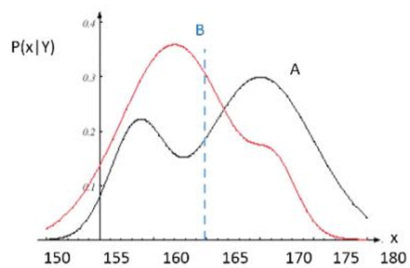
למידה חישובית ממידע: הרצאה 4 – MAP Classifier and Bayesian Learning

למידה בייסיאנית - Bayesian Learning

בקלסיפיקציה נקבל x חדש ונרצה לדעת לסווג אותו לקלאס המתאים לו. בלמידה הסתברותית נתייחס לדגימות שלנו כעל בעלי התפלגות משותפת כלשהי וכך נדע לסווג אותם. נניח שקיבלתי טופס של מבחן ונרצה לנחש האם מדובר במבחן של סטודנט או סטודנטית. לצורך העניין נניח כי יש **60% בנים ו-40% בנות**, לכן לפי נתון זה בלבד ננחש שזהו **מבחן של סטודנט** – מינימום סיכון, **הסתברות לטעות** = **40%**.

- בהינתן שני קלאסים A ו-B וכאשר אנחנו יודעים את ה-prior-probability של המחלקות P(A), P(B)
- נסווג מקרה חדש באופן הבא:
  - Classify A if P(A)>P(B),
  - Classify B otherwise
- נשים לב כי סיווג זה אינו לוקח בחשבון את המידע שיש לנו אודות הדגימה x (פיצ'רים).
- הסתברות הטעות = 1-max(P(A),P(B))
- לרוב לא נרצה לסווג על פי ה-prior אלא נרצה לנקוט בגישה מתקדמת יותר: likelihood

נניח כי ידועות לנו בנוסף, ההסתברויות (P(x | A) ו- P (x | B) למשל: P (height | male) ו- P (height | female) ובגרף מוצגות פונקציות הצפיפות של הסתברויות אלה. נניח כי x = 163cm לכן סביר יותר שהמבחן הוא של סטודנטית מכיוון של ה-likelihood של קלאס B בנקודה x=163 הינה גבוהה יותר מאשר ה-likelihood של A בנקודה זו. P(x=163| B) > P(x=163| A)



מדוע סיווג זה הינו בעייתי? אין התחשבות ב-Prior

P(H>1.9 | NBA) = 0.85   P(H<1.9 | NBA) = 0.15   P(H>1.9 | R) = 0.1   P(H<1.9 | R) = 0.9

פגשנו ברחוב אדם שהוא 1.93, וידוע לנו שההסתברות של שחקן NBA להיות מעל 1.9 היא 85% לכן נסיק שהאדם האקראי הזה הוא שחקן NBA, אבל אחוז שחקני ה-NBA מתוך אוכלוסיית האנשים בעולם הוא זניח ולכן סיווג באופן זה הינו שגוי.

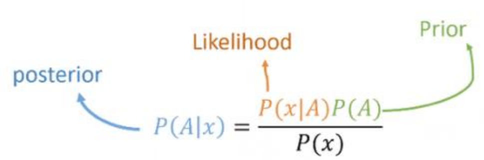
- What we want is the rule:
  - “Classify A if P(A|x) > P(B|x)”
- Not the same – why?
- What about prior probabilities?
- In our example (male/female) $P(A) \cong P(B)$.
  But, in more general cases, even if P(x|A)>P(x|B) it may be the case that $P(A) \ll P(B)$
  (i.e. the probability of A is very very small in the first place even though the specific value x is much more common in A than in B).

אלגוריתם MAP: Maximum A-Posteriori

- אנחנו רוצים להעריך את P(A | x) ואת P(B | x), כלומר – בהינתן x (הדגימה) נרצה לדעת את המצב "האמתי הטבעי" הסביר ביותר.
- המסווג שלנו צריך לסווג באופן הבא:
  -- נסווג A אם P (A | x) > P (B | x)
  -- אחרת נסווג B
- אבל, אנחנו לא יודעים באופן ישיר את ההסתברויות ה"posterior" הללו.

