1. Introduction

1.1 What is Machine Learning?

1.1.1 The Basic Concept

Artificial Intelligence (AI)

בינה מלאכותית הינו תחום בו תוכנות מחשב או מנגנון טכנולוגי אחר מחקה מנגנון חשיבה אנושי. בתחום רחב זה יש רמות שונות של בינה מלאכותית – יש מערכות שמסוגלות ללמוד דפוסי התנהגות ולהתאים את עצמן לשינויים, ואילו יש מערכות שאמנם מחקות מנגנון חשיבה אנושי אך הן לא מתוחכמות מעבר למה שתכנתו אותן בהתחלה. שואב רובוטי היודע לחשב את גודל החדר ואת מסלול הניקוי האופטימלי פועל לפי פרוצדורה ידועה מראש, ואין בו תחכום מעבר לתכנות הראשוני שלו. לעומת זאת תוכנה היודעת לסנן רעשים באופן מסתגל, או להמליץ על שירים בנגן מוזיקה בהתאם לסגנון של המשתמש, משתמשות בבינה מלאכותית ברמה גבוהה יותר, כיוון שהן לומדות עם הזמן דברים חדשים.

המונח בינה מלאכותית מתייחס בדרך כלל למערכת שמחקה התנהגות אנושית, אך היא שגרתית, לא לומדת משהו חדש, ועושה את אותו הדבר כל הזמן. מערכת זו יכולה להיות משוכללת ולחשב דברים מסובכים ואף להסיק מסקנות על דוגמאות חדשות שהיא מעולם לא ראתה, אך תמיד עבור אותו הקלט (Input), יהיה אותו הפלט (Output).

ניקח לדוגמא מערכת סטרימינג של סרטים, למשל Netflix. כחלק משיפור המערכת והגדלת זמני הצפייה, ניתן לבנות מנגנון המלצות הבנוי על היסטוריית השימוש של הלקוחות שלי במערכת – איזה סרטים הם רואים, איזה ז'אנרים ומתי. כשיש מעט צופים ומעט סרטים, ניתן לעשות זאת באופן ידני – למלא טבלאות של הנתונים, לנתח אותם ידנית ולבנות מערכת חוקים שמהווה מנוע המלצות מבוסס AI. ניקח לדוגמא אדם שצופה ב"פארק היורה" וב"אינדיאנה ג'ונס" – סביר שהמערכת תמליץ לו לצפות גם ב-"פולטרגייסט". אדם שצופה לעומת זאת ב-"אהבה בין הכרמים" ו"הבית על האגם", ככל הנראה כדאי להמליץ לו על "הגשרים של מחוז מדיסון".

מערכת זו יכולה לעבוד טוב, אך בנקודה מסוימת כבר לא ניתן לנסח אותה כפרוצדורה מסודרת וכאוסף של חוקים ידוע מראש. מאגר הסרטים גדל, נוספים סוגים נוספים של סרטים (כמו למשל סדרות, תוכניות ריאליטי ועוד) ובנוסף רוצים להתייחס לפרמטרים נוספים – האם הצופה ראה את כל הסרט או הפסיק באמצע, מה גיל הצופה ועוד. מערכת הבנויה באופן קלאסי אינה מסוגלת להתמודד עם כמויות המידע הקיימות, וכמות הכללים שנדרש לחשוב עליהם מראש היא עצומה ומורכבת לחישוב.

נתבונן על דוגמא נוספת – מערכת לניווט רכב. ניתן להגדיר כלל פשוט בו אם משתמש יוצא מתל אביב ורוצה להגיע לפתח תקווה, אז האפליקציה תיקח אותו דרך מסלול ספציפי שנבחר מראש. מסלול זה לא מתחשב בפרמטרים קריטיים כמו מה השעה, האם יש פקקים או חסימות ועוד. כמות הפרמטרים שיש להתייחס אליהם איננה ניתנת לטיפול על ידי מערכת כללים ידועה מראש, וגם הפונקציונאליות המתאפשרת היא מוגבלת מאוד – למשל לא ניתן לחזות מה תהיה שעת ההגעה וכדומה.

Machine Learning (ML)

למידת מכונה הוא תת תחום של בינה מלאכותית, הבא להתמודד עם שני האתגרים שתוארו קודם – היכולת לתכנת מערכת על בסיס מסות של נתונים ופרמטרים, וחיזוי דברים חדשים כתלות בפרמטרים רבים שיכולים להשתנות עם הזמן. מנגנוני ML מנתחים כמויות אדירות של דאטה ומנסות להגיע לאיזו תוצאה. אם מדובר באפליקציית ניווט, המערכת תנתח את כל אותם הפקטורים ותנסה לחשב את משך הנסיעה המשוער. נניח והיא חזתה 20 דקות נסיעה. אם בסופו של דבר הנסיעה ארכה 30 דקות, האלגוריתם ינסה להבין איזה פקטור השתנה במהלך הדרך ומדוע הוא נכשל בחיזוי (למשל: הכביש בן ארבעה נתיבים, אבל במקטע מסיים הוא מצטמצם לאחד וזה מייצר עיכוב, וזה עיכוב קבוע ברוב שעות היממה ולא פקק אקראי). בהינתן מספיק מקרים כאלה, האלגוריתם "מבין" שהוא טועה, והוא פשוט יתקן את עצמו ויכניס למערך החישובים גם פקטור של מספר נתיבים ויוריד אולי את המשקל של הטמפרטורה בחוץ. וככה באופן חזרתי האלגוריתם שוב ושוב מקבל קלט, מוציא פלט ובודק את התוצאה הסופית. לאחר מכן הוא בודק היכן הוא טעה, משנה את עצמו, מתקן את המשקל שהוא נותן לפקטורים שונים ומשתכלל מנסיעה לנסיעה.

במערכות אלה ה-Input נשאר לכאורה קבוע, אבל ה-Output משתנה – עבור זמני יציאה שונים, האלגוריתם יעריך זמני נסיעה שונים, כתלות במגוון הפרמטרים הרלוונטיים.

מערכות ML משמשות את כל רשתות הפרסום הגדולות. כל אחת מנסה בדרכה שלה לחזות למשל, איזה משתמש שהקליק על המודעה צפוי שיבצע רכישה. הפלטפורמות השונות מנסות לזהות כוונה (Intent) על ידי למידה

מניסיון. בהתחלה הן פשוט ניחשו על פי כמה פקטורים שהוזנו להם על ידי בני אדם. נניח, גוגל החליטה שמי שצופה בסרטוני יוטיוב של Unboxing הוא ב-Intent גבוה של רכישה. בהמשך הדרך, בהנחה והמשתמש מבצע רכישה כלשהי, האלגוריתם מקבל "נקודה טובה". אם הוא לא קנה, האלגוריתם מקבל "נקודה רעה". ככל שהוא מקבל יותר נקודות טובות ורעות, האלגוריתם יודע לשפר את עצמו, לתת משקל גדול יותר לפרמטרים טובים ולהזניח פרמטרים פחות משמעותיים. אבל רגע, מי אמר למערכת להסתכל בכלל בסרטוני Unboxing?

