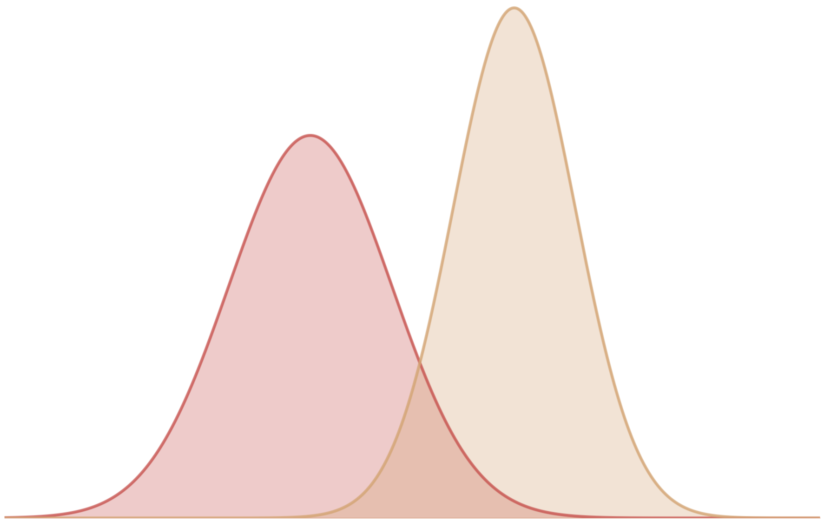


מסווג בייס

יסודות מתמטיים ויישומים מעשיים

Bayes Classifier: Mathematical Foundations and Practical Applications

ד"ר יורם סגל

Dr. Yoram Segal

28-11-2025

מסווג בייס

יסודות מתמטיים ויישומים מעשיים

ד"ר יורם סגל

תוכן העניינים

פרק 1

הכומר שחשב על העתיד

1.1 החלטות

מאות פעמים ביום, מבלי שנשים לב, אנו מבצעים פעולות סיווג. כשנכנסת שיחה בטלפון -- להרים או לדחות? כשמגיע מייל -- לקרוא או למחוק? כשפוגשים אדם ברחוב -- לחייך או להתעלם? כל אחת מהבחירות הללו היא, בעצם, תהליך של סיווג מידע לקטגוריות. המוח האנושי עושה זאת באופן אוטומטי, בלי שנידרש לחשוב על הפרטים הקטנים. אבל כשאנו רוצים ללמד מכונה לבצע את אותן החלטות, עלינו להבין את המנגנון הבסיסי שמאחוריהם.

כדי להמחיש את התהליך, נעיין בתרשים פשוט המתאר את מהות הסיווג. בליבת כל תהליך סיווג עומדת פונקציה מתמטית שמקבלת מידע ומחזירה החלטה.

איור 1.1: זרימת תהליך הסיווג ממאפייני קלט להחלטת פלט

כפי שניתן לראות באיור 1.1, התהליך מורכב משלושה רכיבים מרכזיים: משמאל מגיעים **מאפייני הקלט** (Input Features), המסומנים כוקטור x. מידע זה זורם אל תוך המסווג עצמו -- הקופסה האדומה המרכזית המכילה את הפונקציה $f(x)$. בסופו של דבר, המסווג מייצר **פלט** (Output Class), המסומן כ-$\hat{y}$, המייצג את ההחלטה הסופית. החצים האדומים מדגישים את כיוון הזרימה -- ממידע גולמי, דרך עיבוד אלגוריתמי, אל החלטה ברורה. התהליך הזה, פשוט ככל שייראה, הוא הבסיס לכל מערכת למידת מכונה שמבצעת סיווג.

בבנק, למשל, נציג צריך להחליט: האם לאשר הלוואה ללקוח מסוים? השאלה האמיתית היא -- מה הסיכון שהלקוח לא יחזיר את הכסף? המסווג, במקרה זה, מקבל מאפיינים כמו גיל, הכנסה, היסטוריית אשראי, ומחזיר החלטה: לאשר או לדחות.

הגדרה

מסווג הוא פונקציה שמקבלת וקטור מאפיינים ומחזירה החלטה:

$$\hat{y} = f(\mathbf{x}) \tag{1.1}$$

1.2 הכומר מלונדון

אם ניסה אדם מן המאה ה-18 להבין איך מכונות יסווגו מידע בעתיד, הוא היה נחשב לאדם תמוה. אבל בדיוק זה מה שעשה כומר אנגלי שקט בשם Thomas Bayes. הוא לא חשב על מכונות, כמובן, אלא על שאלה פילוסופית עמוקה יותר: איך אנחנו יודעים מה אנחנו יודעים? איך מידע חדש משנה את הבנתנו לגבי העולם?

רעיונותיו של בייס לא פורסמו בחייו. רק בשנת 1763, שנתיים לאחר מותו, חברו ריצ'רד פרייס פרסם את המאמר שלו. אותה תקופה, כשרוכבים על סוסים עברו ליד חלונו, הם ודאי תהו מה עושה הכומר המוזר הזה עם כל המספרים שהוא רושם לעצמו. הם לא יכלו לדעת שהרעיונות האלה ישמשו יום אחד לסינון דואר זבל, לאבחון מחלות, ולזיהוי טרוריסטים בשדות תעופה.

אבל הרעיון של בייס לא נשאר מוגבל למעבדה האקדמית. הוא התפתח, הורחב, והפך לכלי מרכזי במדע המודרני. כדי להבין את המסע ההיסטורי הזה, נעיין בציר הזמן המתאר את האבולוציה של משפט בייס.

איור 1.2: ציר הזמן של משפט בייס -- מהמסה המקורית ועד ליישומים מודרניים

כפי שניתן לראות באיור 1.2, ציר הזמן מתחיל בשנת 1763 עם פרסום המסה המקורית של בייס (Bayes Essay). כמעט חמישים שנה מאוחר יותר, בשנת 1812, המתמטיקאי הצרפתי פייר-סימון לפלס (Laplace) הרחיב את הרעיון והפך אותו לכלי מתמטי מתוחכם יותר. נדרשו עוד מאה ועשרים שנה עד שבשנת 1936 הסטטיסטיקאי רונלד פישר (Fisher) פיתח את שיטת ה-LDA (ניתוח דיסקרימיננטי ליניארי), שהשתמשה ברעיונות דומים לצורך סיווג. הקפיצה הגדולה הבאה הגיעה עם עידן המחשוב: בסוף שנות ה-90, בשנת 1998, התחילו להשתמש במסווג בייס לסינון דואר זבל (Spam Filtering), והרעיון הישן הזה מצא בית חדש בעולם האינטרנט. ולבסוף, בשנת 2020,

הסמל האחרון על ציר הזמן מייצג את הלמידה המודרנית (Modern ML) -- עידן שבו מסווג בייס הוא אבן יסוד בכל קורס למידת מכונה.

השאלה שטרדה את בייס

אם אני יודע את ההסתברות לתוצאה בהינתן סיבה -- האם אוכל להסיק את ההסתברות לסיבה בהינתן התוצאה?

זוהי הפיכת ההיגיון. בייס לא שאל "מה יקרה אם...?" אלא "מה קרה בגלל...?" הוא רצה לחשב לאחור, מהתוצאה אל הגורם. זה בדיוק מה שאנחנו עושים כל הזמן כבני אדם -- אנחנו רואים תוצאה ומנסים להסיק מה גרם לה.

1.3 המאבטח בנתב"ג

נחשוב על מאבטח בשדה התעופה בן-גוריון. יש לו רשימת שאלות קבועות, ומושג שנקרא "פרופיילינג". הוא מחפש סימנים: מבטע, מראה, התנהגות. לפני שהוא שומע מילה אחת מפיו של הנוסע, כבר יש לו השערה מוקדמת על פי הנתונים הסטטיסטיים -- מה ההסתברות שאדם מסוים מהווה איום.

זוהי הדעה הקדומה -- **הפריור** (Prior). לפני שהמאבטח יודע דבר על הנוסע הספציפי, יש לו השערה מוקדמת שמבוססת על ידע כללי. זה לא נובע משנאה או דעה אישית, אלא ממידע סטטיסטי שנצבר במשך שנים.

עכשיו הנוסע עונה על שאלות. "לאן אתה טס?" "למה?" "כמה זמן?" "עם מי?" כל תשובה היא חתיכת מידע חדשה -- חתיכת **ראייה** (Evidence). השאלה שעומדת עכשיו היא: האם הראייה מאששת או סותרת את ההשערה המקורית?

אם הנוסע עונה בשקט, בביטחון, עם פרטים עקביים -- הראייה מפחיתה את החשד. ההסתברות שהוא מהווה סכנה יורדת. אבל אם הוא מגמגם, סותר את עצמו, נראה עצבני -- הראייה מעלה את החשד. זה התהליך של עדכון האמונה שלנו -- מה שבייס הגדיר כ**פוסטריור** (Posterior): ההסתברות המעודכנת לאחר שראינו את הראיות.

תובנה מרכזית

במכונת בייס, אנחנו לא שואלים "מה דעתך על הבן אדם הזה?" אלא מניחים השערה ובודקים אם הראייה תומכת בה או לא.

1.4 חתול או כלב?

נניח שאני רוצה לבנות מכונה שמפרידה בין תמונות של חתולים לכלבים. איך בייס יעבוד? הסוד הוא בהשערת האפס. המכונה תמיד מניחה שכל תמונה היא חתול -- זו השערת ברירת המחדל, הפריור שלנו. תפקידה של המכונה הוא לאשש או לסתור את ההנחה הזו.

עכשיו המכונה מסתכלת על התמונה. יש אוזניים מחודדות? זנב ארוך? ציפורניים נשלפות? כל אחת מהתכונות הללו היא ראייה. המכונה שואלת את עצמה: "אם אני מניח שזה חתול, מה ההסתברות שאני אראה את המאפיינים האלה?" ואז היא שואלת: "אם אני מניח שזה כלב, מה ההסתברות שאני אראה את המאפיינים האלה?"

אם ההסתברות שהתמונה היא חתול נמוכה מאוד בהינתן הראיות -- המסקנה היא שזה כלב. לא שהמכונה "יודעת" שזה כלב בצורה ישירה, אלא שהיא שוללת את האפשרות שזה חתול. זה הפרדוקס של מסווג בייס: הוא לא מחפש מה משהו הוא, אלא מה הוא לא.

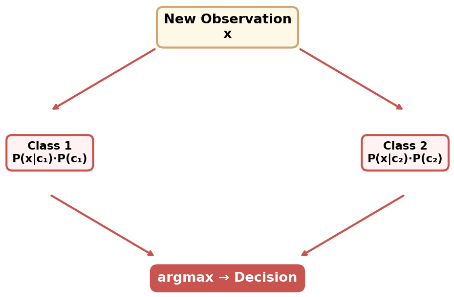
1.5 הקוביות המוטות

"בוא נשחק שש-בש," אומר לך חבר ומוציא קוביות. השערת האפס שלך היא שהקוביות הוגנות -- כי ברוב המקרים זה מה שקורה. אין לך סיבה לחשוד שהחבר שלך מרמה. זה הפריור שלך -- ההנחה הראשונית.

אבל אחרי עשרים משחקים שבהם החבר מנצח, הראייה מצטברת. ההסתברות שהקוביות באמת הוגנות -- הולכת ויורדת. כל זריקה נוספת שמסתיימת בניצחון שלו היא ראייה נוספת נגד ההנחה המקורית. ובשלב מסוים, ההסתברות המעודכנת (הפוסטריור) נמוכה כל כך, שאתה מחליט לבדוק את הקוביות.

זה בדיוק מה שבייס עושה: מעדכן את האמונה שלנו בהתאם לראיות החדשות. כל מידע חדש משנה את התמונה. אנחנו לא מתעקשים על האמונה המקורית שלנו -- אנחנו פתוחים לשנות דעה.

כדי להבין איך בייס מבצע את ההחלטה הסופית, נעיין בתרשים המתאר את תהליך הסיווג הבייסיאני.



איור 1.3: תרשים תהליך ההחלטה במסווג בייס -- מתצפית חדשה לבחירה בין מחלקות

כפי שניתן לראות באיור 1.3, התהליך מתחיל בקופסה העליונה המכילה **תצפית חדשה** (New Observation), המסומנת כ-x. משם, התרשים מתפצל לשני ענפים:

משמאל, המסווג מחשב את המכפלה $P(x|c_1) \cdot P(c_1)$ -- כלומר, את ההסתברות לראות את הנתון x בהינתן שהוא שייך למחלקה c_1, כפול ההסתברות הכללית למחלקה c_1. מימין, נעשה אותו חישוב עבור מחלקה c_2. החצים מובילים את שני החישובים הללו אל הקופסה התחתונה האדומה, שבה מתבצע פעולת argmax -- המסווג בוחר את המחלקה בעלת המכפלה הגבוהה ביותר. זהו תהליך קבלת ההחלטה (Decision): התצפית מסווגת למחלקה שעבורה ההסתברות המעודכנת היא הגבוהה ביותר. התרשים ממחיש בצורה ברורה את הפילוסופיה של בייס -- אנחנו משווים בין השערות, ובוחרים את זו שהראייה תומכת בה בצורה החזקה ביותר.

