- מכיל את מירב האינפורמציה הרלוונטית.
- מכיל את האינפורמציה הרלוונטית בצורה פשוטה (תחת מטריקת המרחב).
- איוריאנטי תחת טרנספורמציות לא רלוונטיות (תחת מטריקת המרחב).
- לא רגיש לרעש (תחת מטריקת המרחב).
- קל (או לפחות אפשרי) לחישוב.

מאפיינים אפשריים: האות המקורי, הספקטרום, פורמנטים.

חילוץ מאפיינים – קללת המימדיות

Curse of dimensionality: ככל שמימד מרחב המאפיינים גבוה יותר, כמות דוגמאות האימון הדרושה גדולה יותר (באופן אקספוננציאלי במימד).

מימד מרחב המאפיינים קטן מאפשר:

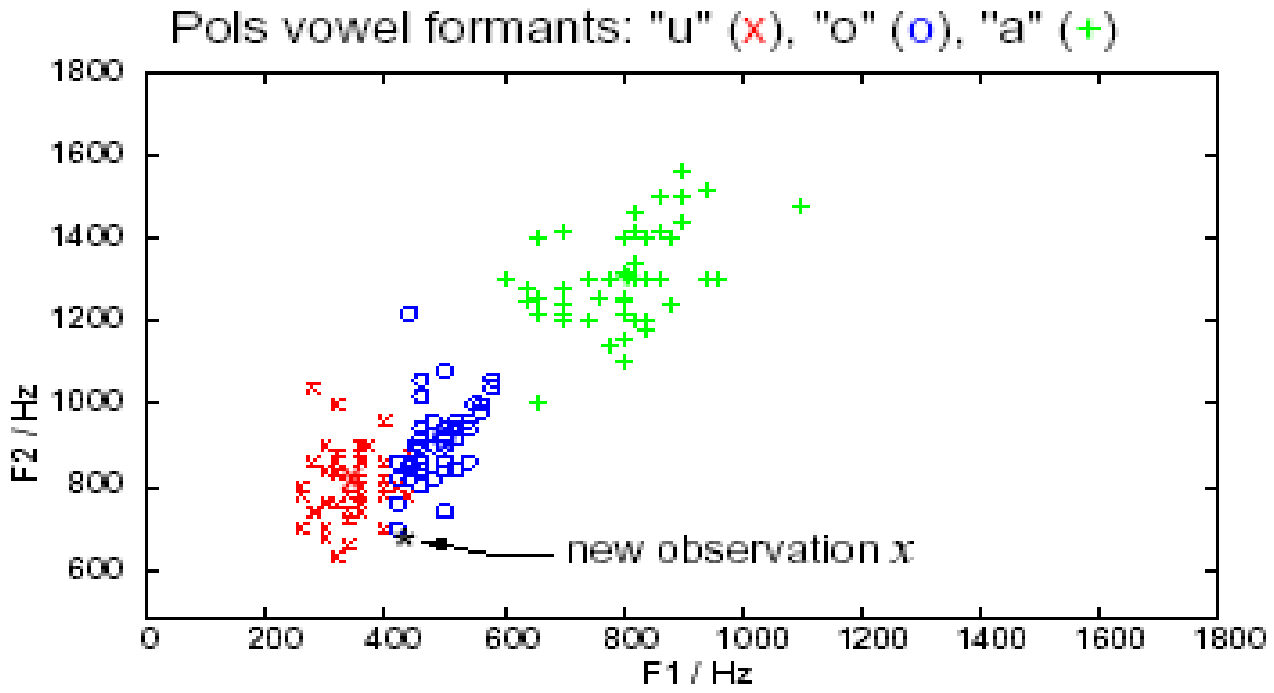
- מודלים פשוטים וקטנים יותר.
- פחות דוגמאות אימון דרושות.
- אימון מהיר יותר.

מסווג פשוט: Minimum distance classification.

האלגוריתם: מצא את דוגמת האימון הקרובה ביותר במרחב.

בחירת המטריקה חשובה ביותר.

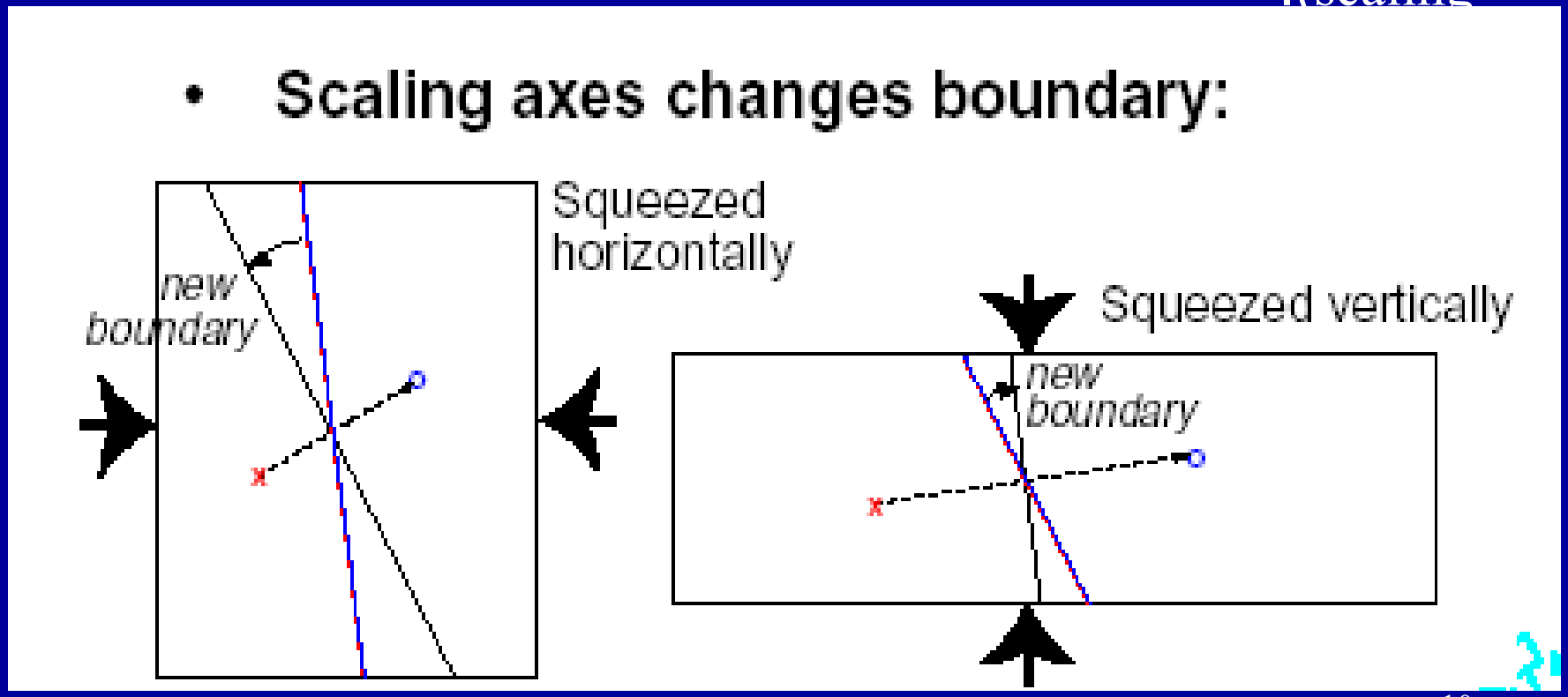
הקטנת דרישות CPU וזכרון: השוואה רק עם נציגים מקבוצת האימון.



חסרונות: רגישות לרעש, אין הכללה. פתרון אפשרי: k-nearest neighbor

מטריקה אפשרית: מרחק אוקלידי.

חסרון מהותי: המרחק האוקלידי רגיש לטרנספורמציית מתיחה (scaling).



קללת המימדיות: גישות לפתרון

הגישה הכללית: הורדת מימד תוך שמירה מירבית על האנפורציה הרלוונטית.

ההבדלים בין השיטות השונות:

- הגדרת המושג "אינפורמציה רלוונטית".
- השיטות להורדת המימד (בדר"כ טרנספורמציה ליניארית).