האמת שהיא שאף אחד. מישהו, בנאדם, אמר למערכת לזהות את כל הסרטונים שמשתמש צופה בהם ביוטיוב, לזהות מתוך הסרטון, האודיו, תיאור הסרטון ומילות המפתח וכו' – איזה סוג סרטון זה. ייתכן ואחרי מיליארדי צפיות בסרטונים, האלגוריתם מתחיל למצוא קשר בין סוג מסוים של סרטונים לבין פעולות כמו רכישה באתר. באופן הזה, גוגל מזינה את האלגוריתם בכל הפעולות שהמשתמש מבצע. המיילים שהוא קורא, המקומות שהוא מסתובב בהם, התמונות שהוא מעלה לענן, ההודעות שהוא שולח, כל מידע שיש אליו גישה. הכל נשפך לתוך מאגר הנתונים העצום בו מנסה גוגל לבנות פרופילים ולמצוא קשר בין פרופיל האדם לבין הסיכוי שלו לרכוש או כל פעולה אחרת שבא לה לזהות.

המכונה המופלאה הזו לומדת כל הזמן דברים חדשים ומנסה כל הזמן למצוא הקשרים, לחזות תוצאה, לבדוק אם היא הצליחה, ואם לא לתקן את עצמה שוב ושוב עד שהיא פוגעת במטרה. חשוב לציין שלמכונה אין סנטימנטים, כל המידע קביל ואם היא תמצא קשר מוכח בין מידת הנעליים של בנאדם לבין סרטונים של בייבי שארק, אז היא תשתמש בו גם אם זה לא נשמע הגיוני.

חשוב לשים לב לעניין המטרה – המטרה היא לא המצאה של האלגוריתם. הוא לא קם בבוקר ומחליט מה האפליקציה שלכם צריכה לעשות. המטרה מוגדרת על ידי היוצר של המערכת. למשל – חישוב זמן נסיעה, בניית מסלול אופטימלי בין A ל-B וכו'. המטרה של גוגל – שמשתמש יבצע רכישה, והכל מתנקז לזה בסוף, כי גוגל בראש ובראשונה היא מערכת פרסום. אגב, גם ההגדרה של מסלול "אופטימלי" היא מעשה ידי אדם. המכונה לא יודעת מה זה אופטימלי, זו רק מילה. אז צריך לעזור לה ולהגיד לה שאופטימלי זה מינימום זמן, מעט עצירות, כמה שפחות רמזורים וכו'. לסיכום, המטרה מאופיינת על ידי האדם ולא על ידי המכונה. המכונה רק חותרת למטרה שהוגדר לה.

יש מנגנוני ML המתבססים על דאטה מסודר ומתויג כמו ב-Netflix, עם כל המאפיינים של הסרטים אבל גם עם המאפיינים של הצופים (מדינה, גיל, שעת צפייה וכו'..). לעומת זאת יש מנגנוני ML שמקבלים טיפה יותר חופש ומתבססים על מידע חלקי מאד (יש להם מידע על כל על הסרטים, אבל אין להם מידע על הצופה). מנגנונים אלו לא בהכרח מנסים לבנות מנוע המלצות אלא מנסים למצוא חוקיות בנתונים, חריגות וכו'.

כך או כך, המערך הסבוך הזה הקרוי ML בנוי מאלגוריתמים שונים המיומנים בניתוח טקסט, אלגוריתמים אחרים המתמקדים בעיבוד אודיו, כאלה המנתחים היסטוריית גלישה או זיהוי מתוך דף ה-Web בו אתם צופים ועוד. עשרות או מאות מנגנונים כאלה מסתובבים ורצים ובונים את המפה השלמה. ככה רוב רשתות הפרסום הגדולות עובדות. ככל שהמכונה של גוגל/פייסבוק תהיה חכמה יותר, ככה היא תדע להציג את המודעה המתאימה למשתמש הנכון, בזמן הנכון ועל ה-Device המתאים.

1.1.2 Data, Tasks and Learning

כאמור, המטרה הבסיסית של למידת מכונה היא היכולת להכליל מתוך הניסיון, ולבצע משימות באופן מדויק ככל הניתן על דאטה חדש שעדיין לא נצפה, על בסיס צבירת ניסיון מדאטה קיים. באופן כללי ניתן לדבר על שלושה סוגים של למידה:

למידה מונחית – הדאטה הקיים הינו אוסף של דוגמאות, ולכל דוגמא יש תווית (label). מטרת האלגוריתמים במקרה זה היא לסווג דוגמאות חדשות שלא נצפו בתהליך הלמידה. באופן פורמלי, עבור דאטה $x \in \mathbb{R}^{n\times d}$, יש סט labels – $y \in \mathbb{R}^{1\times d}$, ומחפשים את האלגוריתם שמבצע את המיפוי $g: X \to Y$ בצורה הטובה ביותר, כלומר בהינתן דוגמא חדשה $x \in \mathbb{R}^n$, המטרה היא למצוא עבורה את ה-y הנכון. המיפוי נמדד ביחס לפונקציות מחיר, כפי שיוסבר בהמשך בנוגע לתהליך הלמידה.

למידה לא מונחית – הדאטה הקיים הינו אוסף של דוגמאות במרחב, בלי שנתון עליהן מידע כלשהו המבחין ביניהן. במקרה זה, בדרך כלל האלגוריתמים יחפשו מודל המסביר את התפלגות הנקודות – למשל חלוקה לקבוצות שונות וכדומה.

למידה באמצעות חיזוקים – הדאטה בו נעזרים אינו מצוי בתחילת התוכנית אלא נאסף עם הזמן. ישנם סוכנים הנמצאים בסביבה מסוימת ומעבירים מידע למשתמש, והוא בתורו ילמד אסטרטגיה בה הסוכנים ינקטו בצעדים הטובים עבורם.

האלגוריתמים השונים של הלמידה מתחלקים לשתי קבוצות – מודלים דיסקרימנטיביים המוציאים פלט על בסיס מידע נתון, אך לא יכולים ליצור מידע חדש בעצמם, ומודלים גנרטיביים, שלא רק לומדים להכליל את הדאטה הנלמד גם עבור דוגמאות חדשות, אלא יכולים גם להבין את מה שהם ראו וליצור מידע חדש על בסיס הדוגמאות שנלמדו.