לכן, נשתמש בנוסחת בייס: $\boxed{P(A|x) = \frac{P(x|A)P(A)}{P(x)}}$

<u>הקומפוננטות של נוסחת בייס</u>



- לא הגדרנו את P(x) מפני שהוא משותף לשני הקלאסים.
- P (A | x) = Posterior
- P (x | A) = Likelihood
- P(A) = Prior

נסווג דגימה בעלת ווקטור פיצ'רים x על ידי (k מספר המחלקות):

$$\boxed{C(x) = \max_{i=1..k} \frac{P(x|A_i)P(A_i)}{P(x)}}$$

אבל נוכל להשמיט את P(x) מהנוסחה, מהיות שהחישוב נעשה ביחס ל-i:

$$C(x) = \max_{i=1..k} P(x|A_i)P(A_i)$$

העקרונות של סיווג בייס / Principles of Bayes Classification

- הסיווג תלוי ב-class conditional information (ההסתברויות בהינתן הקלאס) כמו ה- likelihood ובהתפלגות הפריורית (ה-prior).
- חוק בייס מניב:
  -- נסווג A אם $P(A|x) > P(B|x)$
  -- אחרת נסווג B
  -- נדון בסיווג של multiclasses בהמשך.
- נשים לב כי P(x) מוסר מן המכנה משני הצדדים מפני שבשני הצדדים הוא אותו הדבר.

סיווג בעל שגיאה מינימלית

בכל פעם שאנו מתבוננים בערך x, ההסתברות לשגיאה תהיה:

- אם נחליט B אז $P(error|x) = P(A|x)$
- אם נחליט A אז $P(error|x) = P(B|x)$
- החלטת הבייס היא זו שתביא למינימום את ההסתברות של הטעות (ה-error rate).
- שימוש ב-Byes decision: $P(error|x) = \min[P(B|x), P(A|x)]$

אילו ממש ידענו את מבנה ההסתברות השלמה (ואנחנו לא) היינו יכולים להעריך: $Error_P(h) = \int P(Error|x)\,dP(x)$

פונקציית המחיר / Loss = Cost of Wrong Decision

- נניח שיש לנו k מחלקות (קלאסים) שונות A1,…,Ai,…,Ak.
- על פי התבוננות בדגימה x, עלינו לסווג את הדגימה הזו לאחד הקלאסים Ai על ידי יישום גישת הסיווג Bayes/MAP. נשים לב שהחלטה שגויה מניבה loss!
- Loss יכול להיות תלוי באיזה j סווג באופן שגוי ל-i (מה המחיר לכך שדגימה אשר שייכת במציאות לקלאס j, ושויכה על ידי האלגוריתם שלנו לקלאס i). למשל, דוגמה לפונקציית loss עבור סיווג לקלאסים 0-1:

$$\lambda_{ij} = \lambda(\text{Choose } A_i \mid A_j) = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

- 0 אם צדקנו (המחיר הוא 0), 1 הוא המחיר אם טיענו.
- הסיכון הוא המחיר הצפוי / the expected loss עבור החלטת סיווג, המחושבת לפי הסתברויות אפוסטריוריות.
- **הסיכון להחליט Ai כאשר אנו מתבוננים בדגימה x הינו:**

$$R(\text{choose } A_i|\mathbf{x}) = \sum_{j=1}^{k} \lambda_{ij} P(A_j \mid \mathbf{x}) = \sum_{j \neq i} P(A_j \mid \mathbf{x}) = 1 - P(A_i \mid \mathbf{x})$$

המעבר הראשון נכון לכל פונקציית loss זוהי תוחלת המחיר שאנו נשלם אם נחליט i.