Principal Component Analysis – PCA

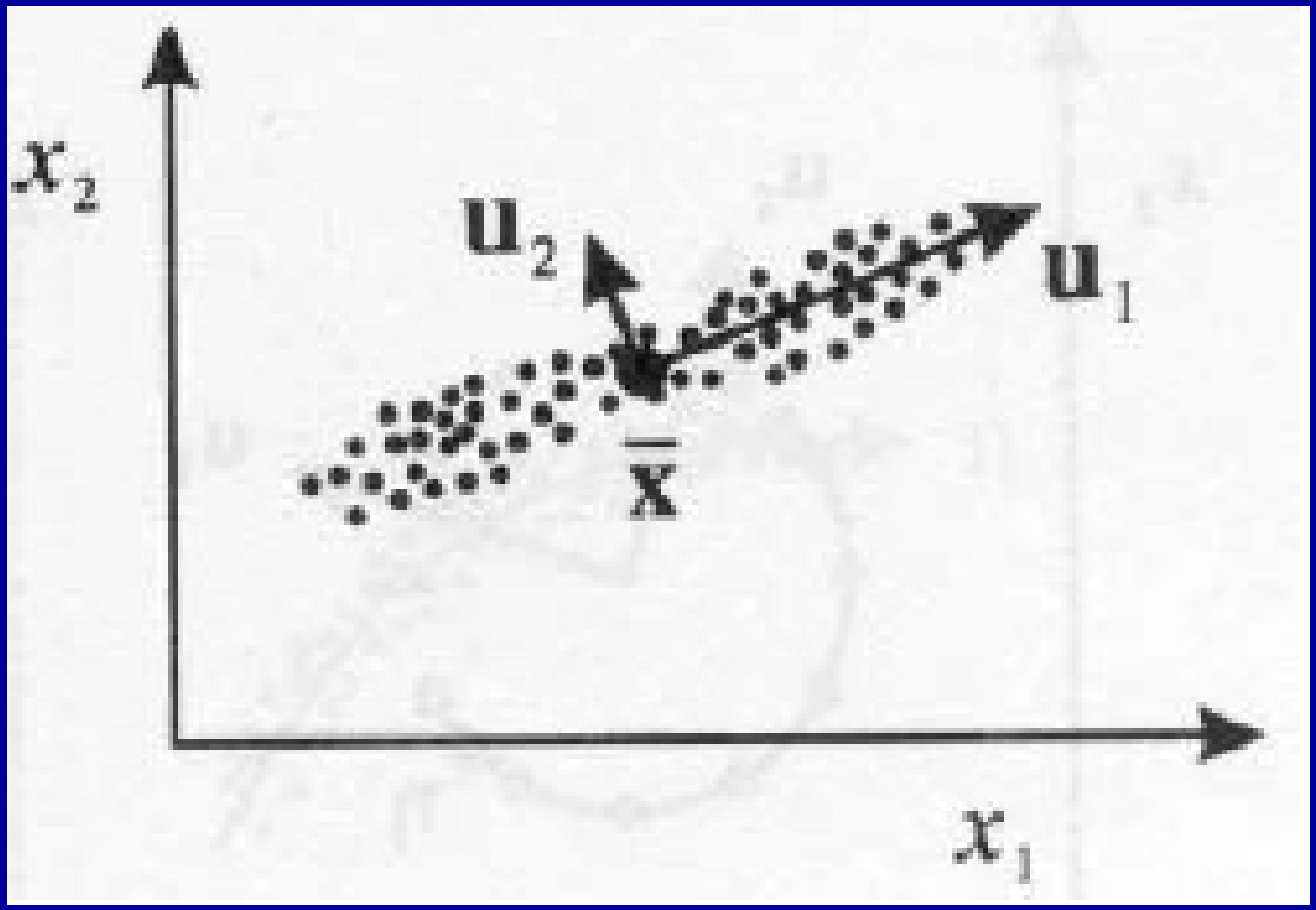
מוטיבציה:

בהינתן מרחב n-מימדי, נרצה למיין את המימדים השונים לפי סדר "החשיבות", ונוותר על m המימדים ה-"פחות חשובים".

כדי שהתהליך יהיה יעיל, נרצה תחילה להחליף את הבסיס של המרחב לבסיס אחר כך ש-m המימדים שנוותר עליהם יהיו "לא חשובים" ככל האפשר.

התהליך של החלפת הבסיס מתבצע ע"י מציאת ווקטורים עצמיים למטריצת הקובריאנס של ווקטורי האימון, וקביעת הבסיס כאוסף הווקטורים העצמיים.

m המימדים שנוותר עליהם הם m הווקטורים העצמיים בעלי הערכים


x_2
u_2
u_1
$\bar{x}$
x_1

נתונים הווקטורים $x_1,\ldots,x_n$ במרחב ווקטורי d-מימדי.

נרצה למצוא טרנספורמציה אורתונורמלית W המעבירה ווקטור x_i ל- y_i, $y_i=Wx_i$ המקיימת תנאי אופטימליות עבור הקריטריון הבא:

לכל ווקטור y_i נמחק את המימדים $m+1,\ldots,d$ ונציב במקומם קבועים $b_{m+1},\ldots,b_d$, כלומר:

$$x_i \to y_i = \sum_{j=1}^{d} z_{i,j} u_j \quad , \quad y'_i = \sum_{j=1}^{m} z_{i,j} u_j + \sum_{j=m+1}^{d} b_j u_j$$

כאשר $u_j=We_i$, $\{e_i\}$ הבסיס סטנדרטי.

נחשב את שגיאת הקירוב עבור x_i:

$$\mathrm{err}_i = x_i - y'_i = \sum_{j=m+1}^{d} (z_{i,j} - b_j) u_j$$

נחשב את ממוצע ריבוע השגיאה:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i - y'_i\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=m+1}^{d}(z_{i,j} - b_j)u_j\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{d}(z_{i,j} - b_j)^2$$

נמצא את ערכי b_j המביאים את E למינימום:

$$\frac{\partial E}{\partial b_j} = 0$$

$$\frac{\partial E}{\partial b_j} = -2\sum_{i=1}^{n}(z_{i,j} - b_j)$$

$$b_j = \frac{1}{n}\sum_{i=1}^{n} z_{i,j} = \frac{1}{n}\sum_{i=1}^{n} u_j^t x_i = u_j^t \overline{x}$$

כלומר b_j הנו הערך הממוצע עבור קואורדינטה j במרחב החדש.

כעת נמצא את W:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i - y'_i\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{d}(z_{i,j} - b_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{d}(u_j^t x_i - u_j^t \bar{x})^2 =$$

$$= \sum_{j=m+1}^{d} u_j^t \Sigma u_j$$

$$\Sigma = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t$$

W הרצוי יביא למינימום את E.

<u>פתרון (ללא הוכחה):</u> $\{u_j\}$ הם הווקטורים העצמיים של Σ.

לפיכך, השגיאה תהיה:

$$E = \sum_{j=m+1}^{d} u_j^t \Sigma u_j = \sum_{j=m+1}^{d} u_j^t \lambda_j u_j = \sum_{j=m+1}^{d} \lambda_j u_j^t u_j = \sum_{j=m+1}^{d} \lambda_j$$

לפיכך נוותר על d-m הווקטורים העצמיים בעלי הערכים-עצמיים הנמוכים ביותר.

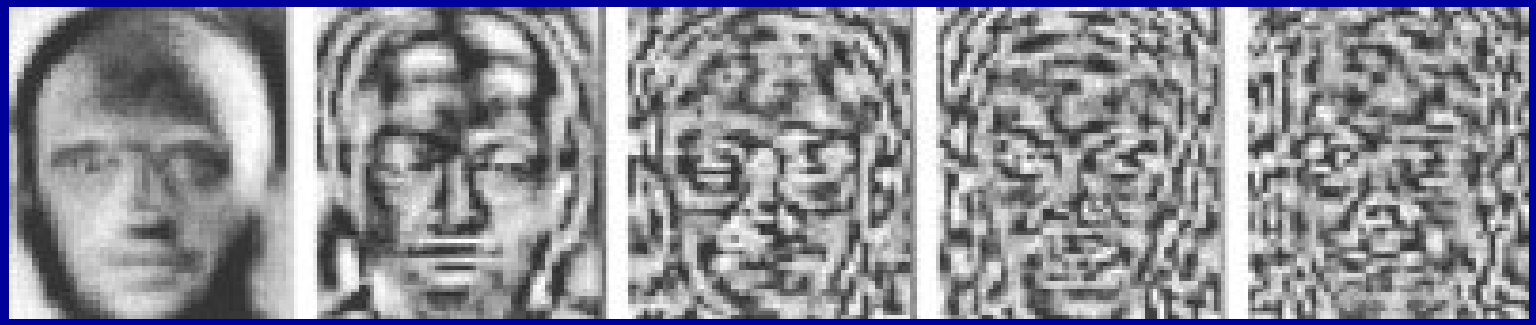
<u>לסיכום, האלגוריתם הוא כדלהלן:</u>

קלט: סידרת ווקטורים x_i.