כאמור, בשביל לבנות מודל יש צורך בדאטה. מודל טוב הוא מודל שמצליח להכליל מהדאטה הקיים גם לדאטה חדש. המודל למעשה מנסה למצוא דפוסים בדאטה הקיים, מהם הוא יוכל להסיק מסקנות גם על דוגמאות חדשות. כדי לוודא שהמודל אכן מצליח להכליל גם על דוגמאות חדשות, בדרך כלל מחלקים את הדאטה הקיים לשניים – סט אימון (training set) וסט מבחן (test set). אם המודל מצליח למצוא דפוסים בסט האימון שנכונים גם עבור הטסט סט, זה סימן שהמודל הצליח למצוא כללים שיכולים להיות נכונים גם לדוגמאות חדשות שיבואו.

מגוון התחומים בהם משתמשים בכלים של למידה הוא עצום, עד כדי כך שכמעט ואין תחום בו לא נכנס השימוש באלגוריתמים לומדים. דוגמאות בולטות למשימות בהם משתמשים באלגוריתמים לומדים: סיווג, רגרסיה (מציאת קשר בין משתנים), חלוקה לקבוצות, מערכת המלצות, הורדת מימד, עיבוד שפה טבעית ועוד.

1.2 Applied Math

האלגוריתמים של למידת מכונה נסמכים בעיקרם על שלושה ענפים מתמטיים; אלגברה לינארית, חשבון דיפרנציאלי והסתברות. בפרק זה נציג את העקרונות הנדרשים בלבד, ללא הרחבה, על מנת להבין את הנושאים הנדונים בספר זה.

1.2.1 Linear Algebra

<u>וקטורים ומרחבים וקטורים</u>

באופן מתמטי מופשט, וקטורים, המסומנים בדרך כלל ע"י $\vec{x}$ או על ידי $\boldsymbol{x}$, הינם אובייקטים הנמצאים במרחב וקטורי $(V,+)$ מעל שדה $\mathbb{F}$. מהו אותו מרחב וקטורי?

ראשית השדה, $\mathbb{F}$, הוא קבוצת מספרים המקיימים תכונות מתמטיות מסוימות. לדיון בספר זה, השדה הוא קבוצת המספרים הממשיים – $\mathbb{R}$, או קבוצת המספרים המרוכבים – $\mathbb{C}$. שנית, נשים לב כי המרחב הוקטורי דורש גם הגדרת פעולת חיבור $(+)$.

כעת, $(V,+)$ היא מרחב וקטורי אם הוא מקיים את התכונות הבאות:

(I) קיים איבר אפס (וקטור אפס) כך שכל $\vec{x}$ בקבוצה V מקיים: $\vec{x}+\vec{0}=\vec{0}+\vec{x}=\vec{x}$.
(II) לכל איבר בשדה a ולכל $\vec{x}$ ו-$\vec{y}$ בקבוצה V, גם $a\cdot\vec{x}+\vec{y}$ הינו גם איבר בקבוצה V.

הערה: קיימות דרישות נוספות למרחב וקטורי, אך הם מעבר לנדרש בספר זה.

דוגמאות:

א. וקטורים גאומטריים:
מערך חד מימדי $\vec{x}=(x_1,x_2\dots x_n)$ (n-יה סדורה) נקרא וקטור גאומטרי n מימדי, כאשר רכיבי הוקטור הם איברים בשדה $\mathbb{F}$. האיבר x_i, המיוצג על ידי האינדקס i מתאר את מיקום האיבר. מרחב זה מסומן ע"י $\mathbb{F}^n$.
נראה שמרחב זה הוא אכן מרחב וקטורי:
<u>חיבור וקטורים:</u>

$$\vec{x}=(x_1,x_2\dots x_n),\quad \vec{y}=(y_1,y_2,\dots,y_n)\quad\rightarrow\quad \vec{x}+\vec{y}=(x_1+y_1,x_2+y_2,\dots,x_n+y_n)$$

<u>וקטור אפס:</u>

$$\vec{0}=(0,0,\dots,0)$$

<u>כפל בסקלר:</u>

$$\vec{x}=(x_1,x_2\dots x_n)\quad\rightarrow\quad a\,\vec{x}=(a\,x_1,a\,x_2,\dots,a\,x_n)$$

הערה: לשם פשטות, בהמשך, נכנה וקטור גאומטרי כ"וקטור" בלבד.

ב. מטריצות:

מערך דו מימדי $\begin{pmatrix} A_{11} & \dots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nm} \end{pmatrix}$, אשר רכיביו הם איברים בשדה $\mathbb{F}$, נקרא מטריצה מסדר $n \times m$, כאשר n הוא מספר השורות ו-m הוא מספר העמודות במערך. האיברים במטריצה A_{ij} מיוצגים ע"י שני אינדקסים – i,j, המתארים את השורה והעמודה בהתאמה. מרחב זה מסומן בדרך כלל ע"י $\mathbb{F}^{n\times m}$.
נוכיח שמרחב זה הוא אכן מרחב וקטורי:

<u>חיבור מטריצות:</u>

$$\hat{A} = \begin{pmatrix} A_{11} & \dots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nm} \end{pmatrix}, \hat{B} = \begin{pmatrix} B_{11} & \dots & B_{1m} \\ \vdots & \ddots & \vdots \\ B_{n1} & \dots & B_{nm} \end{pmatrix} \rightarrow \hat{A} + \hat{B} = \begin{pmatrix} A_{11} + B_{11} & \dots & A_{1n} + B_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} + B_{m1} & \dots & A_{nm} + B_{nm} \end{pmatrix}$$

<u>מטריצת אפס:</u>

$$\hat{0} = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

<u>כפל בסקלר:</u>

$$\hat{A} = \begin{pmatrix} A_{11} & \dots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nm} \end{pmatrix} \quad \rightarrow \quad a\,\hat{A} = \begin{pmatrix} a\,A_{11} & \dots & a\,A_{1m} \\ \vdots & \ddots & \vdots \\ a\,A_{n1} & \dots & a\,A_{nm} \end{pmatrix}$$

ניתן להבחין כי הווקטורים הגיאומטריים שהוגדרו בדוגמא א, הם בעצם מטריצות במימד $n = n \times 1$.

ג. פולינומים:
פולינומים מסדר n הינם ביטויים מהסוג $a_0 + a_1x + a_2x^2 + \cdots a_nx^n$, כאשר n מייצג את החזקה הגדולה ביותר ו-a_i הם איברים בשדה. מרחב זה מסומן בדרך כלל ע"י $P_n(x)$.

בכל הדוגמאות לעיל קל להראות שהן אכן מהוות מרחב וקטורי. רשימה חלקית לדוגמאות נוספות לווקטורים (ולמרחבים וקטורים) כוללת למשל מרחבי פונקציות או אפילו אותות אלקטרומגנטיים. כאן בחרנו להציג רק את הדוגמאות הרלוונטיות לספר זה.

