למידה חישובית – שיעור 4

Bayesian Learning

תחום בלמידה חישובית המשתמש בכלים הסתברותיים לפתור בעיות קלסיפיקציה.

תזכורת, נוסחת בייס

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B|A)}{P(B)}$$

פרדוקס סימפסון

אם נביט על כל קטגוריה בנפרד, הנשים קולעות טוב יותר ולכן נסיק שהן ינצחו, אך אם נסכום את כל הקטגוריות יחדיו נקבל תמונה אחרת בה הגברים קולעים טוב יותר.

	Less than 1.70m	1.70-1.90	Taller than 1.90
Women	4/6	4/6	8/9
Men	1/2	1/2	23/27

נרצה למצוא classifier שייתן לנו חלוקה טובה ככל שניתן. נמדוד את טיב החלוקה על ידי הבאת הטעות למינימום ומדידה באמצעות כלים הסתברותיים. נתייחס אל הfeatureים שלנו כאל משתנים מקריים, כלומר בהינתן feature כלשהו נמצא לו הסתברות לכמה הדיוני שהמתנה שייך ל-class כלשהו.

Prior Probability Only

מה יותר סביר שהclass.

אם יש לי 2 classים: A ו-B. בנוסף, אנו יודעים את ההסתברויות $P(A), P(B)$. נוכל לסווג כך ש:

- אם $P(A) > P(B)$ נסווג ה-instance שלנו להיות A
- אחרת, נסווג את ה-instance שלנו להיות B.

בטכניקה הזו אנו לא משתמשים במידע כלשהו על ה-instance שלנו. אנו רק משתמשים בידע הקודם שיש לנו.

ההסתברות לטעות אם כן היא: $1 - P(B)$

Likelihood

זוהי גישה מתקדמת יותר, בה אנו מודדים את הסבירות לסיווג (חיובי או שלילי) בהתאם לפרמטר כלשהו. לדוגמה: אם נקבל פרמטר גובה= 1.60 נאמר שהסבירות שה-instance הוא אישה, גבוה יותר. יחס הlikelihood **אינו** מדויק כמובן והוא נותן הערכה גסה הקשורה לסבירות של ה-instance לקבל פרמטר כלשהו.

נוכל להשתמש בLikelihood בהנחה ואנו יודעים: $P(x|A)\ and\ P(x|B)$.

הבעיה: אנו מעוניינים לדעת מה ההסתברות ל-class בהינתן ה-instance, ולכן להפך, כלומר את: $P(A|x)\ and\ P(B|x)$. כי אנו מעוניינים להחליט איזה class הוא יותר probable.

MAP -Maximum A- Posteriori

יש לנו מידע על $P(x|A)\ and\ P(x|B)$ ואנו מעוניינים לקבל מידע על $P(A|x)\ and\ P(B|x)$, על מנת להחליט מהו ה-class value של ה-instance הנתון. על מנת לקבל את האינפורמציה הדרושה לנו, נשתמש בנוסחת בייס

$$P(A|x)=\frac{P(x|A)\cdot P(A)}{P(x)}$$

כאשר:

- $\boldsymbol{P(A)}$ - *Prior*. הידע שיש לי את ה-class, בלי קשר לinstance.
- $\boldsymbol{P(x|A)}$ - *Likelihood*.
- $\boldsymbol{P(A|x)}$ - *posterior*. בזה נשתמש כדי לעשות קלסיפיקציה

חוק בייס:

נשים לב שאנו מחשבים את $P(A|x)$ ו-$P(B|x)$, כך ששניהם מחולקים ב-$P(x)$, לכן אנו יכולים "להיפטר" מהמכנה ולחשב:

- אם $P(x|A)\cdot P(A)>P(x|B)\cdot P(B)$, נסווג את ה-instance עם A
- אחרת נסווג עם B.

מה ההסתברות לטעות? (Minimum Error Rate Class)

נרצה לדעת מה ההסתברות לטעות במידה ואנו מסווגים את x.

- אם בחרנו B, ההסתברות לטעות: $P(error)=P(error|x)=P(A|x)$
- אם בחרנו A, ההסתברות לטעות: $P(error)=P(error|x)=P(B|x)$

כלומר נוסחת בייס מצמצמת לי את ההסתברות לטעות. נקבל:

$$P(error|x)=\min\big(P(B|x),P(A|x)\big)$$

<u>הערה:</u> אנחנו עושים מינימום ל- $P(B|x),P(A|x)$ כי אנחנו בוחרים בשביל הקלסיפיקציה את המקסימום של הערכים הללו.