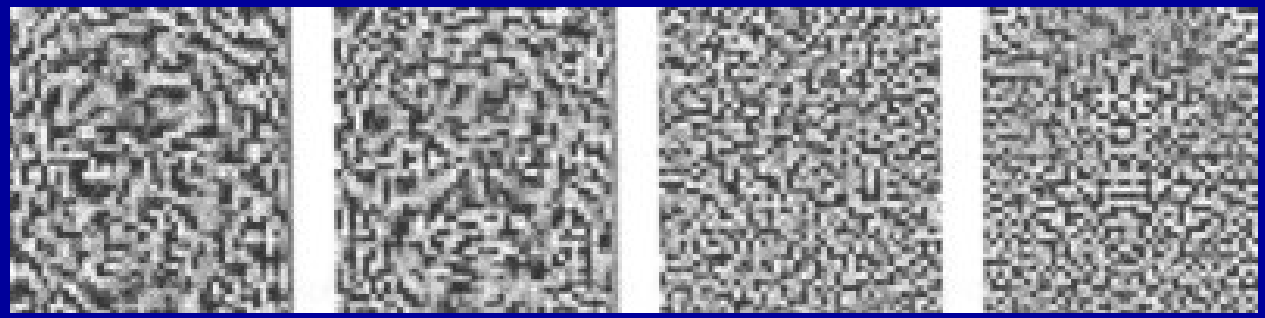
פלט: סידרת ווקטורים z_i שהיא היטל של הקלט על מרחב במימד נמוך יותר.

1. נחשב את ווקטור ממוצע הקלט $\leftarrow Ex$.
2. נחשב את מטריצת הקובריאנס Σ.
3. נמצא את הווקטורים העצמיים של Σ, ונמיינם עפ"י ערכיהם העצמיים.
4. נבחר את m ($m<d$) הווקטורים העצמיים בעלי ע"ע מקסימליים – $\{u_j\}$.
5. לכל ווקטור x_i נחשב את ההיטל שלו על המרחב הנפרש ע"י $\{u_j\}$ $z_i=[(x_i^t-Ex)u_1 \quad \ldots \quad (x_i^t-$ לאחר הפחתת Ex.

הפנים הממוצעות (בצד שמאל) וארבעת ה-Eigenfaces הראשונים



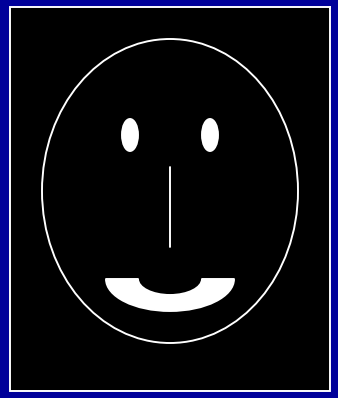
Eigenfaces 15,100,200,250,300 (מצד שמאל)

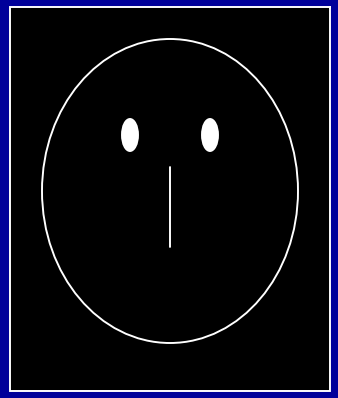


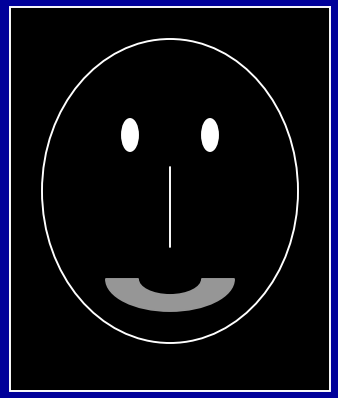
Eigenfaces 400,450,1000,2000 (מצד שמאל)

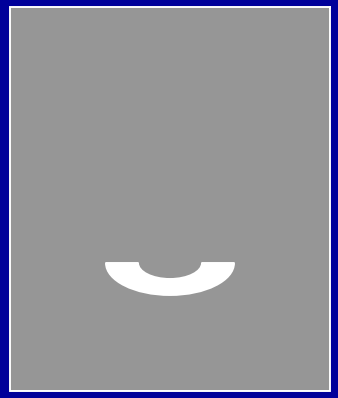
אינטרפטציה לשימוש ב-PCA (eigenfaces) עבור זיהוי פנים:

שיטת הרכבת קלסתרון ע"י המשטרה.