<u>פעולות חשבון על מטריצות ווקטורים:</u>

כמו שנזכר לעיל, הווקטורים הגיאומטריים שהוגדרו בדוגמא א, הם בעצם מטריצות במימד $n = n \times 1$. לכן, פעולות החשבון מוגדרות באופן זהה.

- <u>חיבור וחיסור בין שני מטריצות:</u>

$\hat{A}, \hat{B} \in \mathbb{F}^{n\times m}$ כאשר A_{ij}, B_{ij} הם האיברים בשורה i בעמודה j של המטריצות $\hat{A}, \hat{B}$ בהתאמה. אז, האיבר בשורה i בעמודה j של מטריצת הסכום (או ההפרש) הינו

$$(A \pm B)_{ij} = A_{ij} \pm B_{ij}$$

(הגדרת חיבור המטריצות בעצם כבר ניתנה בדוגמא א לעיל)

שים לב: ניתן לחסר ולחסר מטריצות רק בעלות אותו המימד.

- <u>כפל בין שתי מטריצות:</u>

$\hat{A} \in \mathbb{F}^{n\times k}, \hat{B} \in \mathbb{F}^{k\times m}$ הן שתי מטריצות, כאשר מספר השורות במטריצה $\hat{A}$ שווה למספר העמודות של מטריצה $\hat{B}$ (אך שתי המטריצות אינן בהכרח בעלות אותו מימד). במקרה כזה, מכפלת המטריצות מוגדרת על ידי:

$$\hat{A} \cdot \hat{B} = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} + \cdots A_{1k}B_{k1} & \dots & A_{11}B_{1n} + A_{1k}B_{kn} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} + \cdots + A_{mk}B_{kn} & \dots & A_{n1}B_{1m} + \cdots A_{nk}B_{km} \end{pmatrix}$$

למעשה כל איבר בתוצאה הינו סכום של מכפלת שורה i ממטריצה A בעמודה j ממטריצה B:

$$\left(\hat{A}\cdot\hat{B}\right)_{ij}=\sum_{k}A_{ik}B_{kj}$$

שים לב: על מנת שכפל המטריצות יהיה מוגדר מספר העמודות ב-$\hat{A}$ שווה למספר השורות ב$\hat{B}$.

עבור מטריצות ריבועיות (מסדר $n\times n$), מוגדר גם הכפל $\hat{A}\hat{B}$ וגם הכפל $\hat{B}\hat{A}$, אולם ייתכן ו $\hat{A}\hat{B}\neq\hat{B}\hat{A}$.

- <u>שחלוף:</u>

החלפת שורות בעמודות, או 'סיבוב' המטריצה. נניח מטריצה $\hat{A}\in\mathbb{F}^{n\times m}$, אז השחלוף שלה, המסומן כ-$\hat{A}^T$ הוא:

$$\left(\hat{A}^T\right)_{ij}=A_{ji}$$

ובאופן מפורש:

$$\hat{A}=\begin{pmatrix}A_{11} & \dots & A_{1m}\\ \vdots & \ddots & \vdots\\ A_{n1} & \dots & A_{nm}\end{pmatrix}\ \rightarrow\ \hat{A}^T=\begin{pmatrix}A_{11} & \dots & A_{n1}\\ \vdots & \ddots & \vdots\\ A_{1m} & \dots & A_{nm}\end{pmatrix}$$

שים לב שהמטריצה החדשה; $\hat{A}^T$ הינה במימד $m\times n$. בנוסף ניתן להוכיח כי מתקיים: $\hat{A}=\left(\hat{A}^T\right)^T$.

שחלוף של וקטור שורה, נותן וקטור עמודה ולהפך.

- <u>מטריצת יחידה:</u>

מטריצת יחידה, הינה מטריצה ריבועית (מסדר $n\times n$), המסומנת על ידי $\mathbb{I}_n$ ומוגדרת כך שכל איבריה אפס מלבד איברי האלכסון הראשי המקבלים את הערך 1:

$$(\mathbb{I}_n)_{ij}=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

ובאופן מפורש:

$$\mathbb{I}_n=\begin{pmatrix}1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1\end{pmatrix}$$

מטריצה זו מקיימת $\hat{A}\cdot\mathbb{I}_m=\mathbb{I}_n\cdot\hat{A}=\hat{A}$ לכל מטריצה $\hat{A}$ מסדר $n\times m$.

הערה: לעיתים סדר מטריצת היחידה אינו משנה או טריוויאלי, ולכן המטריצה מסומנת רק על ידי $\mathbb{I}$ ללא ציון המימד.

- <u>מטריצה הופכית:</u>

למטריצות ריבועיות (מטריצות עם מספר זהה של שורות ועמודות; מסדר $n\times n$) ייתכן ויש מטריצה הופכית $\hat{A}^{-1}$ שמקיימת את הקשר:

$$\hat{A}\cdot\hat{A}^{-1}=\hat{A}^{-1}\cdot\hat{A}=\mathbb{I}_n$$

דוגמא: $\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}=\mathbb{I}_2$. לכן , במקרה הזה $\hat{A}=\hat{A}^{-1}=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$.

<u>מערכת משוואות לינאריות:</u>

מערכת משוואות לינאריות מוצגת באופן כללי באופן הבא:

$$\begin{matrix}A_{11}x_1+A_{12}x_2+ & \dots & A_{1n}x_n=b_1\\ \vdots & \ddots & \vdots\quad\ \vdots\\ A_{m1}x_1+A_{m2}x_{2+} & \cdots & A_{mn}x_n=b_m\end{matrix}$$

נשים לב כי מערכת משוואות לינארית ניתנת לייצוג באופן קומפקטי על ידי הפרדה בין רשימת המשתנים, המקדמים של משתנה, והאיבר החופשי, באופן הבא:

$$\hat{A}\,\vec{x} = \vec{b} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

מטריצה $\hat{A} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}$ הינה מטריצת המקדמים מסדר $n \times m$, כאשר n הוא מספר המשתנים, ו-m הוא מספר המשוואות במערכת.

וקטור $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, הינו וקטור עמודה (לעיתים גם מסומן על ידי $(x_1, x_2 \dots x_n)^T$), המייצג את וקטור המשתנים.

וקטור $\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$, הינו וקטור עמודה, שאיבריו הם האיבר החופשי.

הפתרונות של מערכת המשוואות הליניארית, $\hat{A}\,\vec{x} = \vec{b}$, באם הם קיימים ויחידים, נתונים ע"י $\vec{x} = \hat{A}^{-1}\vec{b}$.