המעבר השני מתייחס ספציפית לפונקציית loss 0/1. במעבר השלישי והאחרון נסתמך על כך שהסכום על כל j הינו 1 $\sum_{j=1}^{k} P(A_j \mid \mathbf{x}) = 1$ ונחסיר מסכום זה את ההסתברות של P(Ai|x). **המסווג שלנו מביא למינימום את תוחלת ההפסד שלנו.**

- לכן הסיכון המינימלי לסיווג במקרה שלנו (כאשר פונקציית המחיר היא 0/1 או c/0) יהיה:

$$\boxed{\text{Choose } A_i \text{ such that } P(A_i|x) > P(A_j|x) \text{ for all } j \neq i}$$

- תחת פונקציית zero-one loss השוויונות הבאים שקולים

$$g_i(x) = P(A_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid A_i) P(A_i)}{\sum_{j=1}^{k} P(\mathbf{x} \mid A_j) P(A_j)}$$

$$g_i(x) = P(\mathbf{x} \mid A_i) P(A_i)$$

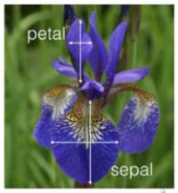
$$g_i(x) = \ln P(\mathbf{x} \mid A_i) + \ln P(A_i)$$

- נשים לב שהחישוב של הprior P(Ai) מחשבים מתוך הדאטה אימון (סופרים כמה אינסטנסים מתוך הדאטה אימון הם בעלי התכונה ומחלקים במספר הדגימות), מסתמכים על כך שהוא מייצג. את ההסתברות P(x|Ai) נתנו לנו למשל בדוגמה של ה-NBA מפורשות, ובדוגמה של המבחנים והגבהים ניתנה לנו פונקציית הצפיפות עבור הסתברות זו.
- עלינו ללמוד את ההתפלגות של ההסתברות המותנית, נבצע חישוב מתוך הקבוצה המקיימת (התנאי). נדגום את הקבוצה המקיימת את התנאי ומתוכה נחשב את ההסתברות המותנית על ידי הדגימות המקיימות את הפיצ'רים.

מקרה פרטי: עבור קלאסים בעלי prior זהה

- כאשר כל הקלאסים השונים הם **בעלי הסתברויות prior שוות** **P(Ai)=P(Aj)** לכל i,j נוכל לוותר על ה"prior" גם ולקבל = MLC **Maximum Likelihood Classifier**: שמשמעותו נבחר Ai עבור כל i אשר שונה מ-j המקיימים $P(x \mid Ai) > P(x \mid Aj)$
- או **Log-Likelihood Classification**: שמשמעותו נבחר Ai עבור כל i אשר שונה מ-j המקיים $\ln(P(x \mid Ai)) > \ln(P(x \mid Aj))$

דוגמה: דאטת האירוסים של פישר

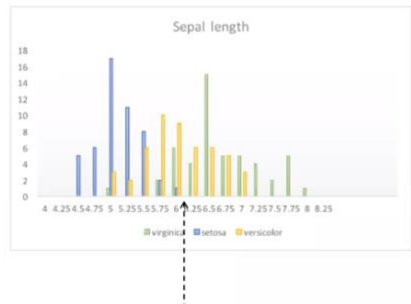


קיימים שלושה סוגים של אירוסים: סטוסה, איריס וירג׳יניקה ואיריס ורסיקולור. אלו הם הקלאסים שלנו ולכן יש לנו 3 קלאסים. נמדדו 4 פיצ׳רים עבור 50 דגימות מכל אחד מהסוגים של הפרחים המצויינים לעיל. ארבעת הפיצ׳רים הם: **אורך ורוחב פטאלי וספאלי** בס"מ. לכן כל דגימה, אינסטנס x, בדאטה אימון שלנו, מחזיקה 4 משתנים (ווקטור של 4 ערכים).

P(sepal length = x | Setosa)

= (count of Setosa w sepal length = x)/(total setosa)

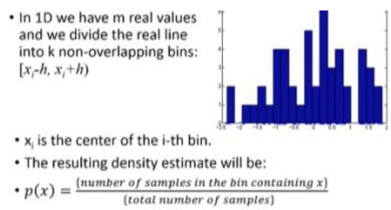
איך נחשב את $P(x \mid Ai)$? הדרך הפשוטה ביותר היא לספור:



What if the sepal length of a new instance is 6.1?