$x_1=$ 

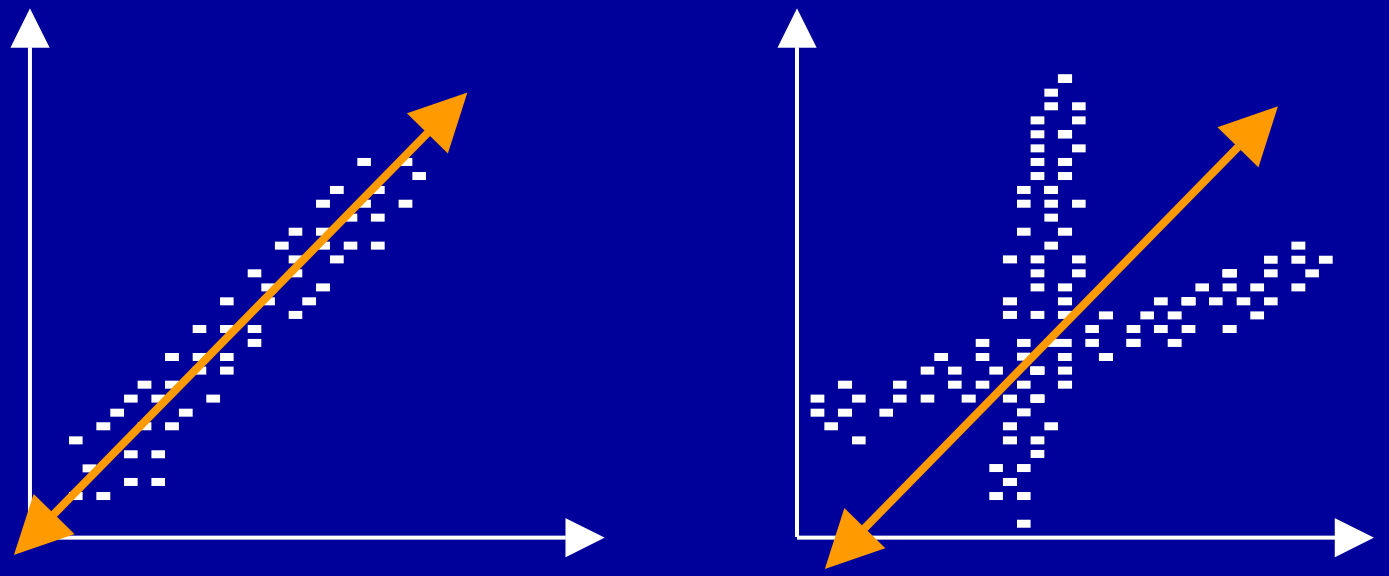
$x_2=$ 

$Ex=(x_1+x_2)/2=$ 

$u_1=$ 

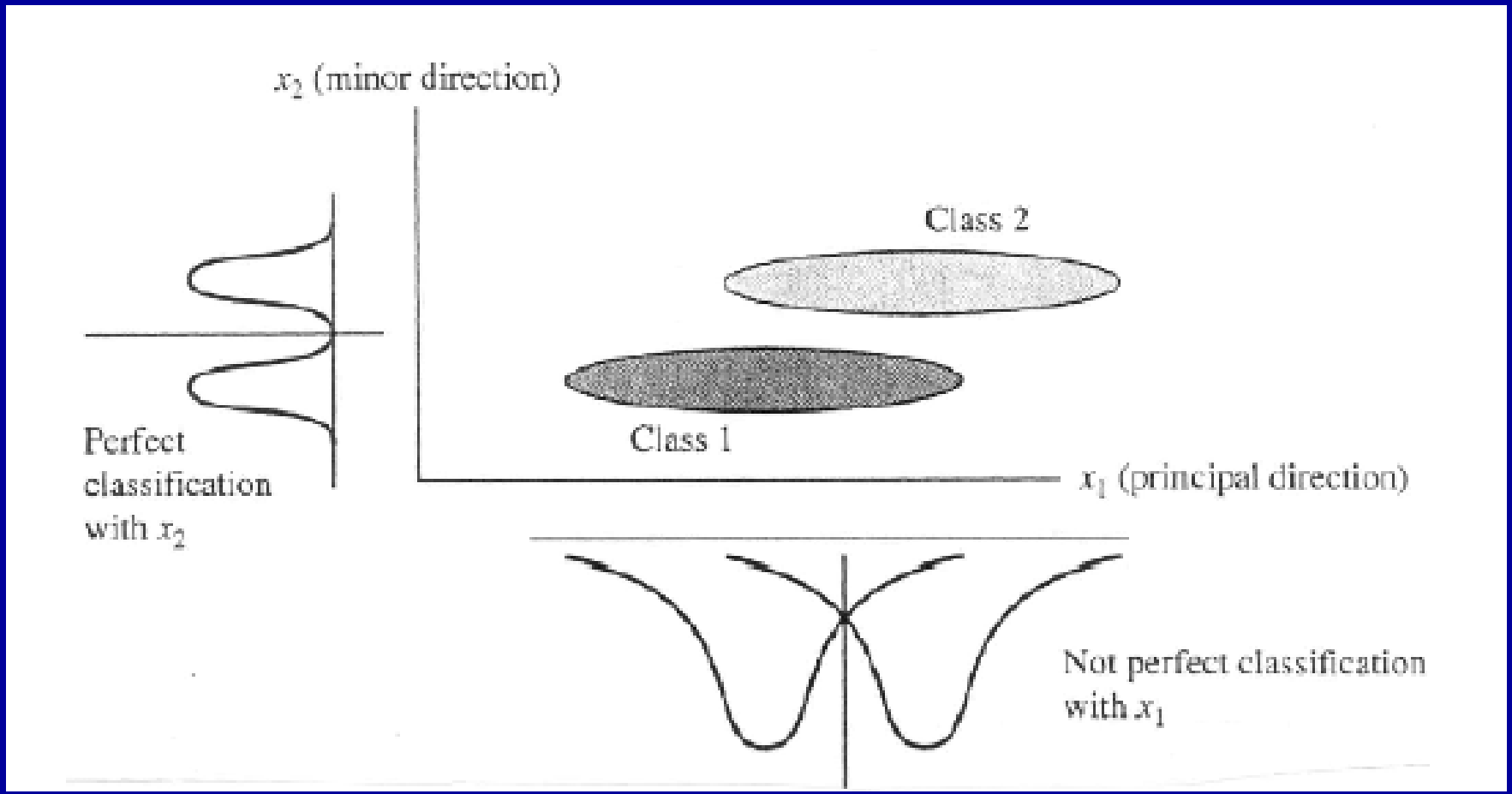
מגבלות שיטת PCA:

1. PCA הנה שיטה ליניארית, ואינה יכולה לאפיין תת-מרחבים לא ליניאריים. (לדוגמא קואורדינטות פולריות).
2. PCA אינה מתאימה לתאר צבירים.



3. PCA רגישה לנרמול הצירים השונים.

4. PCA אינה מנסה להפריד בין מחלקות שונות.



זיהוי תבניות:

מסווגים בייסיאנים

סיווג בייסיאני

- חוק Bayes
- בחירה בין השערות שונות – ML, MAP
- מסווג בייסיאני נאיבי
- למידה פרמטרית / א-פרמטרית
- למידת MAP,ML

חוק בייס (Bayes)

$$\Pr(w|X) = \frac{\Pr(X|w)\Pr(w)}{\Pr(X)}$$

$\Pr(w)$: הסתברות אפריורית של ההשערה w.

$\Pr(X)$: הסתברות אפריורית של התצפיות X.

$\Pr(w|X)$: הסתברות מותנית של ההשערה w בהינתן תצפיות X.

$\Pr(X|w)$: הסתברות מותנית של תצפיות X בהינתן ההשערה w.

בחירת השערות

$$\Pr(w|X) = \frac{\Pr(X|w)\Pr(w)}{\Pr(X)}$$

נרצה לבחור את ההשערה המסתברת ביותר בהינתן התצפיות:

.Maximum a posteriori (MAP)

$$w_{MAP} = \underset{w\in\{W\}}{\operatorname{argmax}} \Pr(w|X) = \underset{w\in\{W\}}{\operatorname{argmax}} \frac{\Pr(X|w)\Pr(w)}{\Pr(X)} = \underset{w\in\{W\}}{\operatorname{argmax}} \Pr(X|w)\Pr(w)$$

<u>הגדרה</u>: נראות (likelihood) היא ההסתברות של התצפיות בהינתן המודל w.

לעיתים נניח כי $\Pr(w)$ מתפלג אחיד, ואז נבחר את ההשערה בעלת הנראות המירבית (Maximum Likelihood):

$$w_{ML} = \underset{w\in\{W\}}{\operatorname{argmax}} \Pr(X|w)$$

דוגמא לבחירת השערות

האם לנבדק יש סרטן?

נבדק מבצע בדיקת מעבדה, והבדיקה יוצאת חיובית.

סטטיסטיקה ידועה על הבדיקה:

- אם יש סרטן, 98% שהבדיקה תהיה חיובית.
- אם אין סרטן, 97% שהבדיקה תהיה שלילית.

בנוסף: ל- 0.8% מן האוכלוסייה יש בסרטן.

נגדיר:

w – יש סרטן / אין סרטן.

X - תוצאת הבדיקה.

$Pr(w = cancer) = 0.008$, $Pr(w = !cancer) = 0.992$

$Pr(X=positive \mid w=cancer) = 0.98$, $Pr(X=negative \mid w = cancer) = 0.02$

$Pr(X=positive \mid w=!cancer)= 0.03$, $Pr(X=negative \mid w = !cancer)= 0.97$

דוגמא לבחירת השערות

$$W_{MAP} = \underset{w \in \{W\}}{\text{argmax}} \Pr(X|w)\Pr(w)$$

$\Pr(X=\text{positive} \mid w=\text{cancer}) * \Pr(w=\text{cancer}) = 0.98 * 0.008 = 0.00784$

$\Pr(X=\text{positive} \mid w=!\text{cancer}) * \Pr(w=!\text{cancer}) = 0.03 * 0.992 = 0.02976$

$W_{MAP} = !\text{cancer}$

$\Pr(w=\text{cancer} \mid X=\text{positive}) = 0.00784/ (0.00784 + 0.02976) = 20.9\%$

$\Pr(w=!\text{cancer} \mid X=\text{positive}) = 0.02976 / (0.00784 + 0.02976) = 79.1\%$

... אבל 20.9% היא הסתברות מספיק גבוהה בשביל לנקוט בפעולות נוספות.

← כדאי לשקלל את ה"עלות" של קבלת כל השערה.

$$w_{MAP\text{-}cost} = \underset{w \in \{W\}}{\text{argmax}} \Pr(w|X) * \text{cost}(\text{decision} = !w \mid w)$$

נניח שהעלות של בחירה "cancer" בטעות היא 1,

והעלות של בחירה "!cancer" בטעות היא 10.

$$\Pr(X=\text{positive} \mid w=\text{cancer}) * \Pr(w=\text{cancer}) * \text{cost}(\text{decision} = \text{!cancer} \mid \text{cancer})$$

$$= 0.98 * 0.008 * 10 = 0.0784$$

$$\Pr(X=\text{positive} \mid w=\text{!cancer}) * \Pr(w=\text{!cancer}) * \text{cost}(\text{decision} = \text{cancer} \mid \text{!cancer})$$

$$= 0.03 * 0.992 * 1 = 0.02976$$

$$W_{\text{MAP-cost}} = \text{cancer}$$

נוסחאות רלוונטיות בהסתברות

1. $\Pr(A,B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$

2. $\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A,B)$

3. $\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i)\Pr(A_i)$ if $\Pr(A_i \cap A_j) = 0 \quad \forall i \neq j,$

$$\Pr(\cup A_i) = 1$$

מסווג Naïve Bayes

בהינתן ווקטור תצפיות $X=x_1,x_2,..x_n$

הנחת "naïve bayes" :- קיימת אי תלות בין התצפיות השונות.

$$\Pr(x_1,\ldots,x_n|w)\cong\prod_{i=1}^{n}\Pr(x_i|w)$$

כך שמקבלים:

$$W_{NB}=\underset{w\in\{W\}}{\operatorname{argmax}}\ \Pr(w)\prod_{i=1}^{n}\Pr(x_i|w)$$

נחזור לדוגמא עם בדיקת הסרטן:

נניח שישנה בדיקה נוספת (Y) בלתי-תלויה שניתן לבצע.