<u>מכפלה פנימית, נורמה, אורתוגונליות</u>

מרחב מכפלה פנימית, מוגדר ע"י מרחב וקטורי V (המוגדר על גבי שדה $\mathbb{F}$) ועל ידי פעולת "מכפלה פנימית". מכפלה פנימית, הנקראת לעיתים רק מכפלה, הינה בעצם פונקציה המקבלת שני וקטורים ממרחב וקטורי V ומחזירה סקלר (=מספר) בשדה $\mathbb{F}$. מכפלה זו, מסומנת בדרך כלל ע"י $\langle\cdot,\cdot\rangle$ (או ע"י $\langle\cdot \mid \cdot\rangle : V \times V \to \mathbb{F}$), חייבת לקיים מספר תכונות.

לכל $\vec{v}, \vec{u}, \vec{w} \in V$ (כל שלושה וקטורים במרחב הוקטורי V), ולכל $\lambda \in \mathbb{F}$ (סקלר בשדה $\mathbb{F}$):

- $\langle \vec{v} + \vec{u}, \vec{w} \rangle = \langle \vec{v}, \vec{w} \rangle + \langle \vec{u}, \vec{w} \rangle$
- $\langle \lambda\vec{v}, \vec{u} \rangle = \lambda \langle \vec{v}, \vec{u} \rangle$
- $\overline{\langle \vec{v}, \vec{u} \rangle} = \langle \vec{u}, \vec{v} \rangle$
- $\langle \vec{v}, \vec{v} \rangle \geq 0$

ההגדרה עצמה של המכפלה משתנה כתלות במרחב הוקטורי הנתון. לדוגמא:

א. מכפלה סקלרית על מרחב הווקטורים הגיאומטריים:

נתון $\vec{u}, \vec{v} \in \mathbb{C}^n$ וקטורים גיאומטריים מסדר n מעל שדה המספרים המרוכבים. מכפלה פנימית בין שני וקטורים אלו, נקראת גם מכפלה סקלרית, מוגדרת ע"י

$$\langle \vec{v}, \vec{u} \rangle = \vec{v}^T \cdot \vec{u} = \sum_{i=1}^{n} \bar{v}_i u_i$$

כאשר $\bar{v}_i$ הינו הצמוד המרוכב של v_i.

ב. מרחב הילברט – מרחב מכפלה פנימית על מרחב הפונקציות:

נניח שני פונקציות מרוכבות $f : \mathbb{C} \to \mathbb{C}$ אינטגרביליות בתחום כלשהו I (כמו שהוזכר לעיל, גם מרחב הפונקציות הוא מרחב וקטורי). אז המכפלה פנימית מוגדרת על ידי

$$\langle f(x), g(x) \rangle = \int_I dx f^*(x) g(x)$$

כאשר $f^*(x)$ הינו הצמוד המרוכב של $f(x)$

ניתן להגדיר גם מרחבי מכפלה פנימית נוספים, נניח עבור מרחב המטריצות.

נורמה:

נורמה, מוגדרת על ידי מכפלה פנימית של וקטור בעצמו, ומסומנת ע"י $||\cdot||$, זאת אומרת:

$$||\vec{v}|| = \sqrt{\langle \vec{v}, \vec{v} \rangle} \geq 0$$

שוויון מתקיים אך ורק עבור וקטור האפס; $||\vec{v}|| = 0 \Leftrightarrow \vec{v} = 0$.

תכונה נוספת, נקראת אי-שיוון קושי-שוורץ (Cauchy–Schwarz inequality) או אי-שיוון המשולש, מתוארת ע"י

$$||\vec{v} + \vec{u}|| \leq ||\vec{v}|| + ||\vec{u}||$$

דוגמא:

א. במרחב הווקטורים הגיאומטריים, הגדרת הנורמה היא בעצם הגדרת אורך (או גודל הוקטור). נניח עבור הווקטורים הגיאומטריים התלת-מימדים, $V = \mathbb{R}^3$, אז עבור $\vec{v} = (x, y, z) \in V$, הנורמה מוגדרת ע"י $||\vec{v}|| = \sqrt{x^2 + y^2 + z^2}$.

ב. במרחב הילברט נורמה של פונקציה $f: \mathbb{C} \to \mathbb{C}$ הינה $||f||^2 = \int_I dx |f(x)|^2$.

אורתוגונליות

הגדרת מכפלה פנימית מאפשרת לנו להגדיר אורתוגונליות (או אנכיות) של שני וקטורים במרחב מכפלה פנימית מסוים. שני וקטורים $\vec{v}, \vec{u} \in V$ נקראים אורתוגונליים זה לזה אם ורק אם המכפלה הפנימית שלהם הינה אפס:

$$\vec{v} \perp \vec{u} \;\Leftrightarrow\; \langle \vec{v}, \vec{u} \rangle = 0$$

כאשר מתייחסים למרחב הווקטורים הגיאומטריים, קל להבין את מושג האורתוגונליות.

אורתוגונליות היא הכללה של תכונת הניצבות המוכרת מגאומטריה. בגאומטריה, שני ישרים במישור האוקלידי ניצבים זה לזה אם הזווית הנוצרת בנקודת החיתוך שלהם היא זווית ישרה (בת 90 מעלות). מושג האורתוגונליות מכליל כונה זו גם למרחבים ווקטורים n-מימדים. על מנת להכליל את מושג הניצבות יש ראשית להגדיר זווית בין שני וקטורים:

$$\cos(\theta) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

כאשר $\langle x, y \rangle$ הינה המכפלה הפנימית בין שני הווקטורים, המוגדרת מעל הטבעיים כך: $\langle x, y \rangle = \sum_i x_i \cdot y_i$, והביטוי במכנה $\|x\| \cdot \|y\|$ הוא מכפלת הנורמות. לפי אי שיוויון קושי שוורץ ניתן להוכיח שהביטוי באגף ימין תמיד קטן או שווה בערכו המוחלט ל-1, ולכן לפי ההגדרה הזו תמיד ניתן לחשב זווית בין שני וקטורים.

לווקטורים אורתוגונליים חשיבות רבה כאשר חוקרים מרחבים וקטורים. לבסיס של מרחב וקטורי יש מספר תכונות נוחות כאשר הוא אורתונורמלי (כל אבריו אורתוגונליים זה לזה ובעלי אורך 1). יתר על כן, מתברר שבהינתן בסיס כלשהו למרחב וקטורי ניתן לקבל ממנו בסיס חדש שכל אבריו אורתוגונליים זה לזה, כך שתמיד ניתן למצוא בסיס נוח שכזה. דבר זה נעשה על ידי תהליך גרם-שמידט.