הבעיה בדרך זו היא: שלכאורה מחר יכול להגיע פרח חדש והאורך שלו יהיה 6.1 ס"מ ולא נראה בדאטה שלנו אורך ספאלי כזה, אזי ההנחה היא שההסתברות לפרח כזה על פי ספירה כזו היא 0. אבל בעולם האמיתי, כמובן שיתכן וימצא פרח כזה.

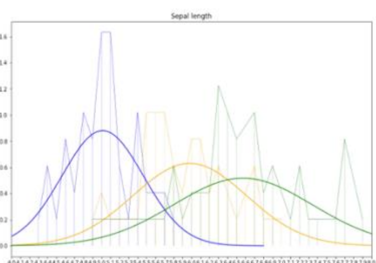
ולכן עלינו "להתחכם" מעט. דרך אחרת לעשות זאת היא להמיר את הדאטה **להיסטוגרמה,** מחלקים את ציר האיקס ל-bin-ים וסופרים כמה מהדאטה שלי נכנס לתוך ה-bin הזה. יש לזה חיסרון כי עלינו לשמור מספרים כמספר ה-bin-ים ששמרנו.

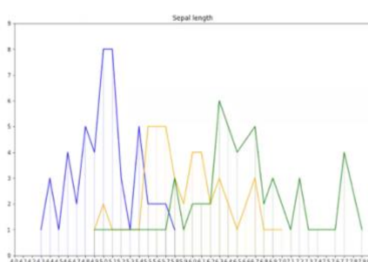


פתרון יותר אלגנטי יהיה:

לנרמל את הדאטה/לבצע קירוב גאוסיאני על הדאטה. ניקח את כל הדאטה הכחול למשל נבצע עליו MLE, נמצא את מיו וסיגמה ובכך ניצור עבורו קירוב גאוסיאני. **להלן דוגמה על הפיצ׳ר אורך ספאלי, מימין הדאטה לפני הנירמול ומשמאל לאחר הקירוב הגאוסיאני.**

תהליך הלמידה הזה אילץ אותנו לשמור 2 משתנים בלבד (מיו וסיגמה) עבור כל סגמנט בדאטה ולכן יש סך הכל 6. ואלה למעשה ה- conditional probabilities שלמדנו מהדאטה, כלומר ה-likelihood.



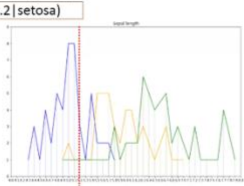


נניח כי מדדנו אורך ספאלי של פרח מסוים וקיבלנו 5.2 ס"מ. איך נדע לסווג לאיזה מבין 3 הסוגים של הפרחים הוא שייך? כאשר אנו משתמשים ב-MAP אנחנו מחפשים אחר הערך הגדול ביותר עבור ההסתברויות הבאות: (posteriors)

- P(versicolor |sepal length = 5.2)
- P(virginica |sepal length = 5.2)
- P(setosa |sepal length = 5.2)

Which one is larger?

1. P(sepal length = 5.2|versicolor)
2. P(sepal length = 5.2|virginica)
3. P(sepal length = 5.2|setosa)



לכן כעת נשתמש בנוסחת בייס כדי לחשב את ההסתברויות הללו. בהנחה שהדאטה מייצג, ה-prior הוא 1/3 עבור כל פלח בדאטה (נלקחו 50 דגימות מכל סוג)ומכיוון שה-prior זהה למעשה נקבע לפי ה-likelihood את הסיווג עבור הדגימה המתוארת: ה-likelihood בנקודה 5.2 הוא הגבוהה ביותר עבור הדאטה הכחול, כלומר בהינתן שהפרח הוא סטוסה, ההסתברות שהאורך הספאלי הוא 5.2 הוא הגבוהה ביותר ועל כן, נסווג על פי הלייקליהוד: **סטוסה**.