הנתונים לגבי הבדיקה:

$\Pr(Y=\text{positive} \mid w=\text{cancer}) = 0.5, \quad \Pr(Y=\text{negative} \mid w=\text{cancer}) = 0.5$

$\Pr(Y=\text{positive} \mid w=!\text{cancer}) = 0.1, \quad \Pr(Y=\text{negative} \mid w=!\text{cancer}) = 0.9$

אפשר לראות שהבדיקה Y פחות מובהקת מן הבדיקה X, אך אי-תלותן משפרת את יכולת החיזוי.

נניח שהבדיקה Y אף היא חיובית עבור הנבדק שלנו.

$\Pr(X,Y=\text{positive} \mid w=\text{cancer}) * \Pr(w=\text{cancer}) = 0.98 * 0.5 * 0.008 = 0.00392$

$\Pr(X,Y=\text{positive} \mid w=!\text{cancer}) * \Pr(w=!\text{cancer}) = 0.03 * 0.1 * 0.992 = 0.002976$

$W_{MAP} = \text{cancer}$

$\Pr(w=\text{cancer} \mid X,Y=\text{positive}) = 0.00392 / (0.00392 + 0.00298) = 56.8\%$

$\Pr(w=!\text{cancer} \mid X,Y=\text{positive}) = 0.00298 / (0.00392 + 0.00298) = 43.2\%$

מסקנה:

אם נבצע הרבה בדיקות בלתי תלויות, נוכל להגיע ליכולת חיזוי גבוהה יותר.

ביקורת:

1. ההסתברות של סרטן באוכלוסייה $Pr(cancer)=0.008$ פחות רלוונטית מההסתברות של סרטן בקרב אוכלוסיית הנבדקים (שהיא כנראה גבוהה יותר).
2. האם אכן הבדיקות בלתי תלויות?

שערוך מודל

- במציאות, ההתפלגויות אינן ידועות מראש.
- יש צורך ללמוד אותן.
- כאשר המשתנים הם בדידים (כמו בדוגמת ה-cancer),
  יש צורך לחשב אוסף סופי של הסתברויות.
- לדוגמא, נרצה לדעת מהו P(cancer).
  בהינתן מדגם (קבוצת אימון) של n אנשים,ידוע על m
  מתוכם החולים בסרטן.
  נניח התפלגות בינומית.
- נעריך את P(cancer) כ- m/n.

שערוך מודל

השיטה הבייסיאנית נותנת מענה אופטימלי אך דורשת מידע מוקדם על כל ההתפלגויות הרלוונטיות.

במציאות ההתפלגויות אינן ידועות מראש – צריך לשערך (estimate) אותן.

לפעולה זו של שערוך ההסתברויות נקרא גם שערוך מודל, או "אימון מודל" (model training, model estimation).

את האימון נבצע באמצעות "דוגמאות אימון".

בקורס נדבר בעיקר על למידה מונחית: כאשר נתון הסיווג של דוגמאות האימון.

לדוגמא, נניח שנתון מאגר מידע על מטופלים בו ידוע על כל מטופל

גישות לשערוך פרמטרים

הגישה הפרמטרית:

1. נניח שפונקצית ההתפלגות היא בעלת צורת התפלגות מוכרת (בינומית, נורמלית, אקספוננציאלית).
2. ננסה "לנחש" את הפרמטרים של ההתפלגות.

הגישה הא-פרמטרית:

1. לא נניח (כמעט) דבר בנוגע לצורת פונקצית ההפלגות.
2. נלמד את פונקצית ההתפלגות ישירות מדוגמאות האימון.

הגישה הפרמטרית

נניח שרוצים לשערך את ההתפלגות $P(x|w)$.

לדוגמא $w=\{male, female\}$ ו-$x=height$.

נציג את $P(x|w)$ כך:

$$P(x|w_i) = f(x,\Theta_i)$$

לדוגמא: $\Theta=\{\mu, \sigma^2\}$

$$P(x|w_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

כל מה שנותר הוא לשערך את ערך הפרמטרים Θ.

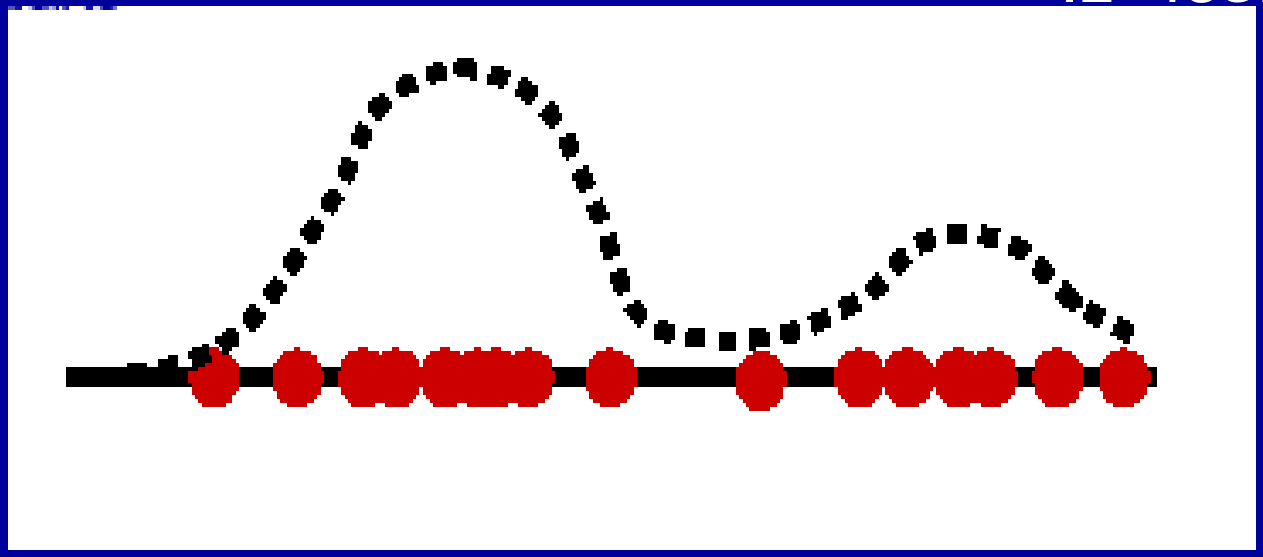
הגישה הא-פרמטרית

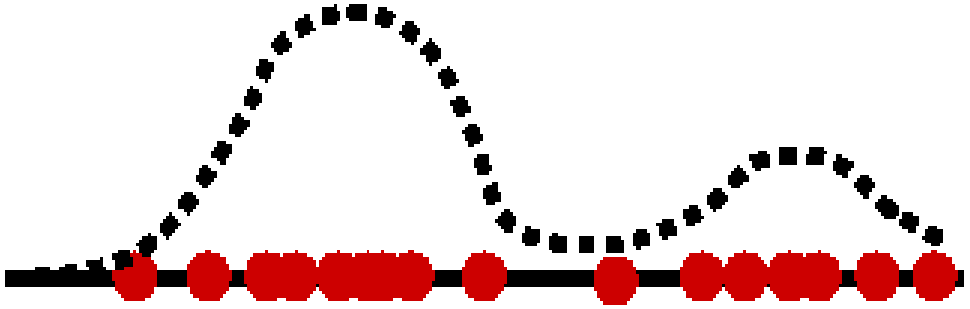
נניח שלא ידועה לנו אופי פונקצית ההתפלגות $P(x|w)$.

שיטה מס' 1: שיטת ההיסטוגרמה

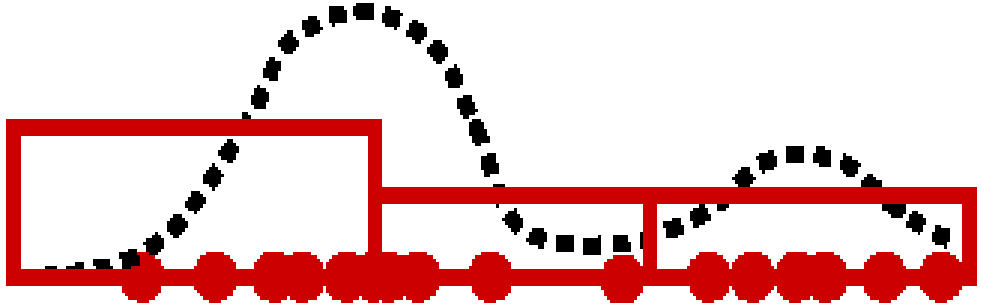
נניח ש- X הוא חד מימדי.

1. נחלק את הטווח של ערכי X ל-M תחומים שווים.
2. נתייחס להתפלגות של X כהטלת מטבע עם M ערכים אפשריים.