בהנחה ויש לנו k classים, נחשב:

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

ושוב, נוכל להשמיט את $P(x)$ וסה"כ נרצה לחשב:

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

מה ההסתברות לטעות? (Cost of Wrong Decision)

נשתמש בפונקציית loss הנקראת zero one loss. שעובדת תחת ההנחות הבאות שה-classifier ייתן לנו את המינימום טעות:

$$\lambda_{ij} = \lambda\left(chosse\ A_i \middle| A_j\right) = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

כלומר אם $i = j$ אז שיעור הטעות שלי הוא 0 ואני לא אשלם כלום, כי בעצם בחרתי ב$class\ A_i$.
אם $i \neq j$ אז שיעור הטעות שלי הוא 1 ואני אשלם 1, כי בחרתי ב$class$ אחר.

ה**Risk** מוגדר להיות הexpected loss שלי, כלומר כמה אני מצפה לטעות במידה ועשיתי קלסיפיקציה שלי. לדוגמה כאשר יש לי רק 2 classים: A ו-B. אם בחרתי ב-A, הexpected loss שלי היא רק ההסתברות שx שייך ל-B.

בהנחה ויש לי k classים, הטעות תהיה:

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

השוויון נובע מכך שאנו משתמשים בפונקציית ה-one zero loss. אם $i = j$ הפונקציה תניב 0 ולכן לא רלוונטי לחישוב

השוויון נובע מתכונת סכום כל ההסתברויות שווה 1

(סוכמים את כל ההסתברויות). פונקציית R הינה פונקציית ה-expected loss שלי, אם יש לי k classים.

ושוב אנו רואים שכאשר נבחר את הclass שנתן לנו את הערך הposterior הגבוה ביותר, קיבלנו את ה error הנמוך ביותר.

כדי למצוא את ה-class של x, כלומר לעשות קלסיפיקציה, נשתמש בנוסחה:

$$g_i(x) = \mathrm{P}(A_i|x) = \frac{P(x|A_i)P(A_i)}{\sum_{j=1}^{k} P\left(x\middle|A_j\right)P(A_j)}$$

וכרגיל, נוכל להשמיט את המכנה ולקבל: $g_i(x) = P(x|A_i)P(A_i)$.
$g_i(x)$ נותן לי ערך פר class. נבחר את ה-i שייתן לי את הערך הגבוה ביותר ל-$g_i(x)$

כעת, מהיות ln פונקציה מונוטונית עולה, נוכל להגדיר:

$$g_i(x) = \ln\left(P(x|A_i)\right) + \ln\left(P(A_i)\right)$$

*בפועל כשנחשב את ההסתברות, זה יהיה נוח יותר לחשב עם ln. זה יותר יציב נומרית, כלומר בעת חישוב הטעות במחשב יש פחות סיכוי שהמספרים יהפכו ל-0 בשלב כלשהו בחישוב. קורה לעיתים קרובות בפייתון ולכן נוח להימנע מהבעיה הזו על ידי העלאת הפונקציה ב-ln.

Maximum Likelihood Classifier

מעין מקרה פרטי של שימוש בנוסחת הlikelihood.

נשתמש בה כאשר לclassים שונים, יש prior probabilities **זהה**, כלומר כאשר $P(A_i) = P(A_j)$ לכל i, j, אז נוכל להשמיט גם את מונח ה-prior ולקבל מסווג מקסימלי (ML- Maximum Likelihood Classifier). במצב זה נסתכל אך ורק על ה-$likelihood$ כלומר על $P(x|A_i)$.

(או שנוכל להסתכל על $\ln(P(x|A_i))$

<u>הערה:</u> זה מקרה פרטי בו לכל הclassים יש את אותה ההסתברות (ולא את אותם הערכים).

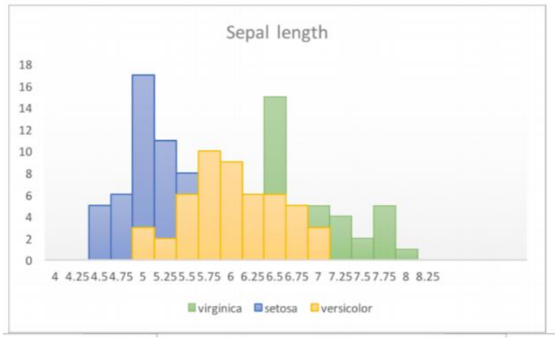
<u>הערה:</u> חישוב $P(x|A_i)$ הוא בדיוק מה שאמרנו בתחילת השיעור שלא מניב לנו את מה שאנו מחפשים, אבל אנו יכולים להסתמך על חישוב זה, בעקבות ההנחה ש: $A_i = A_j, \forall i, j$.