1.6 מה למדנו עד כה

לפני בייס, בקורס למידת מכונה זה, עברנו על אלגוריתמים רבים. כל אחד מהם פותר בעיה מסוימת בעולם הסיווג והחיזוי. **רגרסיה לינארית**, למשל, עוזרת לנו לחזות ערכים רציפים -- כמו מחיר בית על פי גודלו. **רגרסיה לוגיסטית** לוקחת את הרעיון של רגרסיה ומתאימה אותו לבעיות סיווג, באמצעות הפונקציה הסיגמואידית שממפה ערכים לקטע $[0, 1]$ ומאפשרת לנו להחליט בין שתי קטגוריות. **K-Nearest Neighbors** מסתכל על התצפיות הקרובות ביותר במרחב הנתונים ומסווג לפי רוב הקולות של השכנים. **K-Means** פועל בצורה שונה -- הוא מבצע אשכול לא מונחה, ומקבץ נתונים לפי קרבה גיאומטרית בלי לדעת מראש מהן הקטגוריות. **PCA** (ניתוח רכיבים עיקריים) עוזר לנו להוריד ממדים תוך שמירה על השונות המרבית, וזאת על ידי מציאת הכיוונים שבהם הנתונים משתנים הכי הרבה. **LDA** (ניתוח דיסקרימיננטי לינארי) דומה ל-PCA, אך מטרתו שונה -- הוא מחפש את הכיוון שבו ההפרדה בין הקבוצות מקסימלית, לא רק השונות.

PCA לעומת LDA

PCA מחפש את הכיוון שבו השונות הכוללת מקסימלית.
LDA מחפש את הכיוון שבו ההפרדה בין הקבוצות מקסימלית.

עכשיו הגיע תורו של בייס -- שיטה שנולדה במאה ה-18, אך רלוונטית היום יותר מתמיד. מה שמייחד את בייס הוא לא רק היעילות המתמטית, אלא הגישה הפילוסופית: אנחנו מתחילים עם השערה, ומשנים את דעתנו בהתאם לראיות. זו דרך חשיבה שמתאימה לא רק למכונות, אלא לבני אדם.

פרק 2

שפת האי-ודאות

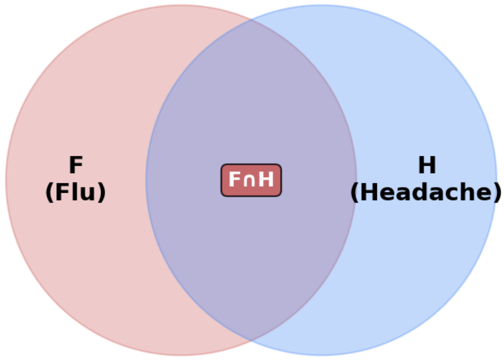
2.1 מספרים בין אפס לאחד

הסתברות היא מספר בין 0 ל-1 שמתאר את מידת האמונה שלנו שמשהו יקרה. אפס -- בלתי אפשרי. אחד -- ודאי. וביניהם -- כל הספקות של החיים.

2.2 שפעת וכאב ראש

בואו נחשוב על שתי תופעות: שפעת וכאב ראש. זו תהיה נקודת המוצא שלנו להבנת הקשר בין אירועים. דמיינו עיגול אחד שמייצג את כל האנשים שחולים בשפעת, ועיגול שני שמייצג את כל האנשים שסובלים מכאב ראש. האם העיגולים האלה נפרדים לחלוטין? האם הם חופפים? ואם כן -- עד כמה?

באוכלוסייה, נניח ש-2.5% מהאנשים חולים בשפעת. זו **הסתברות שולית** -- מידע על העולם, בלי שום ראייה נוספת. במקביל, נניח ש-10% מהאנשים סובלים מכאב ראש. גם זו הסתברות שולית -- עובדה סטטיסטית על האוכלוסייה.



איור 2.1: עיגולי ההסתברות

כפי שניתן לראות באיור 2.1, שני העיגולים חופפים חלקית. העיגול השמאלי (הוורוד) מייצג את כל חולי השפעת -- F (Flu). העיגול הימני (הכחול) מייצג את כל הסובלים מכאב ראש -- H (Headache). האזור שבו שני העיגולים נחתכים, המסומן ב-$F \cap H$, מייצג את האנשים שחולים גם בשפעת וגם סובלים מכאב ראש. זהו הייצוג הגיאומטרי של הסתברות משותפת -- החיתוך בין שני אירועים.

הסתברות שולית

$$P(\text{שפעת}) = 0.025 \tag{2.1}$$
$$P(\text{כאב ראש}) = 0.1 \tag{2.2}$$

2.3 וגם

אם אני שואל "מה הסיכוי שלאדם יש גם שפעת וגם כאב ראש?" -- אני מחפש את האזור שבו שני העיגולים חופפים. אבל כאן יש הבדל עצום בין שני מקרים:

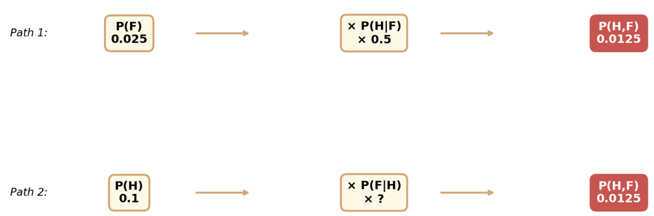
מקרה א': אני שואל על כל האוכלוסייה -- מה הסיכוי שאדם אקראי יהיה גם חולה שפעת וגם יכאב לו הראש?

מקרה ב': אני כבר יודע שמישהו חולה בשפעת, ועכשיו אני שואל -- מתוך חולי השפעת, כמה סובלים מכאב ראש?
ההבדל הזה הוא הלב של בייס.

2.4 גזירת העולם

דמיינו את כל האנושות כמעגל גדול. בתוכו, מעגל קטן יותר של חולי שפעת. ומעגל נוסף של אנשים עם כאב ראש. כשאני שואל על גם וגם ביחס לכל העולם -- המכנה הוא כל האנושות. אבל כשאני אומר "בהינתן שהאדם חולה בשפעת" -- אני **גוזר את העולם**. אני זורק החוצה את כל מי שלא חולה בשפעת, ונשאר רק עם המעגל הקטן.

כיצד מגיעים להסתברות המשותפת הזו? ישנם שני מסלולים אפשריים לחשב את אותה התוצאה. המסלול הראשון מתחיל מהסתברות השפעת ומכפיל אותה בהסתברות לכאב ראש בהינתן שפעת. המסלול השני מתחיל מהסתברות כאב הראש ומכפיל אותה בהסתברות לשפעת בהינתן כאב ראש. שני המסלולים מובילים לאותה נקודה -- ההסתברות המשותפת.



איור 2.2: כלל השרשרת

כפי שניתן לראות באיור 2.2, שני המסלולים מוצגים באופן חזותי. בנתיב העליון (Path 1), מתחילים מ-$P(F) = 0.025$ -- הסתברות השפעת באוכלוסייה. החץ השני מכפיל ב-$P(H|F) = 0.5$ -- ההסתברות לכאב ראש בקרב חולי שפעת. התוצאה: $P(H, F) = 0.0125$ מופיעה בתיבה האדומה בקצה. בנתיב התחתון (Path 2), מתחילים מ-$P(H) = 0.1$ -- הסתברות כאב הראש באוכלוסייה. אבל כאן יש שאלה: מה צריך להכפיל? $P(F|H)$ -- ההסתברות לשפעת בקרב הסובלים מכאב ראש. זה בדיוק מה שבייס בא לגלות. שני הנתיבים מובילים לאותה תוצאה סופית, והשוויון הזה הוא היסוד של כלל השרשרת.

עכשיו, בתוך המעגל הקטן הזה, אני מחפש את אלה שכואב להם הראש.

תובנה קריטית

הסתברות מותנית היא כמו שינוי נקודת המבט. במקום להסתכל על כל העולם, אני מתמקד רק בתת-קבוצה מסוימת.

2.5 הסיסמה הבינארית

הנה דוגמה שתחדד את הרעיון. נניח שיש לי סיסמה בינארית בת 4 ספרות: 1011.

כמה ניסיונות צריך כדי לפצח אותה? $2^4 = 16$ אפשרויות.

עכשיו, מישהו הציץ מהחלון וראה שהספרה הראשונה היא 1. זו **ראייה**.

מה קרה? מרחב האפשרויות ירד מ-16 ל-8. הסיכוי למצוא את הסיסמה -- **עלה** מאחד-ל-16 לאחד-ל-8.

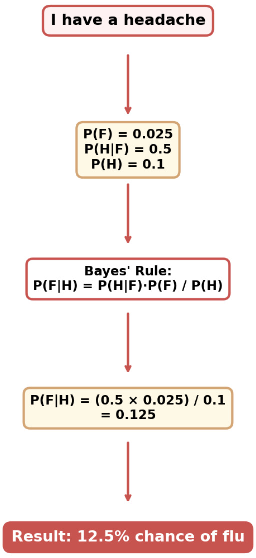
עיקרון יסוד

כל ראייה מקטינה את מרחב האפשרויות. ולכן -- מגדילה את הסיכוי לדעת את האמת.

2.6 נוסחת בייס

עכשיו הגענו לשאלה שטרדה את בייס: אני יודע ש-50% מחולי השפעת סובלים מכאב ראש. זה $P(\text{כאב ראש}|\text{שפעת}) = 0.5$. אבל מה שאני באמת צריך לדעת כרופא הוא ההפך: מגיע אלי חולה עם כאב ראש -- מה הסיכוי שיש לו שפעת?

זהו היפוך הכיוון -- מהראייה להשערה. איך עושים את זה? בואו נעקוב אחרי התהליך צעד אחר צעד. אנחנו מתחילים מהראייה -- החולה שלפנינו מתלונן על כאב ראש. יש לנו שלוש פיסות מידע: ההסתברות הבסיסית לשפעת באוכלוסייה, ההסתברות שחולה שפעת יפתח כאב ראש, וההסתברות הכללית לכאב ראש.



איור 2.3: זרימת החישוב

כפי שניתן לראות באיור 2.3, התהליך מתואר כזרימה ברורה מלמעלה למטה. בקופסה האדומה העליונה -- הראייה שלנו: "ehcadaeh a evah I". החץ הראשון מוביל אותנו לקופסה הבהירה שמכילה את כל המידע שאנחנו יודעים: $P(F) = 0.025$ (שיעור השפעת באוכלוסייה), $P(H|F) = 0.5$ (סיכוי לכאב ראש אצל חולה שפעת), ו-$P(H) = 0.1$ (שיעור כאב הראש הכללי). החץ השני מוביל לקופסה האדומה של משפט בייס עצמו: $P(F|H) = P(H|F) \cdot P(F)/P(H)$ -- הנוסחה שמהפכת את הכיוון. החץ השלישי מוביל לתיבה הבהירה שמציגה את החישוב המפורש: $(0.5 \times 0.025)/0.1 = 0.125$. והחץ הרביעי מגיע לתוצאה הסופית בקופסה האדומה: "5.21 :tluseR

משפט בייס

$$P(\text{כאב ראש}|\text{שפעת}) = \frac{P(\text{שפעת}|\text{כאב ראש}) \cdot P(\text{שפעת})}{P(\text{כאב ראש})} \tag{2.3}$$

נציב את המספרים:

$$P(\text{כאב ראש}|\text{שפעת}) = \frac{0.5 \times 0.025}{0.1} = 0.125 = 12.5\% \tag{2.4}$$

2.7 התוצאה המפתיעה

50% מחולי השפעת סובלים מכאב ראש -- אבל רק 12.5% מאלה שכואב להם הראש

חולים בשפעת! למה? כי שפעת היא מחלה נדירה יחסית. רוב האנשים שכואב להם הראש סובלים מסיבות אחרות לגמרי.

בואו נבין מה באמת קורה כאן. משפט בייס מורכב משלושה רכיבים מרכזיים, וכל אחד מהם תופס תפקיד אחר בחישוב. הרכיב הראשון הוא הפריור -- מה ידענו לפני שהראייה הגיעה. הרכיב השני הוא הנראות (Likelihood) -- עד כמה הראייה מתאימה להשערה שלנו. והרכיב השלישי הוא המכנה -- נורמליזציה שמוודאת שהתוצאה תישאר הסתברות חוקית.

איור 2.4: רכיבי הנוסחה

כפי שניתן לראות באיור 2.4, שלושת הרכיבים מוצגים כשלושה עיגולים הזורמים מימין לשמאל. העיגול הכחול הראשון מימין הוא ה-Prior -- $P(H)$ -- "מה ידענו לפני הראייה". זו ההסתברות הבסיסית להשערה שלנו, לפני שקיבלנו כל מידע חדש. החץ מוביל לעיגול הירוק באמצע -- Likelihood -- $P(E|H)$ -- "עד כמה הראייה מתאימה להשערה". זהו הגורם שמודד עד כמה סביר שנראה את הראייה אם ההשערה נכונה. והחץ השני מוביל לעיגול הוורוד השלישי -- Posterior -- $P(H|E)$ -- "מה אנחנו מאמינים אחרי הראייה". זו התוצאה המעודכנת שלנו, האמונה החדשה שלנו לאחר שספגנו את המידע. הזרימה הזו היא תמציתה של תורת בייס: התחלה מידע קודם, שילוב ראייה חדשה, והגעה לאמונה מעודכנת.

הלקח

מחלות נדירות נשארות נדירות, גם כשיש סימפטום.
שיעור הבסיס (Base Rate) משפיע עצום על התוצאה הסופית.

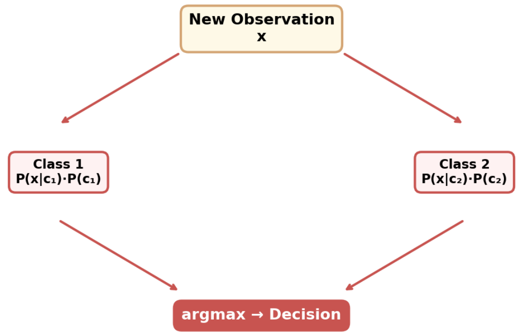
זה מסביר למה לא כדאי "לקפוץ למסקנות". גם אם מצאנו טביעות אצבע של מישהו בזירת רצח -- זה לא אומר בהכרח שהוא הרוצח.