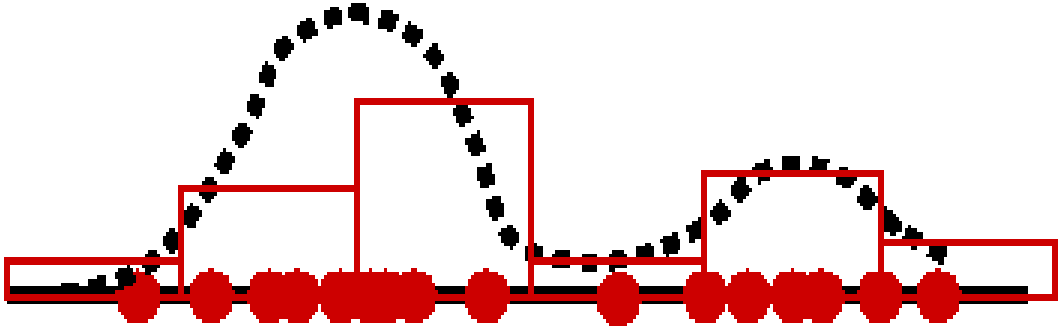




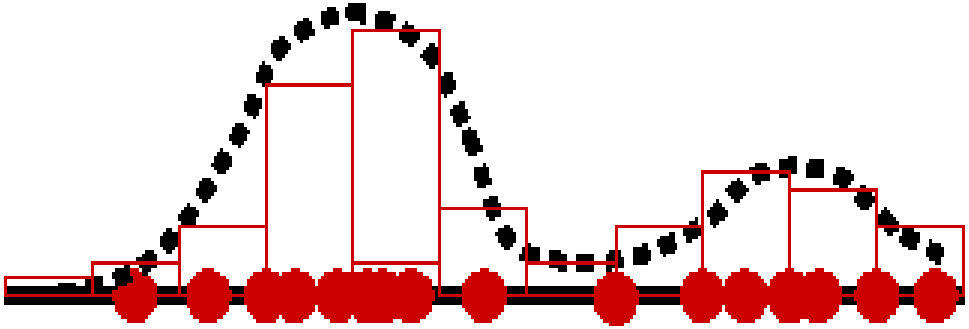
3 bins



7 bins



11 bins



שיטת ההיסטוגרמה

מספר התחומים M משפיע מאוד על איכות השיטה:

M גדול: ההיסטוגרמה "רועשת".

M קטן: מאבדים הרבה אינפורמציה.

גודל ה-M האופטימלי תלוי ב:

- גודל קבוצת האימון.
- מבנה פונקצית ההתפלגות.
- רגישות הבעיה לחוסר דיוק / רעש.

<u>חסרונות עיקריים</u>: במימדים גבוהים יותר נצטרך קבוצת אימון בגודל M^{dim}.

פונקצית ההתפלגות שמקבלים איננה רציפה.

דוגמא:

$w=\{w_1=male, w_2=female\}$ ו-x הינו הגובה.

נחלק את ,תחום x לשני חלקים: $[x<=1.80], [x>1.80]$

$\Theta=\{p_i\}$

$$P(x|\, w_i) = \begin{cases} p_i & \text{iff } x > 1.80 \\ 1-p_i & \text{else} \end{cases}$$

במקרה זה, כל מה שנותר לשערך הוא שתי הסתברויות (p_1, p_2).

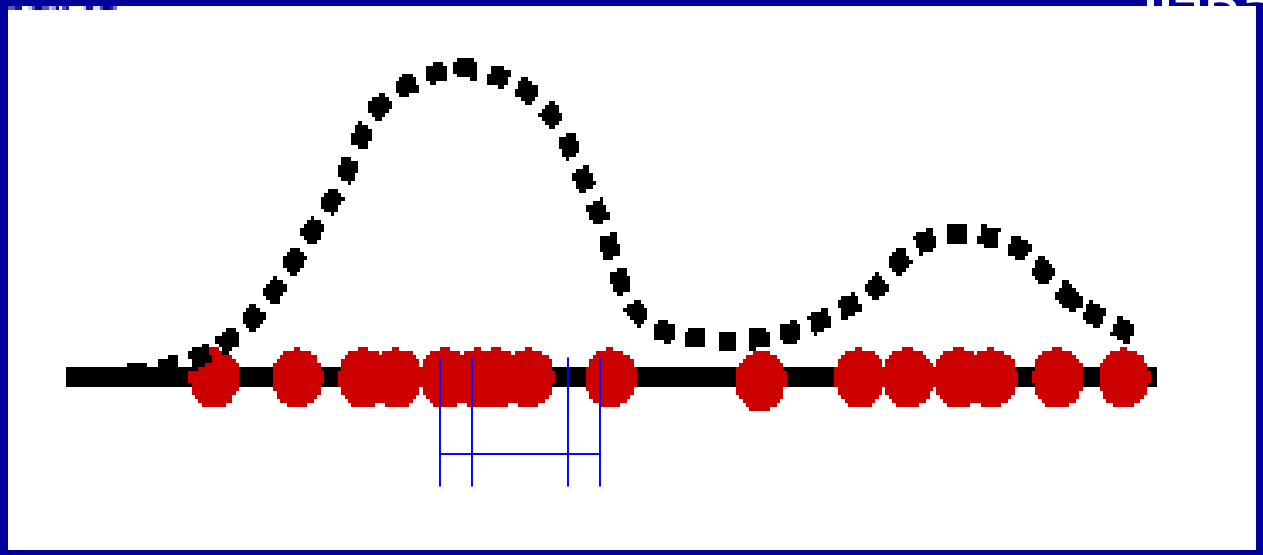
שיפור החסרונות של גישת ההסטוגרמה

גישת החלון

כדי לחשב את צפיפות ההסתברות $P(X|w)$ בנקודה x_0, נחשב את מספר

המקרים מקבוצת האימון שבהם ערך x הוא בחלון כלשהו סביב x_0, ונחלק

בנפח החלון



דוגמא:

$w=\{w_1=male, w_2=female\}$ ו-x הינו הגובה.

נחלק את ,תחום x לטווחים החופפים הבאים:

לכל x_0, נתאים טווח $[x_0 - d, x_0 + d]$.

$$Pr(x = x_0 \mid w_i) = Pr(x \in [x_0 - d, x_0 + d] \mid w_i) \ / \ 2d$$

איך נבחר את גודל החלון d?

- שרירותית
- נבחר גודל משתנה (כפונקציה של מספר הנקודות סביב x_0).

שיטה מס' 2: Vector Quantization

נרצה לבחור קבוצה של k "נציגים" במרחב המאפיינים,

כך שכל ווקטור במרחב יוחלף ע"י אחד הנציגים.

היתרון: עברנו מבעיית סיווג רציפה במימד גבוה לבעיית סיווג בדידה.

בעיית השערוך של ההתפלגות $Pr(X|w)$ הופכת לבעיית שערוך של

ההתפלגות $Pr(i|w)$, כש-i הוא מספר טבעי $1 \le i \le k$.

החסרון: איבוד רזולוציה (דיוק).

ככל שמספר הנציגים יהיה גדול יותר, איבוד הדיוק יהיה קטן יותר, אבל

נצטרך קבוצת אימון גדולה יותר כדי לשערך את $Pr(i|w)$.

אלגוריתם למציאת k הנציגים: k-means

נרצה למצוא k נציגים אשר עבורם איבוד הדיוק הממוצע עבור קבוצת

האימון הוא מינימלי.

נגדיר את העיוות של ווקטור x כריבוע המרחק האוקלידי בינו לבין הנציג

הקרוב ביותר אליו.

הערה: מציאת k נציגים כאלו היא NP-קשה.

אלגוריתם k-means:

1. אתחל את קבוצת k הנציגים (למשל אקראית).

בצע עד להתכנסות:

2. סווג כל ווקטור בקבוצת האימון לנציג הקרוב ביותר.
3. חשב מחדש את הנציגים עפ"י תוחלת הווקטורים המשויכים

ניתוח אלגוריתם k-means

- אלגוריתם k-means מתכנס תמיד.
- האלגוריתם מתכנס למינימום מקומי.
- אתחול אקראי מספק בדרך-כלל.
- בחירת k משמעותית – קטן מדי: אובדן דיוק.
  גדול מדי: over-fitting.
- בדר"כ נבחר את k אמפירית.

שימוש נוסף לאלגוריתם: דחיסה (לא משמרת).

הגישה הפרמטרית: שיטות לשערוך פרמטרים

כדי לשערך את הפרמטרים של מודל כלשהו, יש צורך בשני מרכיבים:

1. קריטריון טיב (= פונקצית ציון) למודל.
2. אלגוריתם חיפוש במרחב הפרמטרים.