כיצד נוכל לשערך את ההסתברות של משתנה מסוים (feature) ב-data שלי?

1. היסטוגרמה

היסטוגרמה היא הצגה גרפית של הנתונים כמלבנים, כך שכל מלבן מייצג $class$ אחר. בין כל 2 classים אין חפיפה. שטח כל מלבן מייצג את השכיחות היחסית המתאימה לו.

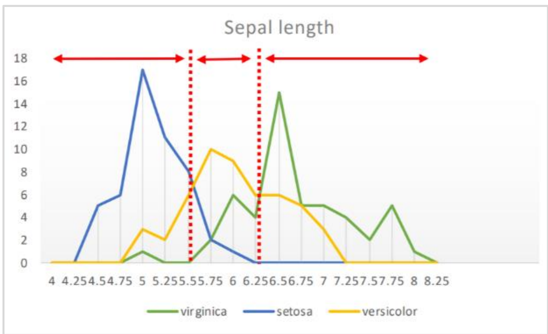
נבנה היסטוגרמה לכל class ונוכל לחשב את ההסתברות (likelihood) לכל כל צבע בclass כלשהו.



2. אינטרפולציה לינארית

שיטה בה אנו מייצרים עקומה לינארית, על ידי חיבור שתי נקודות עוקבות בקו ישר.

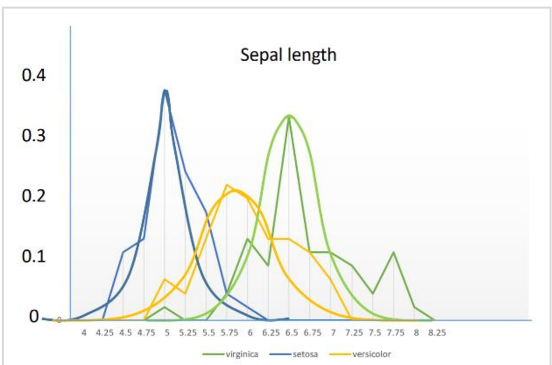
השיטה מאפשרת לי לבנות נקודות חדשות בטווח הנקודות הקיימות שאני כבר מכירה.



3. שיערוך ההסתברויות כהתפלגות נורמלית

בשיטה זו נחשב את הממוצע וסטיית התקן של כל אחד מה-classים שלי ונחשב באמצעותם את ההתפלגות הנורמלית של כל class.

כעת אנו צריכים רק לדעת 2 פרטים על כל class: התוחלת והשונות, ללא תלות בכמות הדוגמאות או כמות ה-classים שיש לי.



איך נמדוד את הclassifier שבנינו?

נשתמש ב-Confusion Matrix:

מטריצה המייצגת לי מה הערך של ה-instance לעומת הפרדיקציה שה-classifier סיפק לי.

	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%)

הערך האמיתי של ה-instance.

הפרדיקציה שהתקבלה מה-classifier.
כל הסיווגים הנכונים של ה-classifier יהיו באלכסון.

באמצעות המטריצה אנו יכולים לראות כמה דוגמאות הצלחנו לסווג נכון, וכמה לא. (סיווג נכון בירוק, סיווג לא נכון באדום).

המטריצה מייצרת לנו הצגה מספרית נוחה שתעזור לנו להבין על איזה הinstanceים נרצה לאסוף עוד מידע. בדוגמה שלנו, אנו יכולים לראות שאנו כמעט ולא טועים בין הvirginica לבין ה-setosa אבל אנו רואים שרוב הטעויות שלנו קשורות לסיווג ל-versicolor. לכן נסיק מכך שעלינו לאסוף מידע נוסף על ה-versicolor כדי שנוכל לדייק בצורה טובה יותר את הקלסיפיקציה על הclass הזה.

Cost of Misclassification

יהיו מקרים בהם אנו נרצה שלטעות אחת תהיה **משקל גבוה** מטעות אחרת. לדוגמה FP לשאלה האם אדם חולה במחלה, גרוע יותר מFN לשאלה זו. במקרים כאלה נשתמש בנוסחת המחיר הבאה:

$$\underset{i}{\operatorname{argmin}} \sum cost(A_i|A_j)P(A_j|x)$$

ייתכן שיהיו מקרים בהם נעדיף classifier אחד על פני אחר כיוון שהוא ייתן לי את הCost הממושקל הנמוך ביותר (גם אם $P(A) < P(B)$ ועדיין נבחר ב$Class\ A$, כי ה$Cost$ בו נמוך יותר).