שני וקטורים אורתוגונליים יסומנו על ידי $\perp$. עבור וקטורים אורתוגונליים מתקיימות התכונות הבאות:

- אם $u \perp v$, אז $v \perp u$.
- אם $u \perp v$, אז לכל סקלר λ גם $\lambda u \perp v$.
- אם $u \perp v$ וגם $w \perp v$, אז $(w + u) \perp v$.
- אם וקטור אורתוגונלי לקבוצה של וקטורים אזי הוא גם אורתוגונלי לכל צירוף לינארי שלהם (נובע משתי התכונות הקודמות).

וקטורים עצמיים וערכים עצמיים

תהי $A \in \mathbb{F}^{n \times n}$ מטריצה ריבועית, וקטור $v \in \mathbb{F}^n$ ו-$\lambda \in \mathbb{F}$ סקלר. λ נקרא ערך עצמי של A ו-$v \neq 0$ נקרא הוקטור העצמי המתאים אם מתקיים:

$$A \cdot v = \lambda \cdot v$$

ניתן להראות שעבור מטריצה A, הוקטורים העצמיים המתאימים לסקלר λ הם כל פתרונות המשוואה ההומוגנית $(A-\lambda\mathbb{I}_n)v=0$.

אם נסמן $V=[v_1,\dots,v_n]$, ו-$\Lambda=[\lambda_1,\dots,\lambda_n]$, אזי מתקיים:

$$A = V\,\mathrm{diag}(\Lambda)\,V^{-1}$$

כאשר $\mathrm{diag}(\Lambda)$ הוא ערכי האלכסון של המטריצה Λ.

1.2.2 Calculus

<u>פונקציה</u>

פונקציה הינה התאמה (או העתקה), המתאימה לכל איבר x (בתחום מסוים), ערך יחיד y, ומסומנת באופן הבא: $y=f(x)$. קבוצת הx-ים, נקראת *תחום*, וקבוצת הy -ים נקראת *טווח*. קבוצות התחום והטווח יכולות להיות רציפות (למשל מספרים ממשיים חיובים) או בדידות (למשל קבוצה $\{0,1\}$). בדרך כלל הסימון מופיע כך: $f:X\to Y$, כאשר X וY הינם התחום והטווח בהתאמה.

דוגמא: $||\cdot||:\mathbb{R}^2\to\mathbb{R}^+$, הינה פונקציה, הלוקחת וקטורים גיאומטריים דו-מימדים, ומחזירה מספר ממשי אי שלילי. הפונקציה עצמה $||\cdot||$ היא הנורמה של הוקטור, כפי שהוגדרה בפרק הקודם.

<u>נגזרת</u>

עבור פונקציות ממשיות, נגזרת מוגדרת על ידי מידת השתנות של הפונקציה $f(x)$ על ידי שינוי קטן (אינפיניטסימלי) בx. באופן גיאומטרי, הנגזרת הינה השיפוע של הפונקציה בנקודה x. נגזרת מסומנת בדרך כלל ע"י $f'(x)=\frac{df}{dx}$.

נגזרות של פונקציות אלמנטריות ניתן לחשב באמצעות כללים ידועים. לדוגמא:

- לכל $n\neq 0$ מתקיים: $\frac{d(x^n)}{dx}=nx^{n-1}$.
- חיבור או חיסור פונקציות: $\frac{d(f(x)+g(x))}{dx}=\frac{df(x)}{dx}+\frac{dg(x)}{dx}$.
- מכפלת שתי פונקציות: $\frac{d(f(x)\cdot g(x))}{dx}=\frac{f(x)dg(x)}{dx}+\frac{g(x)df(x)}{dx}$.
- כלל שרשרת: $\frac{df(g(x))}{dx}=\frac{df}{dg}\frac{dg}{dx}$.

כיוון שנגזרת של פונקציה ממשית מכמתת את קצב שינוי הפונקציה, אז בתחום שבו הפונקציה יורדת הנגזרת שם תהיה שלילית, ובתחום שבו היא עולה הנגזרת תהיה חיובית. ככל שקצב ההשתנות גדול יותר כך ערכה המוחלט של הנגזרת גדל.

הערה: לא לכל פונקציה מוגדרת נגזרת. לספר זה נניח שהפונקציה אנליטית ולכן גזירה.

הערה נוספת: כיון שנגזרת של פונקציה היא גם פונקציה, ניתן גם להגדיר נגזרת שניה או נגזרת מסדרים גבוהים יותר. בדרך כלל הסימון הינו $f''(x)=f^{(2)}(x)=\frac{d^nf}{dx^n}$ לנגזרת מסדר שני וכולי

<u>נקודות אקסטרמום</u>

נקודות אקסטרמום של פונקציה, הן נקודות שבהם הפונקציה מקבלת ערך מקסימום או מינימום באופן מקומי. בנקודות אלו, הנגזרת של הפונקציה "משנה כיוון" (מפונקציה עולה לפונקציה יורדת או להפך) ולכן מקבלת את הערך אפס. יש לשים לב שהתאפסות הנגזרת בנקודות המינימום והמקסימום היא תנאי הכרחי אך לא מספיק. ייתכן שהנגזרת מתאפסת בנקודה מסוימת, אך נקודה זו אינה מינימום או מקסימום מקומי, אלא נקודת פיתול.

לדוגמא: $f(x)=x^3$. נגזרת הפונקציה הינה $f'(x)=3x^2$ והיא מתאפסת בנקודה $x=0$.

<u>גרדיאנט, יעקוביאן והסיאן</u>

עבור פונקציה מרובת משתנים, נגזרת חלקית מוגדרת להיות הנגזרת של הפונקציה לפי אחד המשתנים שלה, והיא מסומנת ב-$\frac{\partial f(x_1,\dots,x_n)}{\partial x_i}$. כאשר גוזרים לפי משתנה מסוים, שאר המשתנים הם קבועים ביחס לנגזרת. בהינתן הפונקציה $f(x_1,\dots,x_n)$, וקטור הנגזרות לפי כל המשתנים נקרא גרדיאנט:

$$\nabla f(x_1,\dots,x_n)=\begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \vdots\\ \frac{\partial f}{\partial x_n}\end{bmatrix}\leftrightarrow [\nabla f]_i=\frac{\partial f}{\partial x_i}$$

עבור m פונקציות התלויות ב-n משתנים, היעקוביאן הוא מטריצת הנגזרות החלקיות:

$$\mathcal{J}_f=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix}_{n\times m}\leftrightarrow [\mathcal{J}_f]_{ij}=\frac{\partial f_i}{\partial x_j}$$

עבור פונקציה $f(x_1,\dots,x_n)$, מטריצת הנגזרות מסדר שני נקראת הסיאן:

$$\mathcal{H}_f=\nabla^2 f=\begin{bmatrix}\frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial x_n\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2}\end{bmatrix}_{n\times n}\leftrightarrow [\mathcal{H}_f]_{ij}=\frac{\partial^2 f}{\partial x_i\partial x_j}$$

1.2.3 Probability

תורת ההסתברות היא תחום המספק כלי ניתוח למאורעות המכילים מימד של אקראיות ואינם דטרמינסטיים. הסתברות של מאורע הוא ערך מספרי למידת הסבירות שהוא יתרחש, כאשר ערך זה נע בין 0 ל-1 – מאורע בלתי אפשרי הוא בעל הסתברות 0, ומאורע ודאי הוא בעל הסתברות 1.