קריטריון הטיב למודל מאפשר להשוות בין שני מודלים ולענות על השאלה:

"איזה מודל טוב יותר"?

אלגוריתם החיפוש משתמש בקריטריון הטיב כדי למצוא מודל אופטימלי

(מודל אשר טוב יותר מכל המודלים האפשריים האחרים).

קריטריוני טיב

הסתברות אפוסטריורית מירבית (MAP - Maximum a-posteriori)

נרצה לבחור את המודל המסתבר ביותר בהינתן תצפיות האימון.

$$w_{MAP} = \underset{w \in \{W\}}{\operatorname{argmax}} \Pr(w|X) = \underset{w \in \{W\}}{\operatorname{argmax}} \frac{\Pr(X|w)\Pr(w)}{\Pr(X)} = \underset{w \in \{W\}}{\operatorname{argmax}} \Pr(X|w)\Pr(w)$$

כזכור, $P(x \mid w) = f(x,\Theta)$, ולכן:

$$\Theta_{MAP} = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} \Pr(\Theta|X) = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} \frac{f(X,\Theta)\Pr(\Theta)}{\Pr(X)} = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} f(X,\Theta)\Pr(\Theta)$$

קריטריוני טיב (2)

נראות מקסימלית (ML - Maximum Likelihood)

נניח שההתפלגות $\Pr(\Theta)$ היא אחידה, כלומר כל המודלים האפשריים הם

סבירים אפריורית באותה מידה.

$$\Theta_{MAP} = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} \Pr(\Theta|X) = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} \frac{f(X, \Theta)\Pr(\Theta)}{\Pr(X)} = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} f(X,\Theta)\Pr(\Theta)$$

$$\Theta_{ML} = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} f(X,\Theta)\Pr(\Theta) = \underset{\Theta \in \{\Theta\}}{\operatorname{argmax}} f(X,\Theta)$$

נראות מקסימלית (ML - Maximum Likelihood)

דוגמא 1: התפלגות בינומית.

דוגמא 2: התפלגות נורמלית.

הסתברות אפוסטריורית מירבית (MAP - Maximum a-posteriori)

דוגמא 1: התפלגות בינומית עם מידע אפריורי.

שיערוך התפלגות בינומית: בעיית שכיחות אפס

נחזור לדוגמת מידול טקסט.

נניח הנחה מרקובית מסדר 2:

$$\Pr(x_i \mid x_1,\ldots,x_{i-1}, w) = \Pr(x_i \mid x_{i-2}, x_{i-1}, w)$$

נצטרך לשערך מתוך קבוצת האימון עבור כל שלשה אפשרית של אותיות

a,b,c את ההסתברות $\Pr(x_i= a| x_{i-2} = b, x_{i-1}= c, w)$.

עפ"י שיטת שערוך Maximum Likelihood, נספור לכל שלשה את

מספר המופעים k של השלשה בקבוצת האימון, ונחלק בגודל קבוצת האימון:

$$\Pr(x_i= a| x_{i-2} = b, x_{i-1}= c, w) = k(a,b,c) / n$$

הבעיה: סביר להניח שתהיינה שלשות שעבורן נקבל: $k(a,b,c)=0$.

שיערוך התפלגות בינומית: בעיית שכיחות אפס (המשך)

מה יקרה עם שלשה כזו תופיע בטקסט המבחן?

... נקבל:

$$\Pr(X|w) = 0$$

<u>פתרון אפשרי:</u>

נעדכן את פונקציית שערוך הפרמטרים לפונקציה הבאה:

$$\Pr(x_i = a \mid x_{i-2} = b, x_{i-1} = c, w) = \frac{k(a,b,c) + \varepsilon}{n + \varepsilon * m^3}$$

m – גודל האלף-בית

n – גודל טקסט האימון

ε - פרמטר שרירותי (בסדר גודל 0.1-1).

שיערוך פרמטרים: הגישה הבייסיאנית

במקום לחפש השמה לפרמטרים (מודל) הממקסמת קריטריון כלשהו,

נחשב את ההסתברות של כל השמה (מודל).

כשנרצה להשתמש במודל לצורך חישוב כלשהו, נשתמש בכל המודלים,

ונמשקלם בהסתברותם.

כלומר נחשב את $\Pr(\Theta|X)$ לכל Θ ונחשב את $\Pr(Y|X)$ כך:

$$\Pr(Y|X)=\int_{\Theta}\Pr(Y|\Theta)\Pr(\Theta|X)d\Theta$$

דוגמא 2: הסתברות לזריחת השמש.

הגישה הפרמטרית: תערובת גאוסיינים (GMM)

ראינו שקיים פתרון סגור לשערוך פרמטרים של התפלגות נורמלית.

הבעיה: במציאות, רוב ההתפלגויות אינם נורמליות.

הפתרון: נקרב כל התפלגות ע"י תערובת של התפלגויות נורמליות –

תערובת גאוסיינים (Gaussian Mixture Model).

$$\Pr(x) = \sum_{k=1}^{n} w_k \Pr(x \mid N(\mu_k, \sigma^2{}_k))$$

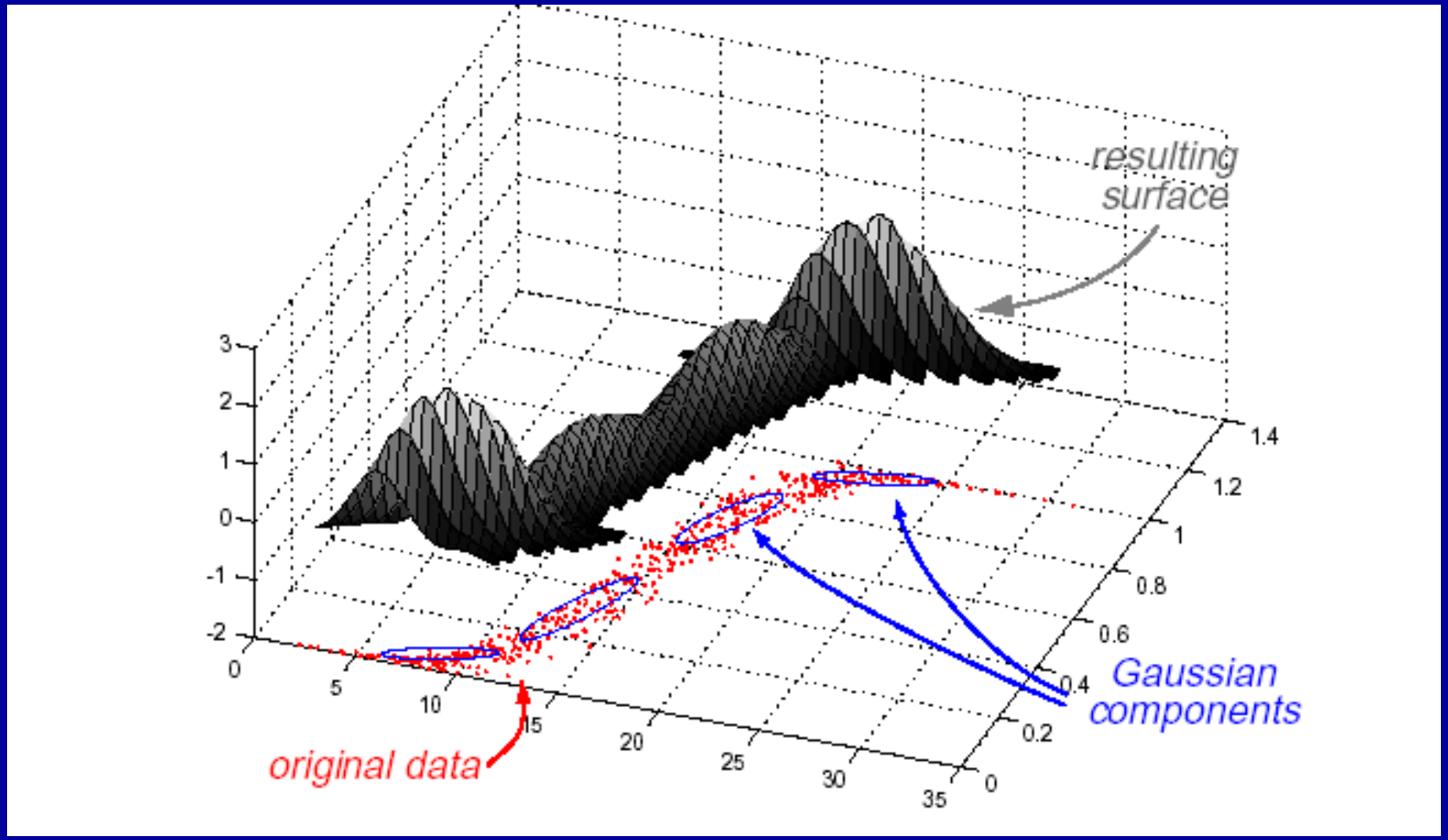
$$\sum_{k=1}^{n} w_k = 1$$

טענה: אפשר להראות של כל התפלגות (סבירה) אפשר לקרב ע"י

תערובת של מספר קטן יחסית התפלגויות נורמליות.