פרק 3

המסווג בפעולה

3.1 כלל ההחלטה

עכשיו, כשיש לנו את נוסחת בייס, איך הופכים אותה למכונת סיווג? הרעיון פשוט: נחשב את ההסתברות עבור כל קבוצה אפשרית, ונבחר את זו עם ההסתברות הגבוהה ביותר. איור 3.1 ממחיש את תהליך ההחלטה הזה. כאשר מגיעה תצפית חדשה x, היא מסתעפת לשתי אפשרויות -- שני מסלולים מקבילים שבהם אנחנו מחשבים את מכפלת הנראות והפריור לכל מחלקה: $P(x|c_1) \cdot P(c_1)$ עבור מחלקה ראשונה, ו-$P(x|c_2) \cdot P(c_2)$ עבור מחלקה שנייה. בסוף המסלולים, שני הערכים נפגשים בפעולת ה-argmax, שבוחרת את המחלקה עם הציון הגבוה יותר ומחזירה את ההחלטה הסופית.



איור 3.1: תהליך הסיווג -- מתצפית חדשה להחלטה סופית דרך חישוב ציונים והשוואתם

כלל בייס לסיווג

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, P(Y = c|X = \mathbf{x}) \tag{3.1}$$

בחר את המחלקה c שמביאה למקסימום את ההסתברות.

3.2 הפטנט של בייס

שימו לב למשהו חשוב. בנוסחה המלאה:

$$P(Y = c|X = \mathbf{x}) = \frac{P(X = \mathbf{x}|Y = c) \cdot P(Y = c)}{P(X = \mathbf{x})} \tag{3.2}$$

המכנה -- $P(X = \mathbf{x})$ -- הוא זהה לכל הקבוצות! כי אנחנו בודקים את אותה דגימה $\mathbf{x}$ ביחס לקבוצות שונות.

לכן, אפשר להתעלם ממנו ולהשוות רק את המונים:

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, P(X = \mathbf{x}|Y = c) \cdot P(Y = c) \tag{3.3}$$

זה כמו שני רצים במרוץ. לא חשוב כמה מהר הם רצים -- חשוב רק מי מהיר יותר.

3.3 הבנקאי המתלבט

נניח שאתם עובדים בבנק ועליכם להחליט: האם לאשר הלוואה ללקוח חדש? יש לכם מידע היסטורי על 200 לקוחות קודמים. לכל אחד רמת הכנסה (נמוכה, בינונית, גבוהה) ואיכות אשראי (טובה או רעה).

איור 3.2 מציג את התפלגות איכות האשראי לפי רמות הכנסה. ציר ה-X מייצג את שלוש רמות ההכנסה -- נמוכה, בינונית, גבוהה. ציר ה-Y מציג את מספר הלקוחות בכל קטגוריה. העמודות הוורודות מייצגות לקוחות עם אשראי רע, והעמודות הירוקות מייצגות לקוחות עם אשראי טוב. ניתן לראות דפוס ברור: ברמת הכנסה נמוכה, רוב הלקוחות (24 מתוך 75) היו בעלי אשראי רע. ברמת הכנסה בינונית, התפלגות איכות האשראי מאוזנת יותר. ברמת הכנסה גבוהה, הרוב המוחלט (05 מתוך 86) היו בעלי אשראי טוב. תבנית זו חושפת קשר ברור בין רמת הכנסה לאיכות אשראי -- ככל שההכנסה גבוהה יותר, כך גדל השיעור של לקוחות עם אשראי טוב.

איור 3.2: התפלגות איכות האשראי לפי רמת הכנסה -- דפוס ברור של מתאם חיובי בין הכנסה לאיכות אשראי

טבלה 3.1: נתוני הלקוחות

הכנסה	**אשראי רע**	**אשראי טוב**	**סה"כ**
נמוכה	42	15	57
בינונית	35	40	75
גבוהה	18	50	68
סה"כ	95	105	200

3.4 שלב האימון

מה עושים עם הנתונים האלה? **סופרים**. קודם כל, ההסתברויות הפריוריות:

$$P(\text{רע}) = \frac{95}{200} = 0.475 \tag{3.4}$$

$$P(\text{טוב}) = \frac{105}{200} = 0.525 \tag{3.5}$$

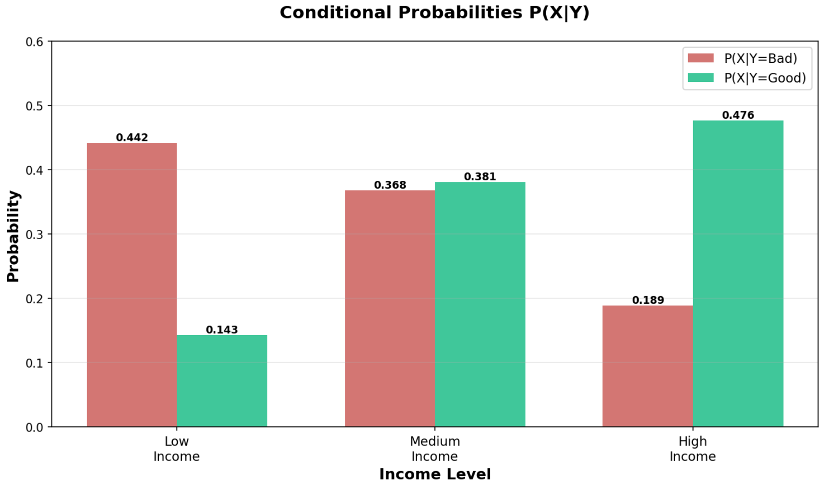
יותר לקוחות היו עם אשראי טוב מאשר רע. זה הפריור.

עכשיו, ההסתברויות המותנות. מתוך אלה עם אשראי רע, כמה היו בעלי הכנסה נמוכה?

$$P(\text{נמוכה}|\text{רע}) = \frac{42}{95} = 0.442 \tag{3.6}$$

איור 3.3 מציג השוואה ויזואלית של הסתברויות מותנות אלה. ציר ה-X מציג את שלוש רמות ההכנסה, וציר ה-Y מציג את הערך ההסתברותי -- מספר בין 0 ל-6.0 המייצג את השכיחות היחסית. העמודות הוורודות מייצגות $P(X|Y = \text{רע})$ --

ההסתברות לרמת הכנסה מסוימת בהינתן שהאשראי רע. העמודות הירוקות מייצגות $P(X|Y = \text{טוב})$ -- ההסתברות לרמת הכנסה בהינתן שהאשראי טוב. שימו לב להבדלים הדרמטיים: עבור הכנסה נמוכה, העמודה הוורודה גבוהה מאוד (244.0) בעוד הירוקה נמוכה (341.0). עבור הכנסה גבוהה, המצב הפוך -- העמודה הירוקה גבוהה (674.0) והוורודה נמוכה (981.0). זה אומר שהכנסה נמוכה היא אינדיקציה חזקה לאשראי רע, בעוד הכנסה גבוהה היא אינדיקציה חזקה לאשראי טוב.



איור 3.3: השוואת הסתברויות מותנות -- הכנסה נמוכה מצביעה על אשראי רע, הכנסה גבוהה על אשראי טוב

טבלה 3.2: טבלת הנראות

הכנסה	$P(X|\text{רע})$	$P(X|\text{טוב})$
נמוכה	0.442	0.143
בינונית	0.368	0.381
גבוהה	0.189	0.476

3.5 הלקוח החדש

עכשיו מגיע לקוח חדש עם הכנסה **נמוכה**. מה יהיה האשראי שלו? נחשב את הציון עבור כל קבוצה:

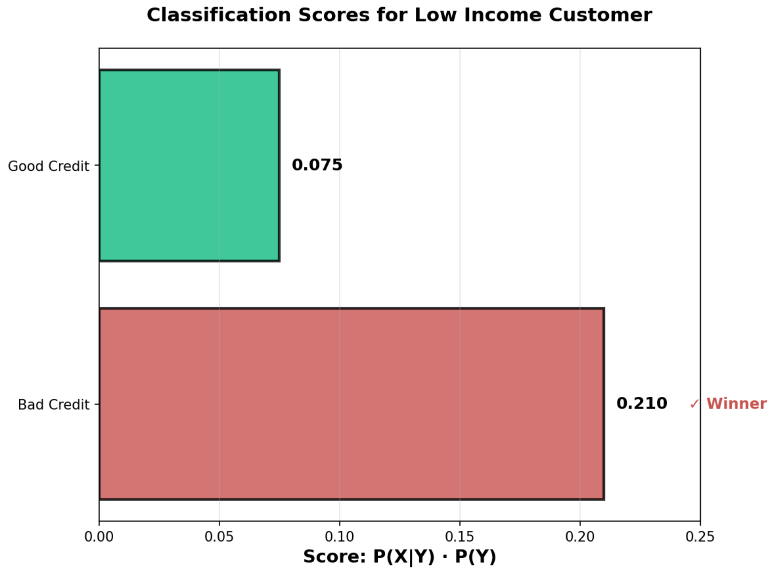
ציון לאשראי רע:

$$0.442 \times 0.475 = 0.210 \tag{3.7}$$

ציון לאשראי טוב:

$$0.143 \times 0.525 = 0.075 \tag{3.8}$$

איור 3.4 מדגים את השוואת הציונים הסופית. ציר ה-X מציג את הציון -- מכפלת הנראות והפריור -- ערך בין 0 ל-52.0. ציר ה-Y מציג את שתי המחלקות האפשריות: אשראי טוב ואשראי רע. העמודה הירוקה (אשראי טוב) קצרה ומגיעה ל-570.0, בעוד העמודה הוורודה (אשראי רע) ארוכה בהרבה ומגיעה ל-012.0. על העמודה הוורודה מופיע סימן V ומילה "Winner" באדום, המצביע על כך שהיא המחלקה הזוכה. הוויזואליזציה הזו הופכת את ההחלטה לברורה לחלוטין -- הציון הגבוה יותר מנצח, והלקוח מסווג כבעל אשראי רע.



איור 3.4: השוואת ציוני הסיווג ללקוח עם הכנסה נמוכה -- הציון הגבוה יותר מנצח

ההחלטה

$0.210 > 0.075$

הסיווג: אשראי רע.

המלצה: לא לאשר את ההלוואה.

3.6 למה זה עבד?

הסבר

למרות שבאוכלוסייה יש יותר אנשים עם אשראי טוב (52.5%), העובדה שהלקוח בעל הכנסה נמוכה היא ראייה חזקה.

44.2% מאלה עם אשראי רע היו בעלי הכנסה נמוכה, לעומת רק 14.3% מאלה עם אשראי טוב.

הראייה הזו מכריעה את הכף.

זה לא מפתיע אינטואיטיבית -- אבל היופי הוא שיש לנו עכשיו **נוסחה מתמטית** שעושה את זה אוטומטית, גם כשיש מאה תכונות ומיליון לקוחות.

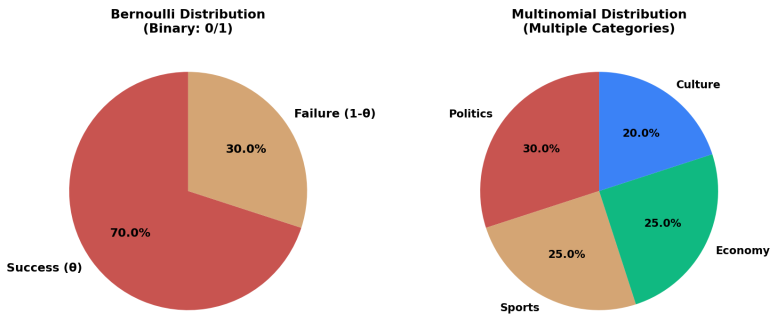
פרק 4

עולם של מספרים

4.1 בדיד ורציף

עד עכשיו עסקנו בעולם פשוט: יש או אין. נמוכה, בינונית, גבוהה. דואר זבל או דואר רגיל. כל מה שסיווגנו עד כה היה מוגדר על ידי קטגוריות ברורות -- מעין מדורי נפרדים שאי אפשר לעבור ביניהם ללא קפיצה. אבל מה קורה כשהמאפיינים שלנו אינם קטגוריות אלא מספרים רציפים? מה נעשה כשהמציאות שלפנינו אינה מדורים נפרדים, אלא קשת רצופה של ערכים -- טמפרטורה, גובה, משקל, מהירות, זמן?

בואו נסתכל על איור 4.1. הגרף משמאל מדגים את העולם הבינארי הפשוט שלנו -- התפלגות ברנולי, שבה יש רק שני אירועים אפשריים: הצלחה או כישלון, 0 או 1. במקרה זה אנחנו רואים שההסתברות להצלחה היא 70%, ולכישלון -- 30%. זהו עולם של ודאויות יחסיות: כן או לא, שחור או לבן. הגרף מימין מציג את העולם הרב-קטגוריאלי -- התפלגות מולטינומיאלית, שבה יש לנו יותר משתי אפשרויות. כאן אנחנו רואים ארבע קטגוריות: פוליטיקה (30%), ספורט (25%), כלכלה (25%), ותרבות (20%). גם זה עדיין עולם בדיד -- מספר סופי של אפשרויות נפרדות. אבל מה קורה כשעולם האפשרויות הוא אינסופי?



איור 4.1: משתנים קטגוריאליים: משמאל -- התפלגות ברנולי (בינארי: הצלחה או כישלון), מימין -- התפלגות מולטינומיאלית (ארבע קטגוריות נפרדות)

4.2 הדואר הזבל

לפני שנגיע לעולם הרציף, בואו נבין לעומק את העולם הבדיד. נניח שאני רוצה לבנות מכונה שמזהה דואר זבל -- מסווג שיודע להבחין בין אימייל לגיטימי לבין הודעות ספאם שמציפות את תיבת הדואר שלנו. איך בונים את מאגר הנתונים? איך לוכדים את מהות הבעיה במספרים?