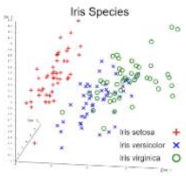
להלן מטריצת הטעות של ניסוי זה ה"confusion matrix"

Classified Species

True Species	versicolor	virginica	setosa
versicolor	31 (20%)	14 (9%)	5 (3%)
virginica	12(8%)	37 (25%)	1 (0.7%)
setosa	11 (7.3%)	0 (0%)	39 (26%)

Wrong classification - error

Correct classification

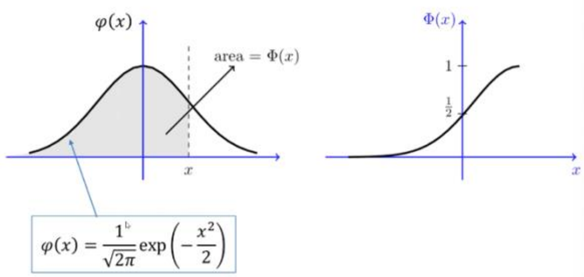


עד כה דיברנו רק על מדד אחד – האורך הספאלי, יש המון אדומים בטבלה ונרצה למתן זאת, עדיין לא השתמשנו בתכונות נוספות מלבד האורך הספאלי. לכן עלינו לפתח את המנגנון שלנו – עלינו ללמוד התפלגויות לא לתכונה בודדת אלא **להסתברויות רב-מימדיות**.

<u>התפלגויות של מספר משתנים – תזכורת / רענון</u>

- הטלת שתי קוביות היא התפלגות רב משתנית. ההתפלגות שלנו מוגדרת מעל מרחב כל הזוגות (i,j) כך ש- i,j = 1,…,6.
- כאשר אנו מניחים ששתי הקוביות הן **הוגנות ובלתי תלויות**, אזי פונקציית ההתפלגות היא יונפורמית מעל 36 תוצאות אפשריות.
- אם אנו לא מניחים שאין תלות, כלומר לכאורה יש תלות, ניתן להגדיר שההתפלגויות השוליות עדיין הוגנות (סוכמות לשישית) אך ההתפלגויות הפנימיות אינן 1/36. כלומר ההתפלגות המשותפת איננה יוניפורמית!

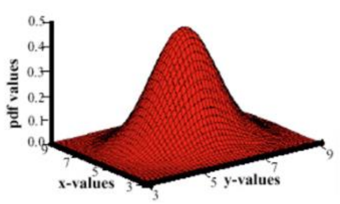
פונקציית הצפיפות הגאוסיאנית



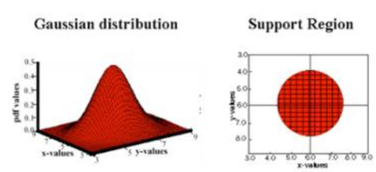
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

נאמר כי איקס מתפלג נורמלית סטנדרטית, אם איקס מתפלג נורמלית ומיו = 0, סיגמה = 1.
האינטגרל של פונקציית צפיפות חד ממדית הוא 1. בפונקציה זו לכל איקס אנו יודעים את הצפיפות.
CDF = פונקציית ההסתברות המצטברת (השטח האפור בגרף מתאר את ההסתברות לערך קטן מאיקס)

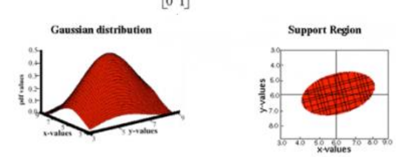
<u>פונקציית הצפיפות הגאוסיאנית עבור וקטור x (רב-מימדית)</u>

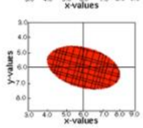


$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}} \exp\left[\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



- Gaussian distribution with identity covariance matrix has equal variances in all directions
- Support region for a Gaussian distribution with identity covariance matrix $C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is a circle





Support region for a Gaussian distribution with covariance matrix $C = \begin{bmatrix} 5 & 0.5 \\ 0.5 & 2 \end{bmatrix}$ is an ellipse rotated at an angle of 9 degrees and -81 degrees with respect to the original axes