אינטרפרטציה: כל תצפית x נוצרת ע"י אחד מן הגאוסיינים הנבחר עפ"י

דוגמא להתפלגות המתוארת ע"י GMM



אם היינו יודעים עבור כל תצפית לאיזה גאוסיין היא "שייכת",

יכולנו לשערך את הפרמטרים של ה-GMM בצורה דומה לשערוך פרמטרי

גאוסיין בודד.

<u>הבעיה</u>: אנו לא יודעים עבור כל תצפית לאיזה גאוסיין היא "שייכת".

מעבר לכך, לא ידוע על שיטה כלשהי המאפשרת למצוא את הערכים ה-ML

של תערובת גאוסיינים.

<u>הפתרון (החלקי):</u> נפעיל שיטה אשר מאפשרת למצוא מקסימום מקומי

לפונקצית ה-likelihood (במקום המקסימום גלובאלי הדרוש).

לשיטה זו קוראים EM.

אלגוריתם EM

אלגוריתם EM הוא אלגוריתם כללי (לאו דווקא ל-GMM) המטפל בבעיה הבאה:

- רוצים לשערך את הפרמטרים Θ של מודל פרמטרי.
- קיימים משתנים חבויים Y אשר באמצעותם קל לשערך את המודל הפרמטרי מתוך התצפיות X.

במקרה שלנו, Y מציין עבור כל תצפית מאיזה גאוסיין היא "נוצרה".

אלגוריתם ה-EM הנו האלגוריתם האיטרטיבי הבא:

1. אתחל את המודל Θ בערכים כלשהם.

חזור עד להתכנסות:

2. חשב את ערכי Y עפ"י Θ ו-X.
3. שערך את Θ עפ"י Y ו-X.

ניתן להוכיח עבור אלגוריתם EM:

1. בכל איטרציה ה-likelihood משתפר.
2. האלגוריתם מתכנס.

אבל... לאו דווקא מגיעים למקסימום גלובאלי.

אלגוריתם EM עבור שיערוך GMM

1. אתחל בצורה כלשהי את הפרמטרים של המודל, למשל ע"י:
   בחירה אקראית של ווקטורים כפרמטרי התוחלת μ_k,
   שימוש בקובריאנס הכללי לאיתחול פרמטרי הקובריאנס σ_k,
   ערך אחיד לפרמטרי המשקל w_k.

2. בצע מס' איטרציות (עד להתכנסות):
   E: חשב את ההסתברות של כל אחד מן הגאוסיינים עפ"י כל ווקטור x :

$$(I) \quad \Pr(x|k) = \Pr(x \mid N(\mu_k, \sigma^2{}_k))$$

$$(II) \quad \Pr(k|x) = \frac{\Pr(x|k)\Pr(k)}{\Pr(x)} = \frac{\Pr(x|k)w_k}{\Pr(x)}$$

$$(III) \quad \sum_j \Pr(j|x) = 1$$

$$\Rightarrow \quad \Pr(k|x) = \frac{\Pr(x|k)w_k}{\sum_j \Pr(x|j)w_k}$$

M: שערך את הפרמטרים של ה-GMM ע"י המשוואות הבאות:

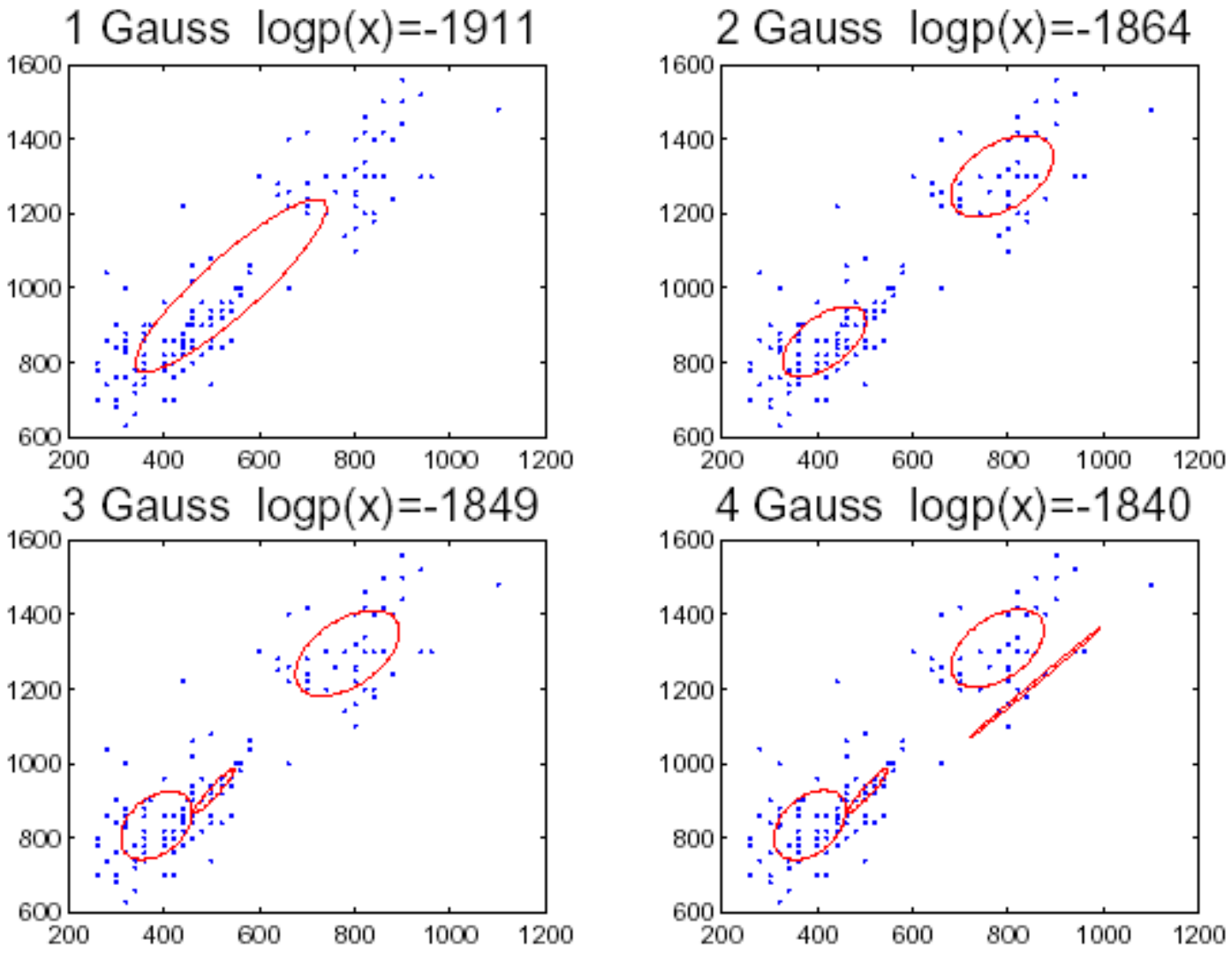
$$\overline{\mu}_k = \frac{\sum_{x \in X} \Pr(k|x) \cdot \overline{x}}{\sum_{x \in X} \Pr(k|x)}$$

$$\sigma_{k_{i,j}} = \frac{\sum_{x \in X} \Pr(k|x) \cdot (x_i - \mu_{k_i}) \cdot (x_j - \mu_{k_j})}{\sum_{x \in X} \Pr(k|x)}$$

$$w_k = \frac{1}{|X|} \sum_{x \in X} \Pr(k|x)$$

GMM examples

- **Vowel data fit with different mixture counts:**



הערות:

1. מודל עשיר מדי (יותר מדי גאוסיינים) ילמד טוב "מדי" את דוגמאות האימון, ולא יכליל.
2. את מס' הגאוסיינים ניתן לקבוע באמצעות קבוצת בקרה (validation set).
3. את מס' הגאוסיינים ניתן לקבוע עפ"י גודל קבוצת האימון: 20-100 דוגמאות/גאוסיין.
4. לעיתים נניח כי מטריצת הקובריאנס אלכסונית.
5. לעיתים נקשור (tie) פרמטרים שונים יחדיו – בעיקר קובריאנס.