קודם כל, צריך לחשוב על מילים חשודות. מילים שמופיעות לעתים קרובות בהודעות ספאם אבל נדירות בדואר רגיל: "זכייה", "נסיך", "תרוויח", "לפרטים", "כנס". אלה הן מילות המפתח שמסגירות את הספאם. עכשיו עוברים על אימיילים ישנים -- מאות אימיילים שכבר סיווגנו ידנית. לכל אימייל בודקים: האם המילה "זכייה" מופיעה בו? אם כן, מסמנים 1. אם לא -- מסמנים 0. וכך לכל מילה ומילה. כך אנחנו הופכים טקסט חופשי -- שפה טבעית מורכבת -- למטריצה נקייה של אפסים ואחדים.

טבלה 4.1: מבנה הדאטה לזיהוי דואר זבל -- כל שורה מייצגת אימייל, כל עמודה מייצגת מילת מפתח, והערכים הם 0 (אינה מופיעה) או 1 (מופיעה)

זבל?	כנס	לפרטים	תרוויח	נסיך	זכייה
כן	0	1	1	0	1
כן	1	1	0	1	0
כן	0	0	0	0	1
לא	0	0	0	0	0

4.3 התפלגות ברנולי

למשתנה בינארי -- כן או לא, 0 או 1 -- יש התפלגות פשוטה מאוד. זוהי התפלגות ברנולי, על שם המתמטיקאי יעקב ברנולי שחי במאה ה-17. ברנולי הבין שכל ניסוי שיש לו רק שתי תוצאות אפשריות -- מטבע שנופל על עץ או פלי, מנורה שדולקת או כבויה, חולה שמחלים או לא -- כולם מצייתים לאותו חוק מתמטי פשוט:

ברנולי

$$P(X = x|\theta) = \theta^x (1-\theta)^{1-x} \tag{4.1}$$

כאשר θ היא ההסתברות להצלחה, ו-x יכול להיות רק 0 או 1.

איך מחשבים את θ? פשוט סופרים. אנחנו עוברים על כל הנתונים, סופרים כמה פעמים קרתה "הצלחה" -- כלומר, כמה פעמים ראינו $x = 1$ -- ומחלקים במספר הכולל של הניסיונות:

$$\hat{\theta} = \frac{\text{מספר ההצלחות}}{\text{סך הניסיונות}} \tag{4.2}$$

דוגמה 1.4. נניח שעברנו על מאגר של 100 אימיילים שכולם מסווגים כספאם. בדקנו כמה מהם מכילים את המילה "זכייה", וגילינו שב-70 מהם המילה הזאת אכן מופיעה.

אז ההסתברות שמילה זו תופיע באימייל ספאם היא:

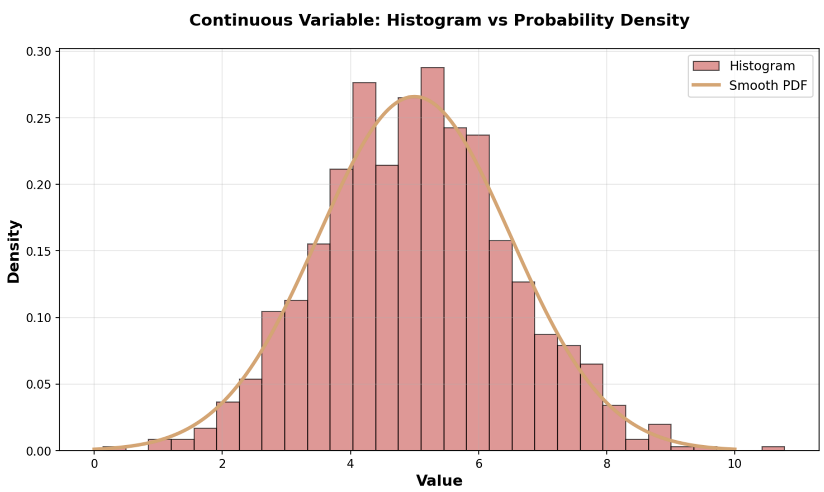
$$\hat{\theta}_{\text{זכייה}} = \frac{70}{100} = 0.7 \tag{4.3}$$

כלומר, 70% מהודעות הספאם מכילות את המילה "זכייה". זהו המידע שנשתמש בו כדי לסווג אימיילים חדשים.

4.4 כשהמספרים רציפים

עד כה דיברנו על ערכים בדידים -- דברים שאפשר למנות, לספור, לסווג למדורים נפרדים. אבל מה עושים כשהפיצ'ר שלנו הוא, למשל, אורך עלה של פרח? המספר יכול להיות 2.3 ס"מ או 2.31 ס"מ או 2.314 ס"מ -- יש כאן אינסוף אפשרויות על ציר המספרים הממשיים. איך מתארים התפלגות של משהו כזה?

בואו נסתכל על איור 4.2. הגרף מציג משתנה רציף שערכיו פרושים על ציר ה-X מ-0 עד 10. ציר ה-Y מייצג את הצפיפות -- כלומר, עד כמה "תפוסה" כל נקודה בערכים. העמודות הוורודות מייצגות היסטוגרמה -- אנחנו מחלקים את הציר לתאים (למשל, מ-0 עד 0.5, מ-0.5 עד 1, וכן הלאה), וסופרים כמה תצפיות נפלו בכל תא. ככל שהעמודה גבוהה יותר, כך יותר תצפיות נמצאו בטווח הזה. הקו הכתום המלופף מייצג את פונקציית צפיפות ההסתברות (PDF – Probability Density Function) -- עקומה חלקה שמקרבת את ההיסטוגרמה הדיסקרטית. שימו לב שהעקומה מגיעה לשיא סביב הערך 5 -- זהו הממוצע, המקום שבו מתרכזים רוב הערכים.



איור 4.2: מהיסטוגרמה לפונקציית צפיפות הסתברות: ציר ה-X מייצג את ערכי המשתנה הרציף, ציר ה-Y מייצג את הצפיפות (תכיפות יחסית). העמודות הוורודות -- היסטוגרמה בדידה; הקו הכתום -- עקומת הצפיפות המתמטית (Smooth PDF)

אז יש לנו שני פתרונות אפשריים למשתנים רציפים:

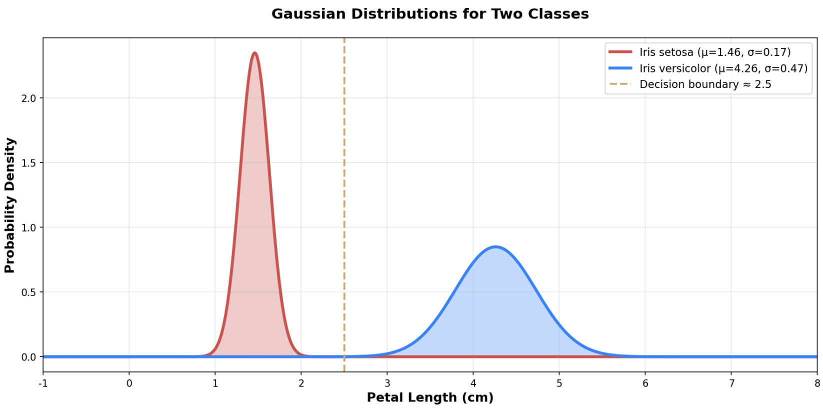
פתרון א': היסטוגרמה -- מחלקים את הטווח לתאים קטנים, וסופרים כמה תצפיות נפלו בכל תא. זהו פתרון מעשי אבל גס: הוא תלוי במספר התאים שבחרנו ובגבולות שלהם. אם נבחר תאים רחבים מדי, נאבד פרטים חשובים; אם נבחר תאים צרים מדי, יהיו לנו תאים ריקים.

פתרון ב': התפלגות נורמלית -- מניחים שהמשתנה מתפלג לפי עקומת פעמון, ההתפלגות הגאוסית הנפלאה. זהו פתרון אלגנטי ועמוק, שמבוסס על אחת התופעות היפות ביותר במתמטיקה -- משפט הגבול המרכזי.

4.5 עקומת הפעמון

למה דווקא גאוסיאן? למה ההתפלגות הנורמלית מופיעה שוב ושוב בטבע, בחברה, במדע? התשובה טמונה במשפט הגבול המרכזי: כל דבר שמושפע מהרבה גורמים קטנים ובלתי תלויים -- גובה אדם (מושפע מאלפי גנים), טמפרטורה יומית (מושפעת ממאות תהליכים אטמוספריים), אורך עלה של פרח (מושפע משילוב של גנטיקה, מזג אויר, אדמה, מים) -- יתפלג בערך כמו פעמון. זהו חוק אוניברסלי שמאחד תחתיו תופעות שונות לכאורה.

בואו נסתכל על איור 4.3. הגרף מציג שתי עקומות פעמון שונות על אותו ציר. ציר ה-X מייצג את אורך עלי הכותרת בס"מ, וציר ה-Y מייצג את צפיפות ההסתברות. העקומה האדומה -- Iris setosa -- מרוכזת סביב $\mu = 1.46$ ס"מ עם סטיית תקן $\sigma = 0.17$ ס"מ. זוהי עקומה גבוהה וצרה: העלים של זן זה קצרים מאוד, והשונות ביניהם קטנה -- כמעט כל העלים באורך דומה. העקומה הכחולה -- Iris versicolor -- מרוכזת סביב $\mu = 4.26$ ס"מ עם סטיית תקן $\sigma = 0.47$ ס"מ. זוהי עקומה נמוכה ורחבה יותר: העלים של זן זה ארוכים הרבה יותר, והשונות ביניהם גדולה יותר. הקו המקווקו החום מסמן את גבול ההחלטה -- הנקודה $x \approx 2.5$ ס"מ שבה אנחנו עוברים מלהעדיף זן אחד לזן השני.



איור 4.3: שתי התפלגויות נורמליות של שני זני פרחי אירוס: העקומה האדומה -- Iris setosa עם ממוצע $\mu = 1.46$ וסטיית תקן $\sigma = 0.17$ (צרה וגבוהה); העקומה הכחולה -- Iris versicolor עם ממוצע $\mu = 4.26$ וסטיית תקן $\sigma = 0.47$ (רחבה ונמוכה). הקו המקווקו מסמן את גבול ההחלטה

ההתפלגות הנורמלית

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.4}$$

יש לה שני פרמטרים בלבד:

- μ -- הממוצע (איפה מרכז הפעמון). זוהי הנקודה שבה העקומה מגיעה לשיא.
- σ^2 -- השונות (כמה רחב הפעמון). ככל ש-σ גדול יותר, העקומה רחבה ונמוכה יותר; ככל ש-σ קטן יותר, העקומה צרה וגבוהה יותר.

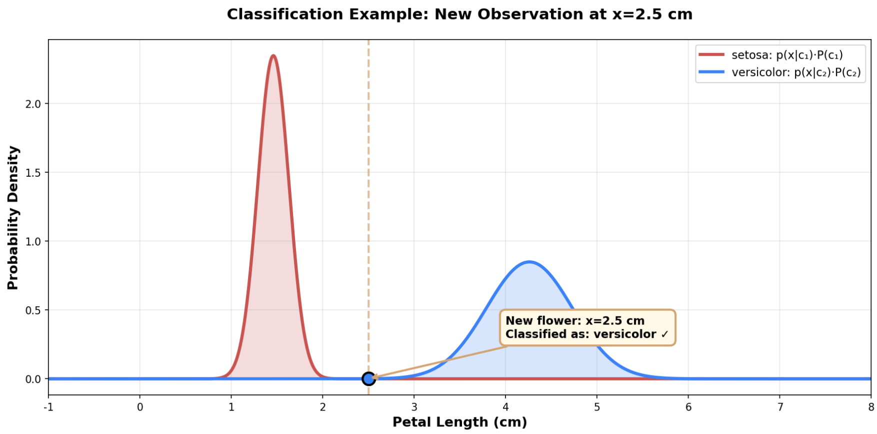
4.6 פרחי האירוס

יש מאגר נתונים מפורסם של פרחי אירוס שנאסף על ידי הבוטנאי אדגר אנדרסון ב-1936. לכל פרח נמדדו ארבע תכונות: אורך ורוחב גביע הפרח, ואורך ורוחב עלי הכותרת. יש שם שלושה זנים שונים של אירוס: Iris setosa, Iris versicolor, ו-Iris virginica. מאגר זה הפך לאבן בוחן קלאסית לאלגוריתמי סיווג -- כמעט כל סטודנט למדעי הנתונים מתנסה בו בשלב כלשהו.

בואו נפשט את הבעיה ונתמקד בתכונה אחת: אורך עלי הכותרת (Petal Length). ובשני זנים בלבד: setosa ו-versicolor.

איור 4.4 מציג את אותה בעיה שראינו קודם, אבל עכשיו עם דוגמה קונקרטית. ציר ה-X מייצג את אורך עלי הכותרת בס"מ. ציר ה-Y מייצג את צפיפות ההסתברות. העקומה האדומה -- setosa -- מרוכזת סביב 1.46 ס"מ, והעקומה הכחולה -- versicolor

-- מרוכזת סביב 4.26 ס"מ. הנקודה השחורה המסומנת על הציר מציינת פרח חדש שאורך עליו $x = 2.5$ ס"מ. התיבה הכתומה מכריזה על התשובה: "הפרח החדש: $x = 2.5$ ס"מ. מסווג כ-versicolor" (עם סימן V ירוק).