- **מיו מייצג את וקטור התוחלות של ההתפלגויות השוליות (d-ממדי)**
- **סיגמה מתאר את מטריצת השונות המשותפת (covariance) מטריצה ריבועית dxd, נדרוש שתהיה הפיכה משום שאנו מבצעים עליה חזקת 1- = inverse. המטריצה הזו היא תמיד סימטרית, חיובית ו-semidefinite.**
- **d מייצג את מספר המימדים.**
- **לפני האקספוננט, במכנה, הסיגמה בערך מוחלט זו הדטרמיננטה של מטריצת השונות המשותפת בחזקת חצי = שורש הדטרמיננטה של מטריצת השונות המשותפת.**
- **x היא דגימה אשר מיוצגת על ידי וקטור d-ממדי**
- **x – miu הוא וקטור d-ממדי מפני שזהו חיסור בין שני וקטורים d-ממדיים, מופיע פעם אחת בצורתו המקורית ולפני מופיע בצורה transpose (ועל כן חזקת t)**

הנפח מתחת לפונקציה הזו הוא 1. ערך הפונקצייה (גובה האוהל) לא תמיד יהיה קטן שווה מאחד, אבל במקרה זה כן (נורמלית סטנדרטית).
פונקציה זו מקבלת וקטור (עם מספר ערכים) ועליה להחזיר סקלאר (כלומר זוהי פונקצייה הממפה מ-R^n ל-R).
הסקלאר הוא הגובה באותה הנקודה וזוהי הצפיפות של הווקטור.
ישנם בסופו של דבר **2 פרמטרים** אשר מגדירים את ההתפלגות הזו גם כאן: בפונקציית צפיפות חד-ממדית מיו מזיז את הפעמון לאורך ציר האיקס (ימינה ושמאלה) וסיגמה מגדירה את הרוחב (פעמון צר או רחב). בפונקציית צפיפות רב מימדית **מיו יקבע את מרכז "האוהל"**. סיגמה = **מטריצת השונות המשותפת, מגדירה את הצורה של "האוהל"**. אם המטריצה היא מטריצת היחידה או אלכסונית בעלת ערכים שווים באלכסון, אזי האוהל יהיה כיפה מושלמת. עבור מטריצה שאינה אלכסונית – הערכים באלכסון יקבעו את היחס בין שני הצירים של האליפסה, הזווית של האליפסה נקבעת על פי הערכים שמחוץ לאלכסון.

<u>דוגמה לחישוב ערך הצפיפות בממד d=2</u>

$$p(x;\mu,\Sigma)=\frac{1}{2\pi\begin{vmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{vmatrix}^{1/2}}\exp\left(-\frac{1}{2}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}^T\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\right)$$

$$=\frac{1}{2\pi(\sigma_1^2\cdot\sigma_2^2-0\cdot 0)^{1/2}}\exp\left(-\frac{1}{2}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}^T\begin{bmatrix}\frac{1}{\sigma_1^2} & 0\\ 0 & \frac{1}{\sigma_2^2}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\right)$$

$$=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right)$$

$$=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right)\cdot\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right).$$

- למעשה אנחנו מחשבים את גובה האוהל שרואה הווקטור x (דו ממדי) ולכן נצפה לקבל סקלאר
- נשים לב כי באקספוננט, אנחנו מכפילים קודם את המטריצה באינברס בווקטור x-miu ואז נקבל ווקטור במימד d שנכפיל אותו בוקטור x-miu בצורת transpose כך שלמעשה מדובר במכפלה פנימית של שני הווקטורים הללו.
- נשים לב כי "הרשינו" לעצמנו לשים במטריצה ערכים ריבועיים (סיגמה1 בריבוע למשל) מהיות שאנו יודעים כי מטריצת השונות המשותפת מחוייבת להיות סימטרית, חיובית ו-semidefinite.
- התקבלה מכפלה של פונקציית צפיפות גאוסיאנית.