<u>הגדרות בסיסיות</u>

Ω = מרחב המדגם – מכלול האפשריות השונות של ניסוי. לדוגמא עבור הטלת קוביה: $\Omega=\{1,2,3,4,5,6\}$.

קבוצה – חלק ממרחב המדגם. לדוגמא עבור הטלת קוביה: $A=\{2,4,6\}=even\ number$.

מאורע – תוצאה אפשרית של ניסוי.

הסתברות – סיכוי של מאורע להתרחש. עבור תת קבוצה A של מרחב המדגם Ω, ההסתברות לקיום מאורע מקבוצת A שווה לחלק היחסי של מספר איברי הקבוצה מתוך קבוצת המדגם:

$$p(A)=\frac{\#A}{\#\Omega},0\le p(A)\le 1$$

$A\cup B$ = איחוד – איחוד של שתי קבוצות הוא אוסף האברים של שתי הקבוצות. איחוד של הקבוצות A ו-B הוא אוסף האיברים המופיעים לפחות באחת משתי הקבוצות A או B. לדוגמא עבור הטלת קוביה:

$$A=\text{even number}=\{2,4,6\},B=\text{lower than }4=\{1,2,3\}$$

$$\rightarrow A\cup B=\{1,2,3,4,6\},\qquad p(A\cup B)=\frac{5}{6}$$

$A\cap B$ = חיתוך – חיתוך של שתי קבוצות הוא אוסף האיברים המופיעים בשתי הקבוצות. חיתוך של הקבוצות A ו-B הוא אוסף האיברים המופיעים גם ב-A וגם ב-B. עבור הדוגמא הקודמת:

$$A\cap B=\{2\},p(A\cup B)=\frac{1}{6}$$

מאורעות זרים – מאורעות שהחיתוך שלהם ריק, כלומר אין להם איברים משותפים:

$$A\cap B=\emptyset,p(A\cap B)=0$$

מאורע משלים – מאורע המכיל את כל האיברים שאינם נמצאים בקבוצה מסוימת:

$$A \cup A^c = \Omega \to p(A \cup A^c) = 1, p(A) = 1 - p(A^c)$$

מאורעות בלתי תלויים: $P(A \cap B) = P(A) \cdot P(B)$. באופן אינטואיטיבי ניתן לחשוב על כך שבמקרה כזה ידיעת האחד אינה משפיעה על הסיכוי של השני.

אם המאורעות זרים (והם בעלי סיכוי שונה מ-0), הם בהכרח תלויים:

$$P(A \cap B) = 0 \neq P(A) \cdot P(B) > 0$$

$p(A|B)$ = הסתברות מותנית – בהינתן מידע מסוים, מה ההסתברות של מאורע כלשהו:

$$p(A|B) = \frac{p(A \cup B)}{p(B)} \leftrightarrow p(A|B) \cdot p(B) = p(A \cup B) = p(B|A) \cdot p(A)$$

בעזרת ההגדרה של הסתברות מותנית ניתן לתת הגדרה נוספת למאורעות בלתי תלויים:

$$A, B \text{ בלתי תלויים } \leftrightarrow p(A|B) = p(A)$$

נשים לב שהמשמעות של שתי ההגדרות זהה – המידע על B לא משנה את חישוב ההסתברות של A.

<u>נוסחת ההסתברות השלמה וחוק בייס</u>

נוסחת ההסתברות השלמה היא נוסחה פשוטה המאפשרת לחשב מאורעות מסובכים. ניתן לפרק מרחב הסתברות לאיברים זרים, ואז לחשב את ההסתברות של כל איבר בפני עצמו. אם ניקח את כל ההסתברויות המתקבלות, ונכפיל כל אחת מהן במשקל של אותו איבר, נקבל את נוסחת ההסתברות השלמה:

$$P(B) = \sum_i P(B|A_i) \cdot P(A_i)$$

מתוך נוסחה זו מגיעים בקלות לחוק בייס, המאפשרת לחשב הסתברות מותנית באמצעות ההתניה ההפוכה:

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

<u>משפט ההכלה וההדחה:</u>

כדי לספור עצמים בקבוצה, אפשר לכלול ולהוציא את אותו עצם שוב ושוב, כל עוד בסוף ההליך נספר כל עצם פעם אחת. עקרון פשוט זה מתורגם לנוסחה הבאה:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{1 \le i < j \le n} |A_i \cap A_j| + \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n-1} |A_1 \cap \ldots \cap A_j|$$

עבור 2 קבוצות הנוסחה נהיית יותר פשוטה:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

במקרה זה, כאשר A, B זרות, אז $|A \cap B| = 0$.

עבור שלוש קבוצות מתקבלת הנוסחה:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|$$

<u>משתנים אקראיים</u>

$X: \Omega \to \mathbb{R}$ משתנה מקרי – פונקציה המתאימה לכל מאורע השייך למרחב ההסתברות ערך מספרי, המהווה את הסיכוי של המאורע להתרחש.

פונקציית ההסתברות של משתנה מקרי X נותנת את הסיכוי של כל x אפשרי:

$$f_X: \mathbb{R} \to [0,1] = p(X = x)$$

פונקציה זו מקיימת שלוש אקסיומות:

- הסתברות של כל מאורע במרחב המדגם גדולה או שווה ל-0.
- סכום ההסתברויות של כל המאורעות במרחב שווה ל-1: $\sum p(X = x) = p(\Omega) = 1$.
- סכום ההסתברויות של שני מאורעות זרים שווה להסתברות של איחוד המאורעות.