איור 4.4: סיווג פרחי אירוס: ציר ה-X -- אורך עלי הכותרת בס"מ; ציר ה-Y -- צפיפות הסתברות. העקומה האדומה -- Iris setosa; העקומה הכחולה -- Iris versicolor. הנקודה השחורה מסמנת פרח חדש באורך 2.5 ס"מ, המסווג כ-versicolor

הנתונים:

- Iris setosa: $\mu_1 = 1.46$ ס"מ, $\sigma_1 = 0.17$ ס"מ
- Iris versicolor: $\mu_2 = 4.26$ ס"מ, $\sigma_2 = 0.47$ ס"מ

4.7 הפרח המסתורי

מגיע פרח חדש שאורך עלי הכותרת שלו 2.5 ס"מ. לאיזה זן הוא שייך? זהו השאלה שמסווג בייס נועד לענות עליה. נניח שההסתברות הפריורית לכל זן היא 0.5 -- כלומר, אין לנו העדפה מוקדמת. עכשיו נחשב את הציון לכל זן באמצעות הנוסחה הגאוסית:

ציון ל-Setosa:

$$\mathcal{N}(2.5|1.46, 0.17^2) \times 0.5 \approx 0.0001 \times 0.5 = 0.00005 \tag{4.5}$$

ציון ל-Versicolor:

$$\mathcal{N}(2.5|4.26, 0.47^2) \times 0.5 \approx 0.21 \times 0.5 = 0.105 \tag{4.6}$$

ההחלטה

$0.105 \gg 0.00005$
הפרח הוא **Iris versicolor**.

למה?

הערך 2.5 ס"מ רחוק כ-6 סטיות תקן מהממוצע של setosa -- מספר אסטרונומי.
בשפה מתמטית:

$$\frac{2.5 - 1.46}{0.17} \approx 6.1 \tag{4.7}$$

זהו אירוע כל כך נדיר שההסתברות שלו כמעט אפסית.
לעומת זאת, הערך 2.5 ס"מ רחוק רק כ-3.7 סטיות תקן מהממוצע של versicolor:

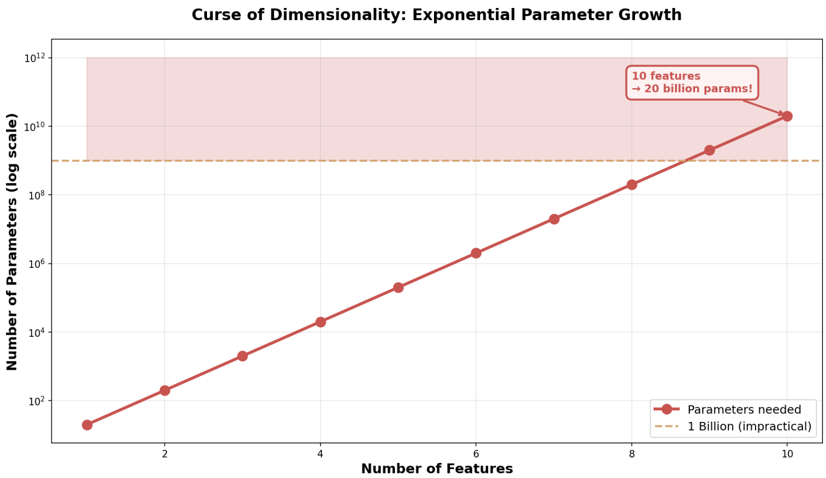
$$\frac{4.26 - 2.5}{0.47} \approx 3.7 \tag{4.8}$$

זה עדיין לא קרוב לממוצע -- זהו ערך בקצה השמאלי של ההתפלגות -- אבל זה הרבה יותר סביר מאשר ב-setosa.
ההבדל הזה מכריע. מסווג בייס אינו שואל "האם זה קרוב לממוצע?" אלא "מה יותר סביר?" -- והתשובה ברורה.

פרק 5

ההנחה הנאיבית

5.1 קללת הממדים



איור 5.1: גידול אקספוננציאלי במספר הפרמטרים

יש בעיה. בעיה גדולה. בעיה שמאיימת לקרוס את כל מה שבנינו עד כה.

עד עכשיו התעסקנו עם תכונה אחת או שתיים. אבל במציאות? במציאות יש עשרות, מאות, לפעמים אלפי תכונות. כל תכונה מוסיפה ממד נוסף למרחב הנתונים שלנו. וכאן מתחילה הקללה.

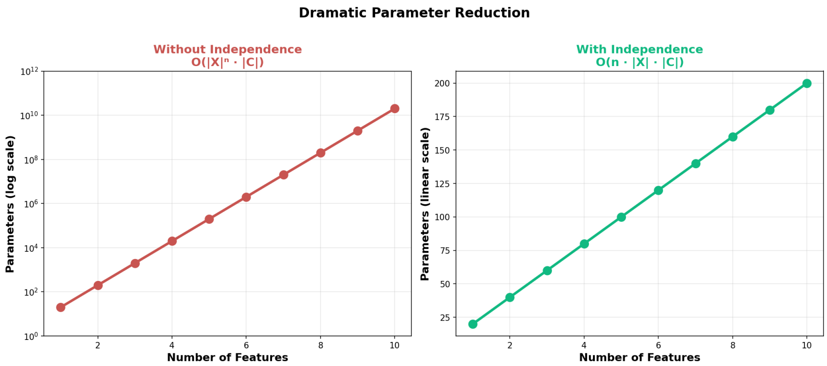
איור 5.1 מציג את הבעיה בצורה דרמטית. הגרף מראה כיצד מספר הפרמטרים הנדרשים גדל באופן אקספוננציאלי עם מספר התכונות. בציר האופקי רואים את מספר התכונות, בציר האנכי -- בסקלה לוגריתמית -- את מספר הפרמטרים שצריך לאמוד. העקומה האדומה טומנת בחובה סיפור מפחיד: תכונה אחת דורשת מספר קטן של פרמטרים, שתי תכונות -- כבר מאות, שלוש -- אלפים. כשמגיעים לעשר תכונות, הנקודה האדומה הימנית מסגירה את האמת הקשה: עשרים מיליארד פרמטרים.

זו לא סתם דוגמה תיאורטית. נניח שלכל תכונה יש 10 ערכים אפשריים, ויש לנו 10 תכונות. כמה קומבינציות ייחודיות של תכונות קיימות?

$$10^{10} = 10,000,000,000 \tag{5.1}$$

עשרה מיליארד אפשרויות. לכל אחת צריך לאמוד הסתברות. אין מספיק נתונים בעולם כדי לכסות את כולן בצורה מהימנה. זו קללת הממדים -- ככל שמוסיפים ממדים, המרחב הופך לדליל יותר ויותר, והנתונים שלנו הופכים לנדירים יותר ויותר.

5.2 הפתרון המפתיע



איור 5.2: הפחתה דרמטית במספר הפרמטרים באמצעות הנחת האי-תלות

איור 5.2 מציג את הפתרון המבריק. שני גרפים זה לצד זה מספרים סיפור של שינוי פרדיגמה. הגרף השמאלי, באדום, חוזר על הבעיה: גידול אקספוננציאלי, סקלה לוגריתמית שמגיעה ל-10^{12} פרמטרים. אבל הגרף הימני, בירוק, מציג מהפכה: באותו מספר תכונות, רק 200 פרמטרים. לא מיליארדים -- מאתיים.

כיצד זה אפשרי? הפתרון הוא הנחה פשוטה להפליא, שלכאורה לא אמורה לעבוד: נניח שהתכונות **בלתי תלויות זו בזו.**

זאת אומרת, אם יש קשר בין לחץ דם למשקל -- נתעלם ממנו. אם יש קורלציה בין הכנסה להשכלה -- נתעלם גם ממנה. נתייחס לכל תכונה בנפרד, כאילו היא חיה בעולם משלה.

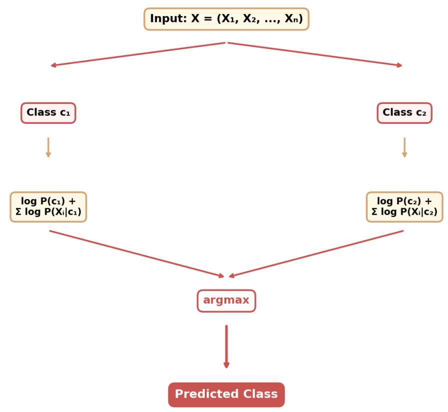
ההשוואה החזותית בין שני הגרפים בלתי נשכחת: מהסקלה הלוגריתמית האדומה שמטפסת לגבהים בלתי אפשריים, לסקלה הליניארית הירוקה שנשארת מעשית ומנוהלת. מעשרים מיליארד פרמטרים למאתיים -- הפחתה של שמונה סדרי גודל.

הנחת נאיב בייס

$$P(X_1, X_2, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \tag{5.2}$$

ההסתברות המשותפת של כל התכונות נתון הסיווג, שווה למכפלת ההסתברויות הבודדות של כל תכונה.

5.3 למה "נאיבי"?



איור 5.3: תרשים זרימת המסווג הנאיבי

איור 5.3 מציג את תהליך הסיווג הנאיבי בצורה ויזואלית. בראש הדיאגרמה -- הקלט: וקטור תכונות $X = (X_1, X_2, \ldots, X_n)$. משם מסתעפות שתי חיצים אדומות לשתי קבוצות אפשריות: c_1 ו-c_2. לכל קבוצה מחושב ציון: $\log P(c_i) + \sum \log P(X_i|c_i)$ -- הלוגריתם של הסתברות הקבוצה המוקדמת, בתוספת סכום הלוגריתמים של כל התכונות. שתי החיצים מתכנסות למטה לפעולת argmax, שבוחרת את הקבוצה בעלת הציון הגבוה ביותר. התוצאה -- הסיווג החזוי.

ההנחה הזו נאיבית -- כי היא כמעט תמיד לא נכונה. ברור שיש קשר בין תכונות! בין גובה למשקל יש מתאם חזק. בין הכנסה להשכלה -- גם כן. בין גיל לבריאות -- בוודאי. כל חוקר סטטיסטיקה יודע שהעולם מלא בתלויות, בקורלציות, בקשרים מורכבים בין משתנים.

אבל הנה הפלא: למרות שההנחה שגויה, האלגוריתם **עובד**. ועובד טוב. מאוד טוב.

בשנת 1998, חוקרי Microsoft הראו שמסווג נאיב בייס פשוט, ללא כל פיתוחים מתוחכמים, מצליח לזהות דואר זבל בדיוק של מעל 95%. זה לא מזל. זה לא חריג סטטיסטי. זו תופעה שחוזרת על עצמה שוב ושוב במגוון רחב של יישומים.

למה זה עובד?

לסיווג נכון לא צריך לדעת את ההסתברות **המדויקת** -- מספיק לדעת איזו קבוצה יותר סבירה.
אפילו אם ההסתברויות מעוותות, כל עוד **הסדר היחסי** בין הקבוצות נשמר -- הסיווג יצליח. נאיב בייס לא טוען לדייק מושלם בהסתברויות, הוא רק צריך לדרג נכון.

5.4 המהירות

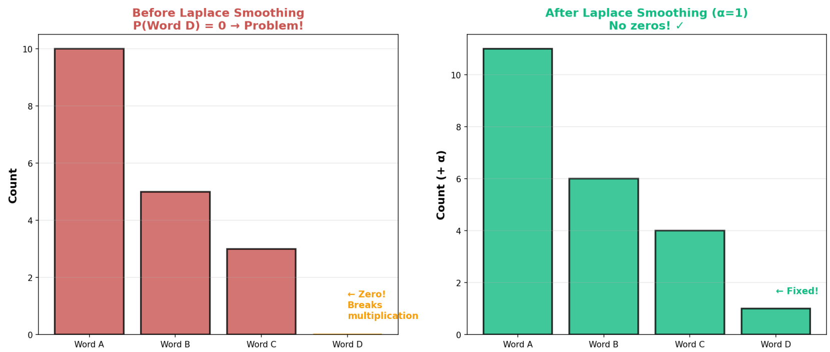
איור 5.4: השוואת מהירות אימון יחסית בין אלגוריתמים שונים

איור 5.4 חושף את אחד היתרונות הגדולים ביותר של נאיב בייס: המהירות. גרף עמודות אופקי משווה בין חמישה אלגוריתמי למידה פופולריים. בתחתית, בעמודה ירוקה זעירה, נאיב בייס זוכה לתואר "**המהיר ביותר!**" -- סימן הקריאה הוא חלק מהוויזואליזציה. מעליו, Logistic Regression -- פי 2.5 יותר איטי. SVM -- פי 8 יותר איטי. Random Forest -- פי 15 יותר איטי. ובקצה, Neural Network -- פי 30 יותר איטי, בעמודה ארוכה אדומה שמשתרעת על פני כמעט כל הגרף.

היתרון הזה לא מקרי. במקום לחשב 10^{10} קומבינציות של תכונות, נאיב בייס צריך לחשב רק $10 \times 10 = 100$ הסתברויות נפרדות. האימון הוא פשוט ספירה: לכל תכונה, לכל קבוצה -- סופרים כמה פעמים כל ערך הופיע, ובונים היסטוגרמה. אין אופטימיזציה מורכבת, אין חישובים איטרטיביים, אין גרדיאנטים.

זה הופך את נאיב בייס לאידיאלי למערכות בזמן אמת, לנתונים זורמים, ולמקרים שבהם צריך לאמן מחדש את המודל בתדירות גבוהה.