נשתמש בכלים שלמדנו כדי באמת לסווג את האירוסים לפי אורך ורוחב פטאלי וספאלי לאחד מ-3 הסוגים. עלינו ללמוד **גאוסיאן 4-ממדי**, לשם כך נלמד 4 miu-ים, לאחר מכן עלינו למצוא MLE ללמידת המטריצה סיגמה. לכן עלינו ללמוד 4 miu-ים לכל סוג פרח, **סיגמה היא סימטרית** ולכן עלינו למצוא 10 פרמטרים (האלכסון, 4 פרמטרים, והערכים מעליו) למטריצה של סוג אחד. סה"כ 14 ערכים לכל סוג ולכן בטוטאל עלינו ללמוד 3*14 ערכים. הרבה פעמים, הדאטה שלנו לא מספיק גדול כדי ללמוד את כל הערכים הללו ולכן אנחנו מניחים אי תלות בגישה שנקראית **Naïve Bayes.**

- MAP $\Rightarrow \arg\max_i P(\mathrm{A_i}\mid\mathbf{x})=\arg\max_i \dfrac{P(\mathbf{x}\mid\mathrm{A_i})P(\mathrm{A_i})}{\sum_{j=1}^{k}P(\mathbf{x}\mid\mathrm{A_j})P(\mathrm{A_j})}$
- Dropping $P(\mathbf{x})\Rightarrow \arg\max_i\{P(\mathbf{x}\mid\mathrm{A_i})P(\mathrm{A_i})\}$
- ML - Assuming $P(A_i)=P(A_j)\Rightarrow \arg\max_i\{P(\mathbf{x}\mid\mathrm{A_i})\}$
- Using log probability $\Rightarrow \arg\max_i\{\ln P(\mathbf{x}\mid\mathrm{A_i})+\ln P(\mathrm{A_i})\}$
- <u>Now</u> : Naive Bayes - assuming $P(\vec{\mathbf{x}}\mid\mathrm{A_i})=\prod_j P(x_j\mid\mathrm{A_i})\Rightarrow$ $\arg\max_i\{P(\mathrm{A_i})\prod_j P(x_j\mid\mathrm{A_i})\}$

<u>נגדיר אי תלות מותנית:</u>

דוגמה: מספר הסרטים שהבנאדם צפה במשך חייו לעומת רמת הכולסטרול בדם, מבחינת כל הדאטה וכל הגילאים שני המשתנים המתוארים הם **תלויים בגיל**. **הם בלתי תלויים בהינתן הגיל של הבנאדם**, ככל שאתה מבוגר יותר ככל הנראה שצפית ביותר סרטים וכן ככל שאדם מבוגר יותר הסבירות שרמת הכולסטרול בדם שלו הינה גבוהה משל אדם צעיר. לכן כאשר אנו ממקדים את קבוצת המדגם שלנו בקבוצת גיל מסויימת (למשל רק אנשים בני 30-32) הם בלתי תלויים, ועל כן הם **משתנים בלתי תלויים בתנאי**.

<u>סימון:</u> x בלתי תלוי ב-y בהינתן הקלאס (גיל וכו'...) $X\perp Y\mid C$

<u>הגדרה:</u> הפיצ'רים הינם בלתי תלויים בתנאי (conditionally independent) בהינתן הקלאס, אם לכל וקטור (רב-מימדי) ערכי פיצ'רים x ולכל ערכי הקלאסים האפשריים i, מתקיים:

$$P\big((x_1,x_2,\dots,x_d)\mid A_i\big)=\prod_{j=1\dots d}P\big(x_j\mid A_i\big)$$

ההסתברות של חיתוך הערכים בהינתן הקלאס שווה למכפלת ההסתברויות של כל אחד מהערכים בהינתן הקלאס. זוהי הנחת אי תלות. לכן עלינו ללמוד עבור כל מכפלה בצד ימין גאוסיאן חד ממדי ולכן לכל מכפלה יש ללמוד 2 פרמטרים (צד ימין). בצד שמאל עלינו ללמוד 55 פרמטרים. לכן ברור שהלמידה של צד ימין הינה יותר יעילה!

בסיווג על ידי Naïve Bayes אנו עושים הנחה שימושית ומפשטת, כי ערכי הפיצ'רים הן בלתי תלויים בתנאי בהינתן הקלאס. **אבל הנחת אי תלות בתנאי הינה לא תמיד נכונה**. ישנה דוגמה בשיעורי הבית.