5.5 בעיית האפס



איור 5.5: כשההסתברות היא אפס -- בעיה קטסטרופלית

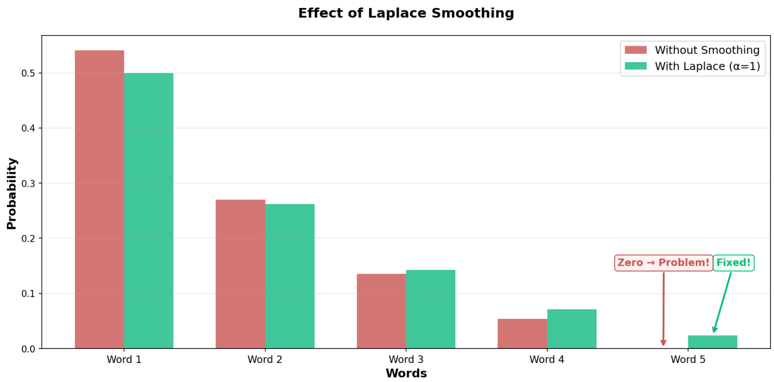
איור 5.5 מציג בעיה טכנית שיכולה לקרוס את כל המודל. שני היסטוגרמות זו לצד זו: משמאל, בורדו-אדום, המצב הבעייתי לפני תיקון. ארבע מילים (Word A, B, C, D) עם ספירות שונות: 10, 5, 3, ואז -- Word D עם ספירה של אפס. עמודה שלא קיימת. חץ כתום מצביע על הבעיה: "אפס! שובר את המכפלה."

מה הבעיה? במסווג נאיב בייס אנחנו מכפילים הסתברויות. אם תכונה מסוימת מעולם לא הופיעה בקבוצה מסוימת, ההסתברות שלה היא אפס. למשל: בכל האימיילים הלגיטימיים שראינו במהלך האימון, אף אחד לא הכיל את המילה "נסיך". לכן:

$$P(\text{נסיך}|\text{לגיטימי}) = 0$$

עכשיו מגיע אימייל חדש עם המילה "נסיך" ועם עשרות מילים אחרות. כשנחשב את ההסתברות של הקבוצה "לגיטימי", נכפיל את כל ההסתברויות של כל המילים. ואז נגיע למילה "נסיך" -- ונכפול באפס. התוצאה? אפס מוחלט. כל המידע מכל התכונות האחרות -- אבד. הסתברות אפס גורמת לכך שכל הקבוצה נדחית באופן מיידי, ללא קשר לכל השאר.

5.6 החלקת לפלס



איור 5.6: החלקת לפלס -- הפתרון לבעיית האפס

איור 5.6 מציג את הפתרון האלגנטי. שתי היסטוגרמות זו לצד זו: משמאל, באדום-ורוד, המצב לפני החלקה -- "בעיה!" כתוב למעלה. חמש מילים, כאשר Word 5 היא אפס מוחלט. חץ אדום מצביע מטה אל הבעיה. מימין, בטורקיז-ירוק, המצב אחרי החלקת לפלס ($\alpha = 1$) -- "אין אפסים!" כתוב בשמחה. אותן חמש מילים, אבל עכשיו גם Word 5 קיבלה עמודה קטנה -- הסתברות קטנה אך קיימת. חץ ירוק מצביע: "תוקן!"

ההשוואה החזותית מדהימה: לא רק ש-Word 5 קיבלה ערך חיובי, אלא גם שאר העמודות השתנו מעט -- הן ירדו בגובהן. זהו האפקט של החלקה: אנחנו "משאילים" קצת הסתברות מהמילים השכיחות ונותנים למילים הנדירות.

הפתרון פשוט להפליא: מוסיפים 1 (או כללי α) לכל הספירות. כך אף ספירה לא תישאר אפס.

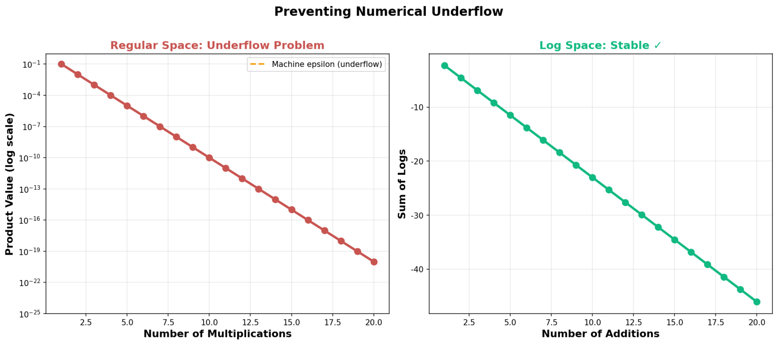
החלקת לפלס

$$P(X_i = x|Y = c) = \frac{\text{count}(X_i = x, Y = c) + \alpha}{\text{count}(Y = c) + \alpha|V|} \tag{5.3}$$

כאשר $|V|$ הוא מספר הערכים האפשריים לתכונה X_i, ו-α הוא פרמטר ההחלקה (בדרך כלל $\alpha = 1$).

במקום להתחיל מאפס, מתחילים מאחד. ככה אף הסתברות לא תהיה אפס לעולם. זה לא סתם טריק טכני -- זו למעשה דרך להביע ידע מוקדם: "גם אם לא ראיתי משהו בעבר, אני מניח שהוא אפשרי."

5.7 בעיית ה-Underflow



איור 5.7: בעיית Underflow והפתרון באמצעות לוגריתמים

איור 5.7 חושף בעיה נומרית עמוקה יותר. שני גרפים זה לצד זה מספרים סיפור של יציבות מספרית. הגרף השמאלי, באדום, מציג את הבעיה: "מרחב רגיל -- בעיית Underflow". הציר האנכי בסקלה לוגריתמית מתחיל ב-10^{-1} ויורד עד 10^{-25}. העקומה האדומה יורדת בתלילות. קו מקווקו אופקי מסמן את "סף המכונה (underflow)" -- הנקודה שבה המחשב כבר לא יכול לייצג מספרים כה קטנים. אחרי 15-20 כפולות, העקומה חוצה את הסף ונעלמת לתוך אזור הבלתי-ייצוגי.

הגרף הימני, בירוק, מציג את הפתרון: "מרחב לוגריתם -- יציב". הציר האנכי עכשיו הוא "סכום לוגריתמים", והוא פשוט יורד באופן ליניארי מ-0 ל--50 ואפילו יותר. אין סף, אין קריסה -- רק ירידה מסודרת ויציבה. החץ הירוק מצביע: "תוקן!"

המשמעות: כשיש הרבה תכונות, המכפלה של הסתברויות קטנות ($0.1 \times 0.05 \times 0.02 \times \ldots$) הופכת למספר זעיר כל כך שהמחשב לא יכול לייצג אותו. זהו Underflow -- זרימה מתחת לרזולוציה של המספרים הצפים. התוצאה היא אפס מספרי, אפילו אם מתמטית הערך לא אפס.

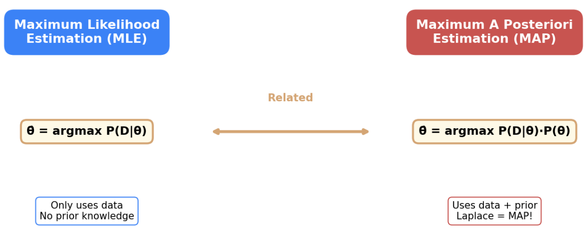
הפתרון: לעבור ללוגריתם. במקום לכפול הסתברויות, מחברים את הלוגריתמים שלהן. $\log(a \cdot b) = \log(a) + \log(b)$.

Log-Naive Bayes

$$\hat{y} = \underset{c}{\operatorname{argmax}} \left[\log P(Y=c) + \sum_{i=1}^{n} \log P(X_i|Y=c) \right] \tag{5.4}$$

מכפלה הופכת לסכום. סכום של לוגריתמים יציב הרבה יותר -- ניתן לחבר אלפי לוגריתמים מבלי לחצות את גבולות הייצוג המספרי.

5.8 MLE לעומת MAP



איור 5.8: השוואה בין שתי גישות לאמידת פרמטרים

איור 5.8 מציג דיאגרמה מושגית של שתי פילוסופיות שונות לאמידת פרמטרים. בראש, שני תיבות: משמאל, בכחול, Maximum Likelihood Estimation (MLE) -- "אמידת נראות מקסימלית". מימין, באדום, Maximum A Posteriori Estimation (MAP) -- "אמידה פוסטריורית מקסימלית". מתחת לכל אחת, נוסחה מתמטית שמסבירה את העיקרון. במרכז, חץ דו-כיווני עם הכיתוב "Related" -- קשורות זו לזו.

תחת MLE: $\theta = \text{argmax}\, P(D|\theta)$ -- "משתמש רק בנתונים, אין ידע מוקדם". תיבה כחולה מבהירה: האמידה מבוססת רק על הנתונים שראינו.

תחת MAP: $\theta = \text{argmax}\, P(D|\theta) \cdot P(\theta)$ -- "משתמש בנתונים ובידע מוקדם". תיבה אדומה מבהירה: "החלקת לפלס = MAP!" -- החלקת לפלס היא בעצם יישום של גישת MAP עם פריור אחיד.

יש שתי גישות פילוסופיות לאמידת פרמטרים:

MLE -- noitamitsE doohilekiL mumixaM

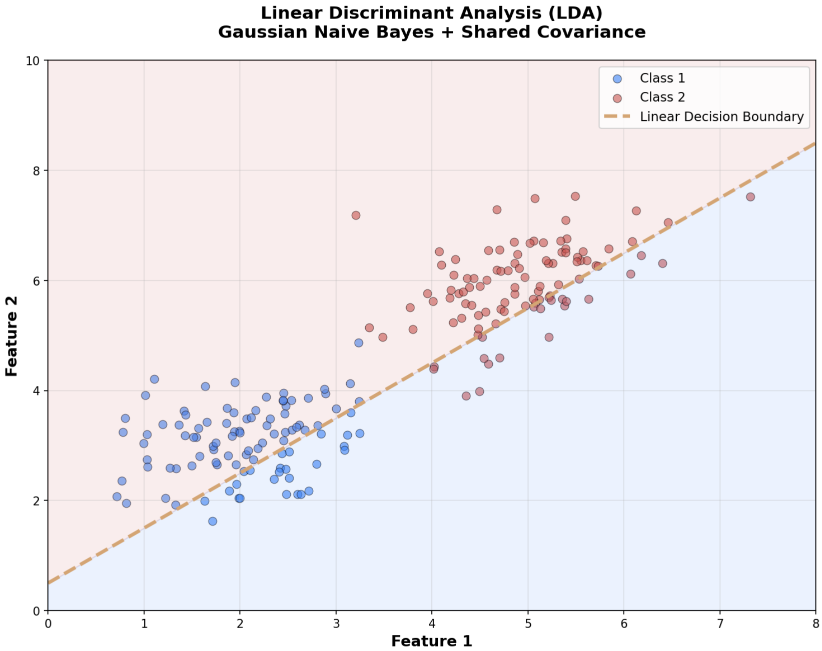
מחפש את הפרמטרים שממקסמים את ההסתברות של הנתונים שראינו. אם ראינו 7 ראשים ב-10 הטלות מטבע, נאמוד ש-$P(\text{ראש}) = 0.7$. זה הגיוני -- הפרמטר שהכי מסביר את הנתונים.

MAP -- iroiretsoP A mumixaM

משלב גם ידע מוקדם על הפרמטרים. אם אנחנו יודעים שהמטבע סימטרי, לא נקפוץ מיד למסקנה ש-$P(\text{ראש}) = 0.7$ אחרי 10 הטלות בלבד. נשלב את הידע המוקדם (Prior) שמטבעות נוטים להיות הוגנים.

החלקת לפלס שראינו קודם היא למעשה MAP עם פריור אחיד -- אנחנו מניחים שכל הערכים שווים בהסתברות מראש, ומוסיפים "ספירה וירטואלית" לכל אחד.

5.9 LDA -- האח המתוחכם



איור 5.9: גבול ההחלטה הליניארי של LDA

איור 5.9 מציג ויזואליזציה מרהיבה של שתי קבוצות במרחב דו-ממדי. הכותרת מעל הגרף מבהירה: "Linear Discriminant Analysis (LDA) -- Gaussian Naive Bayes + Shared Covariance". בציר האופקי: Feature 1, בציר האנכי: Feature 2. המרחב מחולק לשני אזורים: אזור כחול בהיר משמאל-למטה, ואזור אדום-ורוד מימין-למעלה. הגבול ביניהם -- קו מקווקו כתום -- הוא גבול ההחלטה הליניארי.

על הגרף מפוזרות נקודות: נקודות כחולות (Class 1) מתרכזות באזור השמאלי-תחתון, נקודות אדומות (Class 2) באזור הימני-עליון. הקו המקווקו עובר בדיוק באזור הביניים, מפריד בין שתי הקבוצות. זהו גבול ההחלטה -- כל נקודה משמאל לקו תסווג ככחולה, כל נקודה מימין -- כאדומה.

נאיב בייס מניח אי-תלות מוחלטת בין התכונות. במילים אחרות, הוא מתייחס לכל ממד כאילו הוא בלתי תלוי לחלוטין בממדים האחרים. זו ההנחה הנאיבית -- ולפעמים היא פשוט רחוקה מדי מהמציאות.

Linear Discriminant Analysis מרפה את ההנחה הזו. במקום להתעלם מהקשרים בין תכונות, LDA משתמש במטריצת קווריאנס **משותפת** -- מטריצה שמודדת כיצד תכונות משתנות ביחד. זה מאפשר למודל לתפוס קשרים ליניאריים בין תכונות, ולייצר גבול החלטה ליניארי מדויק יותר.

הגרף מראה את היופי של גבול ליניארי: קו ישר פשוט, אבל שמפריד בצורה יעילה בין שתי התפלגויות גאוסיאניות. כל נקודה מסווגת על פי האזור שבו היא נמצאת. רוב הנקודות הכחולות באזור הכחול, רוב האדומות באזור האדום -- הפרדה מוצלחת.

הקשר בין נאיב בייס ל-LDA

נאיב בייס גאוסיאני הוא מקרה פרטי של LDA, כאשר מטריצת הקוואריאנס היא אלכסונית -- כלומר, אין קורלציה בין תכונות.
LDA מרפה את ההגבלה הזו ומאפשר קורלציות, אך בתמורה דורש יותר נתונים ויותר זמן חישוב.

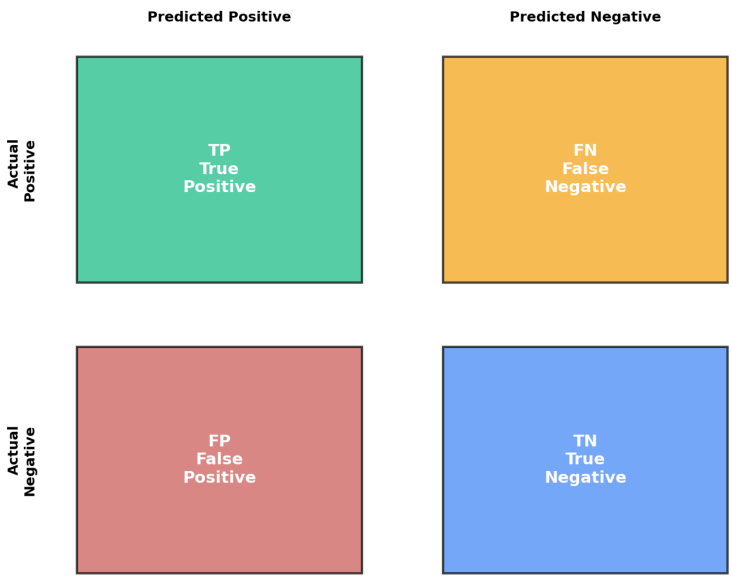
פרק 6

מדידת ההצלחה

6.1 מעבר לדיוק

בנינו מסווג. עכשיו השאלה היא: האם הוא טוב? התשובה האינטואיטיבית היא לספור כמה פעמים צדקנו -- אבל המציאות, כפי שנראה, מורכבת הרבה יותר.

כשאנחנו מעריכים מערכת סיווג, אנחנו זקוקים למפה שתראה לנו את כל הפינות. מפה שלא רק תספר לנו כמה פעמים היינו צודקים, אלא גם איפה טעינו, איך טעינו, ומה המחיר של כל טעות. איור 6.1 מציג את המדדים השונים שפותחו לצורך כך.

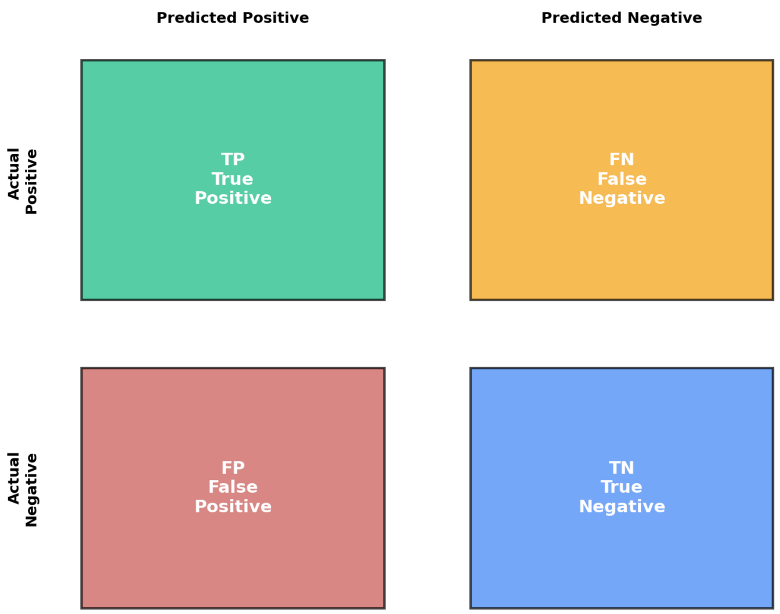


איור 6.1: מדדי הערכה למסווגים: מבט-על הכלים שלנו

6.2 מטריצת הבלבול

כדי להבין איך מסווג עובד באמת, אנחנו צריכים לפרק כל תוצאה לאחד מארבעה מצבים אפשריים. זו מטריצת הבלבול -- Confusion Matrix -- ושמה נובע מכך שהיא חושפת בדיוק איפה המסווג "מתבלבל".

תחשבו על איור 6.2 כעל מפת האפשרויות הבסיסית שלנו. כל תוצאה שהמסווג מייצר נופלת לאחד מארבעה ריבועים צבעוניים.



איור 6.2: ארבעת הרבעים של מטריצת הבלבול: כל תוצאה שייכת לאחד מהם

המטריצה מחולקת לפי שני צירים. ציר אחד -- האמת. מה המצב האמיתי? חולה או בריא? ספאם או מייל לגיטימי? זו התשובה הנכונה, זה מה שקורה בעולם האמיתי. הציר השני -- התחזית שלנו. מה המסווג אמר? מה הוא חזה?

כשמצליבים את שני הצירים האלה, מקבלים ארבע אפשרויות:

טבלה 6.1: ארבע הקטגוריות הבסיסיות של מטריצת הבלבול

	חיזוי: חיובי	**חיזוי: שלילי**
אמת: חיובי	נכון חיובי -- TP	פספוס -- FN
אמת: שלילי	התראת שווא -- FP	נכון שלילי -- TN

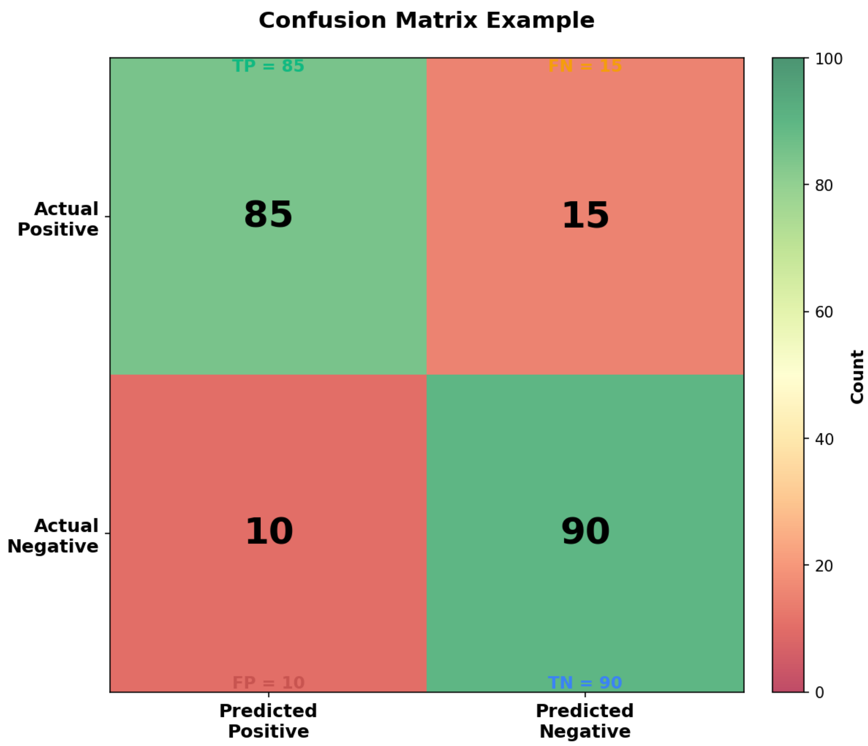
True Positive (TP) -- הרבע הירוק העליון משמאל. חזינו "חיובי" והמציאות אכן חיובית. איש חולה, וזיהינו אותו כחולה. זו הצלחה טהורה.

True Negative (TN) -- הרבע הכחול התחתון מימין. חזינו "שלילי" והמציאות אכן שלילית. איש בריא, וזיהינו אותו כבריא. גם זו הצלחה.

False Positive (FP) -- הרבע האדום התחתון משמאל. חזינו "חיובי" אבל המציאות שלילית. איש בריא, אבל טעינו וזיהינו אותו כחולה. זו התראת שווא -- False Alarm.

False Negative (FN) -- הרבע הכתום העליון מימין. חזינו "שלילי" אבל המציאות חיובית. איש חולה, ופספסנו אותו. זה פספוס מסוכן.

עכשיו בואו נסתכל על דוגמה קונקרטית. איור 6.3 מציג מטריצת בלבול אמיתית, עם מספרים.



איור 6.3: דוגמה למטריצת בלבול עם נתונים ממשיים: 002 מקרי בדיקה

במטריצה הזו אנחנו רואים 002 מקרי בדיקה. בואו נפרק אותה לפי השורות והעמודות:

השורות -- מייצגות את האמת. השורה העליונה היא כל המקרים שבאמת חיוביים (במקרה הזה, 001 מקרים). השורה התחתונה היא כל המקרים שבאמת שליליים (עוד 001 מקרים).

העמודות -- מייצגות את התחזית שלנו. העמודה השמאלית היא כל המקרים שחזינו כחיוביים. העמודה הימנית היא כל המקרים שחזינו כשליליים.

הריבוע הירוק העליון משמאל: 58 מקרים. אלו אנשים שהיו חולים באמת וזיהינו אותם נכון.

הריבוע הכתום העליון מימין: 51 מקרים. אלו אנשים שהיו חולים באמת, אבל פספסנו אותם ואמרנו שהם בריאים.

הריבוע האדום התחתון משמאל: 01 מקרים. אלו אנשים שהיו בריאים, אבל טעינו ואמרנו שהם חולים.

הריבוע הכחול התחתון מימין: 09 מקרים. אלו אנשים שהיו בריאים וזיהינו אותם נכון.

סכימת השורות והעמודות מגלה לנו דברים חשובים. השורה העליונה מסתכמת ל-001 (51+58) -- סך כל המקרים החיוביים האמיתיים. העמודה השמאלית מסתכמת ל-59 (01+58) -- סך כל המקרים שחזינו כחיוביים. והמספרים האלה יהיו הבסיס לחישוב כל המדדים שלנו.

6.3 הדיוק -- ומגבלותיו

המדד הכי פשוט הוא Accuracy -- דיוק. כמה פעמים צדקנו, מתוך כל המקרים?

דיוק

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6.1}$$

במילים פשוטות: סך ההצלחות (הנכונים החיוביים והנכונים השליליים), חלקי סך כל המקרים. נשמע הגיוני -- אבל הנה הבעיה.

נניח שאנחנו בודקים מחלה נדירה. מתוך 1000 אנשים, רק 10 חולים. עכשיו בואו ניצור מסווג "טיפש" -- מסווג שתמיד אומר "בריא", ללא יוצא מן הכלל. הוא אפילו לא מסתכל על הנתונים. הוא פשוט תמיד עונה "בריא".

מה הדיוק של המסווג הזה? 990 מתוך 1000. כלומר, 99%! מסווג עם דיוק של תשעים ותשעה אחוז -- נשמע מעולה, נכון?

לא. המסווג הזה חסר ערך לחלוטין. הוא מפספס את כל עשרת החולים. אם המטרה שלנו היא לזהות חולים כדי לטפל בהם, המסווג הזה כושל באופן מוחלט. זו בדיוק הסיבה שדיוק לבדו לא מספיק. אנחנו צריכים מדדים שמסתכלים על כל אחת מהקטגוריות בנפרד.

6.4 רגישות ודיוק חיובי

אז אילו שאלות אנחנו באמת רוצים לשאול?

שאלה ראשונה: מתוך כל האנשים שבאמת חולים -- כמה מהם המסווג שלי מזהה? זה נקרא **רגישות** או Recall:

רגישות

רגישות (Recall):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6.2}$$

מתוך החולים האמיתיים -- כמה זיהינו?

המכנה כאן הוא $\text{TP} + \text{FN}$ -- כלומר, סך כל המקרים החיוביים האמיתיים (אלה שזיהינו נכון ואלה שפספסנו). המונה הוא TP -- אלה שזיהינו נכון. אז Recall עונה על השאלה: מתוך 001 חולים, כמה תפסתי?

רגישות גבוהה משמעותה שאנחנו לא מפספסים הרבה. אם יש לנו Recall של 95%, זה אומר שתפסנו 59 מתוך 001 חולים. רק 5 חמקו לנו.

שאלה שנייה: מתוך כל האנשים שהמסווג זיהה כחולים -- כמה באמת חולים? זה נקרא **דיוק חיובי** או Precision:

דיוק חיובי

דיוק חיובי (Precision):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6.3}$$

מתוך אלה שזיהינו כחולים -- כמה באמת חולים?

המכנה כאן הוא $\text{TP}+\text{FP}$ -- כלומר, סך כל המקרים שזיהינו כחיוביים (נכון ובטעות). המונה הוא TP -- אלה שזיהינו נכון. אז Precision עונה על השאלה: אם המסווג אומר "חולה", מה הסיכוי שהוא באמת חולה?

דיוק חיובי גבוה משמעותו שיש לנו מעט התראות שווא. אם יש לנו Precision של 90%, זה אומר שמתוך כל 001 אנשים שזיהינו כחולים, 09 באמת חולים ורק 01 היו בריאים (התראת שווא).

6.5 האיזון

יש מתח טבעי בין שני המדדים האלה. לא ניתן לקבל את שניהם במקסימום בו-זמנית -- תמיד יש פשרה.