Learning

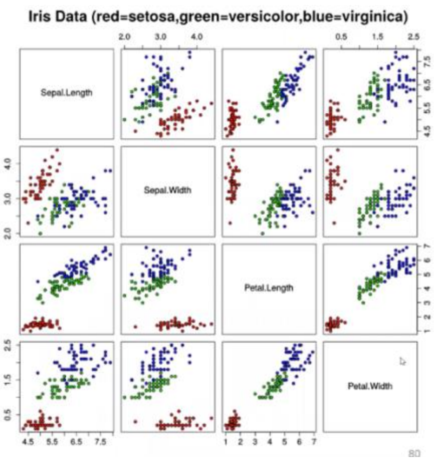
Naive_Bayes_Learn(D – examples)
For each target class A_i
$\hat{P}(A_i)\leftarrow$ estimate $P(A_i)$ from D
For each attribute value $x_j\in V_j$
$\hat{P}(x_j\mid A_i)\leftarrow$ estimate $P(x_j\mid A_i)$ from D

Execution

Naive_Bayes_Classify(x)
return $A_{NB}=\arg\max_i \hat{P}(A_i)\prod_{x_j\in x}\hat{P}(x_j\mid A_i)$

האלגוריתם הלומד של נאיב בייס Naïve Bayes Algorithm

- עלינו ללמוד הסתברויות: נרצה לקחת את הדאטה, להסיק ממנו את ה-priors.
- ולכל ערך באטריביוט ללמוד גאוסיאן חד מימדי (2d פרמטרים המגדירים את הגאוסיאן)



הדאטה של פישר (scatterplots של הדאטה אודות האירוסים):

- **האם האורך הפטאלי והרוחב הפטאלי הם בלתי תלויים בתנאי?**
- ראשית כל נשים לב, הם אינם בלתי תלויים מפני שדאטה בלתי תלוי נראה כמו ענגה, וכאן נראה יחסית קו ישר. אז אינם בלתי תלויים
- אבל האם הם בלתי תלויים בתנאי? בשביל שהם יהיו בלתי תלויים בתנאי – שכל אחד מהצבעים יראו כמו עננים. הענן הירוק דומה מידי לקו, הכחול והאדום יחסית נראים מעוננים.
- **גם אורך פטאלי ורוחב ספאלי אינם בלתי תלויים בתנאי** משום שכל אחד מהצבעים נראה במגמה של קו ולאו דווקא כענן מפוזר.
- **למרות שנראה שאין כאן בהכרח אי תלות בתנאי, גישת נאיב בייס עדיין הייתה עובדת על הדאטה של פישר.**

לרוב, ההנחה הנאיבית מופרת (למשל כפי שנאמר לעיל רוחב פטאלי ואורך פטאלי אינם נראים בהכרח בלתי תלויים בתנאי מפני שאם היה שוויון היה נראה יותר ענני) : $\hat{P}(x_1, x_2 \ldots x_n \mid A_j) \neq \prod_i \hat{P}(x_i \mid A_j)$

אבל, בפרקטיקה, המשערך הזה עובד באופן מפתיע דיי טוב. נשים לב, כי בפועל, אנחנו לא באמת צריכים שההנחה הזו תתקיים ותהיה נכונה. אנו רק צריכים שהבא יתקיים (תנאי יותר חלש) :

$$\boxed{\arg\max_j \hat{P}(A_j)\prod_i \hat{P}(x_i \mid A_j) = \arg\max_j \hat{P}(A_j)\hat{P}(x_1 \ldots, x_n \mid A_j)}$$

Naïve Bayes Full Bayes

אנחנו צריכים שה-j שמתקבל בנאיב בייס ומקיים הסתברות מקסימלית, יהיה אותו ה-j שמתקבל ללא ההנחה של נאיב בייס (בפול בייס). כלומר אנחנו לא צריכים שכל אחד מהאיברים j יהיה שווה, אם אכן מתקיים השיוויון, אז ברור שהמקסימום מתקבל באותו המקום וערכו זהה. אנחנו צריכים שהמקסימום יתקבל עבור j שהוא מתקבל בשתי הגישות וזו דרישה הרבה יותר חלשה.

בשיעורי בית נראה מקרה בו נאיב בייס באמת לא יעבוד.