נניח שאני רוצה Recall גבוה מאוד -- רגישות מקסימלית. אני רוצה לתפוס את כל החולים, בלי לפספס אף אחד. מה אעשה? אזהיר על כל דבר. אפחית את הסף. כל ספק קטן -- "חולה". בגישה הזו, אתפוס את כל החולים (או כמעט כולם), אבל המחיר יהיה הרבה התראות שווא. הרבה אנשים בריאים יסומנו בטעות כחולים. כלומר, Precision נמוך.

מצד שני, נניח שאני רוצה Precision גבוה מאוד -- דיוק חיובי מקסימלי. אני רוצה שכל פעם שאני אומר "חולה", אני באמת בטוח. מה אעשה? אזהיר רק כשאני בטוח

מאוד. ארים את הסף. רק במקרים ברורים מאוד -- "חולה". בגישה הזו, כמעט לא יהיו לי התראות שווא, אבל המחיר יהיה שאני אפספס חולים. כלומר, Recall נמוך.

איך מאזנים בין השניים? אחת הדרכים היא **F1-Score** -- ממוצע הרמוני של שני המדדים:

F1-Score

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

למה ממוצע הרמוני ולא ממוצע רגיל? כי ממוצע הרמוני "קשה יותר לרצות". אם אחד המדדים נמוך מאוד, הממוצע ההרמוני יהיה נמוך -- גם אם המדד השני גבוה. זה מאלץ את המסווג להיות טוב בשני המדדים יחד, לא רק באחד מהם.

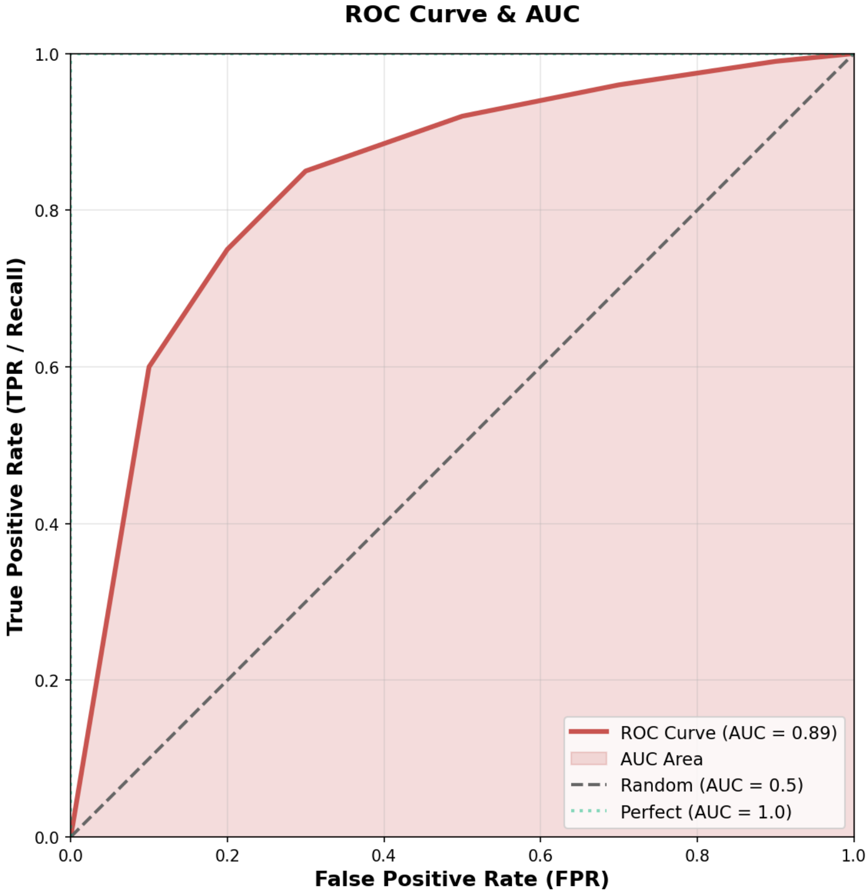
6.6 עקומת ROC

אבל רגע -- איך אנחנו "מורידים את הסף" או "מעלים את הסף"? מה זה בכלל אומר?

רוב המסווגים לא פשוט עונים "כן" או "לא". הם נותנים ציון -- הסתברות. "הסיכוי שהאדם הזה חולה הוא 37.0". עכשיו אנחנו צריכים להחליט: מתי נאמר "חולה"? אם הציון מעל 5.0? מעל 7.0? מעל 9.0?

זה הסף -- Threshold. וכל סף שונה ייתן לנו תוצאות שונות. סף נמוך -- הרבה חולים שנזהה, אבל גם הרבה התראות שווא. סף גבוה -- מעט התראות שווא, אבל גם חולים שנפספס.

עקומת ROC (Receiver Operating Characteristic) מראה לנו את כל האפשרויות האלה בגרף אחד. איור 6.4 מציג את העקומה הזו.



איור 6.4: עקומת ROC: התמורה בין רגישות להתראות שווא בסיפי החלטה שונים

בואו נפרק את הגרף הזה לרכיבים:

ציר ה-X (האופקי): False Positive Rate (FPR) -- שיעור ההתראות השווא. זה מחושב כ-$\frac{\text{FP}}{\text{FP+TN}}$. כלומר, מתוך כל האנשים הבריאים, כמה זיהינו בטעות כחולים?

ציר ה-Y (האנכי): True Positive Rate (TPR) -- שיעור החיובים האמיתיים. זה בדיוק Recall! $\frac{\text{TP}}{\text{TP+FN}}$. מתוך כל החולים, כמה זיהינו?

העקומה האדומה: זו עקומת ROC של המסווג שלנו. כל נקודה על העקומה מתאימה לסף החלטה אחר. בקצה השמאלי התחתון -- סף מאוד גבוה (אנחנו כמעט לא אומרים "חולה"). בקצה הימני העליון -- סף מאוד נמוך (אנחנו כמעט תמיד אומרים "חולה").

הקו המקווקו האלכסוני: זה מסווג אקראי. מסווג שמטיל מטבע. אם אנחנו מטילים מטבע, אז שיעור החיובים האמיתיים יהיה שווה לשיעור ההתראות השווא -- קו ישר באלכסון. זה קו הבסיס שלנו.

הנקודה המושלמת: היא בפינה השמאלית העליונה -- $(0,1)$. כלומר, אפס התראות שווא ומאה אחוז תפיסה של חולים. זה מסווג מושלם, שכמעט אף פעם לא קיים במציאות.

ככל שהעקומה קרובה יותר לפינה השמאלית העליונה, המסווג טוב יותר. ככל שהיא קרובה יותר לקו המקווקו האלכסוני, המסווג חסר ערך יותר (כמו הטלת מטבע).
שטח מתחת לעקומה -- AUC (Area Under Curve): זהו המדד המספרי שמסכם את כל העקומה למספר אחד.

- $\text{AUC} = 1.0$ -- מסווג מושלם. העקומה עוברת בדיוק דרך הפינה השמאלית העליונה.
- $\text{AUC} = 0.5$ -- מסווג אקראי. העקומה היא הקו האלכסוני.
- $\text{AUC} > 0.9$ -- מסווג מצוין. זה בדרך כלל נחשב לביצועים טובים מאוד.
- $\text{AUC} > 0.8$ -- מסווג טוב. זה לרוב מספיק ליישומים רבים.
- $\text{AUC} < 0.7$ -- מסווג בינוני או חלש.

בדוגמה שלנו, $\text{AUC} = 0.89$ -- זה מסווג טוב. השטח הוורוד מסומן בגרף, והוא מכסה 98 אחוז משטח הריבוע.
למה AUC חשוב? כי הוא עצמאי בסף. הוא לא תלוי באיזה סף החלטה בחרנו. הוא מודד את היכולת הכוללת של המסווג להפריד בין חיוביים לשליליים, בכל הספים האפשריים.

6.7 דוגמה מעשית

בואו ניישם את כל הכלים האלה על דוגמה קונקרטית: מסווג דואר זבל. בדקנו 1000 הודעות דוא"ל. התוצאות:

טבלה 6.2: מטריצת בלבול למסווג דואר זבל

	חיזוי: ספאם	**חיזוי: רגיל**
אמת: ספאם	180	20
אמת: רגיל	30	770

בואו נחשב את כל המדדים:

$$\text{Accuracy} = \frac{180 + 770}{1000} = \frac{950}{1000} = 95\% \tag{6.5}$$

$$\text{Recall} = \frac{180}{180 + 20} = \frac{180}{200} = 90\% \tag{6.6}$$

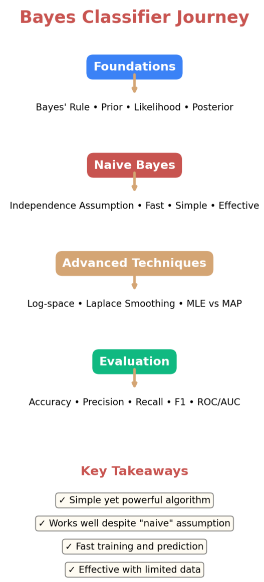
$$\text{Precision} = \frac{180}{180 + 30} = \frac{180}{210} = 86\% \tag{6.7}$$

$$F_1 = 2 \cdot \frac{0.86 \cdot 0.90}{0.86 + 0.90} = 2 \cdot \frac{0.774}{1.76} = 0.88 \tag{6.8}$$

דיוק כללי של 95% -- נשמע מצוין. אבל בואו נעמיק: תפסנו 90% מהספאם (רגישות טובה), אבל גם סימנו בטעות 14% ממיילים רגילים כספאם (דיוק חיובי של 86%). זה מקובל? תלוי בהקשר. אם המחיר של פספוס ספאם נמוך (סתם מעצבן), אבל המחיר של לאבד מייל חשוב גבוה -- אולי נרצה לשפר את ה-Precision.

6.8 המשימה שלכם

עכשיו שיש לנו את כל הכלים -- מהתיאוריה של בייס ועד למדדי ההערכה -- הגיע הזמן לחבר את כל החלקים. איור 6.5 מסכם את המסע שלנו.



איור 6.5: מסע מסווג בייס: מיסודות התיאוריה ועד למדידת ההצלחה

התרשים מציג ארבעה שלבים שעברנו יחד:

יסודות (Foundations): חוק בייס, פריור, נראות (Likelihood), ופוסטריור. אלו אבני היסוד המתמטיות.

נאיב בייס (Naive Bayes): ההנחה של אי-תלות בין המאפיינים -- הנחה "נאיבית" שמפשטת את החישוב ועובדת בפועל. מסווג מהיר, פשוט, ויעיל.

טכניקות מתקדמות (Advanced Techniques): מרחב לוגריתמי כדי למנוע Under-flow, החלקת לפלס (Laplace Smoothing) כדי לטפל בבעיית האפס, וההשוואה בין MLE ל-MAP.

הערכה (Evaluation): דיוק (Accuracy), דיוק חיובי (Precision), רגישות (Recall), F1, ועקומות ROC/AUC. כלים שמאפשרים לנו למדוד באמת אם המסווג שלנו עובד.

התרשים גם מפרט את המסרים המרכזיים (Key Takeaways):

- אלגוריתם פשוט אבל חזק

- עובד היטב למרות ההנחה הנאיבית
- מהיר באימון ובחיזוי
- יעיל גם עם מעט נתונים

ועכשיו -- הגיע הזמן לבנות בעצמכם.

תרגיל: סיווג פרחי אירוס

1. הורידו את מאגר הנתונים של Iris (051 דגימות, 3 מינים, 4 מאפיינים)
2. חלקו את הנתונים: 75% לאימון, 25% לבדיקה
3. ממשו מסווג נאיב בייס ב-NumPy בלבד -- ללא ספריות למידת מכונה
4. ממשו את אותו המסווג שוב באמצעות scikit-learn
5. השוו את התוצאות: האם קיבלתם דיוק דומה? חשבו מדוע
6. חשבו את כל המדדים: Accuracy, Precision, Recall, F1
7. צרו מטריצת בלבול ונתחו: איפה המסווג טועה?

המטרה של התרגיל הזה אינה רק לקבל תוצאות טובות. המטרה היא להבין את הנוסחאות מבפנים -- לא רק להפעיל פונקציה מספרייה. כשאתם ממשים בעצמכם את החישוב של הפריור, הנראות, והפוסטריור, אתם רואים איך המספרים באמת זורמים דרך המערכת. איך הסתברות קטנה אחת משפיעה על התוצאה. איך החלקת לפלס משנה את ההתנהגות במקרי קיצון.

ורק אז, כשאתם ממשים את אותו הדבר עם scikit-learn, אתם מבינים באמת מה הספרייה עושה מאחורי הקלעים. אתם לא עוד משתמשים ב-"קופסה שחורה" -- אתם יודעים בדיוק מה קורה בפנים.

זו הדרך היחידה ללמוד באמת.

סיכום

הספר הזה סיפר סיפור. סיפור של כומר אנגלי בן המאה ה-18, שגילה דרך לחשוב על העתיד בעזרת העבר.

- **הכומר שחשב על העתיד** -- פגשנו את תומס בייס ואת הרעיון הפשוט: לעדכן אמונות לאור ראיות
- **שפת האי-ודאות** -- למדנו לדבר בשפה של מספרים בין אפס לאחד
- **המסווג בפעולה** -- בנינו מכונת החלטות אוטומטית, מנוסחה למעשה
- **עולם של מספרים** -- עברנו מעולם של כן/לא לעולם של רציפים
- **ההנחה הנאיבית** -- גילינו שלפעמים הנחה שגויה נותנת תוצאות נכונות
- **מדידת ההצלחה** -- למדנו לשאול את השאלות הנכונות על ביצועים

עכשיו התור שלכם. קחו את הנוסחאות, בנו מסווג, ותראו בעצמכם את הקסם.