עבור משתנה מקרי רציף יש אינסוף מאורעות אפשריים, לכן ההסתברות של כל מאורע יחיד היא 0. לכן עבור משתנה מקרי רציף מכלילים את פונקציית ההסתברות לפונקציה הנקראת פונקציית ההתפלגות (או פונקציית הצפיפות המצטברת), המחשבת את ההסתברות שמאורע יהיה קטן מערך מסוים:

$$F_X(a) = p(X \leq a) = \int_{-\infty}^{a} f_X(x)dx$$

ניתן לחשב בעזרת פונקציה זו את ההסתברות שמאורע יהיה בטווח מסוים:

$$p(a \leq X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x)dx$$

פונקציית ההתפלגות מקיימת את התכונות הבאות:

- $\lim_{a \to -\infty} F_X(a) = 0$.
- $\lim_{a \to \infty} F_X(a) = 1$.
- $\int_{-\infty}^{\infty} f_x(x)dx = 1$.
- הפונקציה מונוטונית עולה במובן החלש: לכל $a \leq b$ מתקיים $F_X(a) \leq F_X(b)$.
- $p(X \geq a) = 1 - F_X(a)$

<u>תכונות ופרמטרים עבור משתנה מקרי:</u>

תוחלת – ממוצע משוקלל של כל הערכים האפשריים, כל אחד מוכפל בהסתברות שלו:

$$\mathbb{E}[X] = \sum_i x_i P(X = x_i) = \int_{-\infty}^{\infty} x f(x)dx$$

תכונות:

- $\mathbb{E}[c] = c$.
- $\mathbb{E}\big[\mathbb{E}[X]\big] = \mathbb{E}[X]$.
- לינאריות התוחלת: $\mathbb{E}[aX + b],\ \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

שונות – מדד פיזור הערכים ביחס לממוצע המשוקלל (-התוחלת):

$$Var[x] = E[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - (\mathbb{E}[X])^2$$

סטיית תקן מוגדרת להיות שורש השונות: $\sigma = \sqrt{Var[X]}$.

תכונות:

- אי שליליות: $Var[x] \geq 0$.
- $Var\ [aX + b] = a^2 V[X]$.

שונות משותפת – מדד ליחס אפשרי בין שני משתנים מקריים:

$$cov(X, Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

כאשר: $\mathbb{E}[X \cdot Y] = \sum_j \sum_i x_i y_j P(X = x_i \cap Y = y_j)$. אם המשתנים בלתי תלויים אז מתקיים $\mathbb{E}[X \cdot Y] = 0$.

מקדם המתאם – נרמול של השונות המשותפת: $\rho(X,Y)=\frac{cov(X,Y)}{\sqrt{V(X)V(Y)}}$. המקדם מקיים: $|\rho|\le 1$.

שני משתנים מקריים מוגדרים בלתי מתואמים אם $cov(X,Y)=0$. אם המשתנים בלתי תלויים אז הם בהכרח בלתי מתואמים.

בעזרת השונות המשותפת ניתן לכתוב: $V[X+Y]=V[X]+V[Y]+2\cdot cov(X,Y)$.

פונקציה יוצרת מומנטים (התמרת לפלס של פונקציית הצפיפות):

$$M_X(t)=\mathbb{E}[e^{tX}]=\begin{cases}\sum_n e^{t\cdot a}p_X(a)\\ \int_s e^{t\cdot a}f_X(x)dx\end{cases}$$

בעזרת פונקציה זו ניתן ליצור מומנטים, שמסייעים ללמוד על המשתנים:

$$\frac{\partial^n M_X(t)}{\partial t^n}_{t=0}=\mathbb{E}[X^n]$$

המומנט הראשון הוא התוחלת והמומנט השני הוא השונות.

<u>התפלגויות מיוחדות (בדיד):</u>

ישנן כל מיני התפלגויות מיוחדות, שמופיעות בטבע בכל מיני מקרים ויש להן נוסחאות ידועות.

<u>התפלגות ברנולי:</u> $X\sim Ber(p)$

ניסוי בעל שתי תוצאות אפשריות "הצלחה" או "כישלון". המשתנה המקרי מקבל שני ערכים בלבד – 0 או 1, בהתאם להצלחה וכישלון.

$$P(X=k)=\begin{cases}1,\mathrm{p}\\0,q\end{cases},\mathbb{E}[X]=p,V[X]=pq=p(1-p)$$

<u>התפלגות בינומית:</u> $X\sim B(n,p)$

בהתפלגות בינומית חוזרים על אותו ניסוי ברנולי n פעמים באופן בלתי תלוי זה בזה. מגדירים את X להיות מספר ההצלחות שהתקבלו בסה"כ. נסמן ב-p סיכוי להצלחה בניסוי בודד וב-q סיכוי לכישלון בניסוי בודד.

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\mathbb{E}[X]=np,Var[X]=npq$$

צריך לוודא 3 דברים: 1) חוזרים על אותו ניסוי באופן בלתי תלוי. 2) חוזרים על הניסוי n פעמים. 3) X מוגדר כמספר ההצלחות המתקבלות בסה"כ.

<u>התפלגות גיאומטרית:</u> $X\sim G(p)$

חוזרים על ניסוי ברנולי. כאשר X מבטא את מספר הניסויים שבוצעו עד ההצלחה הראשונה. p מסמן את הסתברות ההצלחה בניסוי בודד.

$$P(X=k)=pq^{k-1},\mathbb{E}[X]=\frac{1}{p},Var[X]=\frac{q}{p^2}$$

להתפלגות זו יש שתי תכונות נוספות מיוחדות:

1) "תכונת חוסר זיכרון": $P[X=(n+k)|X>k]=P(n)$.

2) ההסתברות שיעברו k ניסויים ללא הצלחה: $P(X>k)=q^k$.

כמו כן, אם מעוניינים לדעת את מספר הניסיונות הממוצע הנדרש עד להצלחה ראשונה – יש לחשב את התוחלת של המשתנה המקרי X.

התפלגות אחידה: $X \sim U[a,b]$

בהתפלגות זו לכל תוצאה יש את אותה הסתברות. הערכים המתקבלים בהתפלגויות החל מ-a ועד b הינם בקפיצות של יחידה אחת (לדוגמא הגרלה של מספר שלם בין 1-100):

$$P(X=k)=\frac{1}{b-a+1}, k=a,a+1,\dots,b\ , \mathbb{E}[X]=\frac{a+b}{2}, Var[X]=\frac{(b-a+1)^2-1}{12}$$

התפלגות פואסונית: $X \sim poi(\lambda)$

התפלגות זאת מתאפיינת במספר אירועים ליחידת זמן כאשר λ הוא פרמטר המייצג את קצב האירועים ליחידת זמן הנבחרת.

$$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}, k=1\dots\infty, \mathbb{E}[X]=Var[X]=\lambda$$

יש לשים לב שכאן ההתפלגות נמדדת ליחידת זמן.

התפלגות היפר גאומטרית: $X \sim H(N,D,n)$

נתונה אוכלוסייה שמכילה N פריטים סה"כ, מתוכה D פריטים "מיוחדים" בעלי תכונה מסוימת. בוחרים מאותה אוכלוסייה n פריטים ללא החזרה. מגדירים את X להיות מספר הפריטים ה-"מיוחדים" שנדגמו.

$$P(X=k)=\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}, \mathbb{E}[X]=\frac{nD}{N}, Var[X]=\frac{nD}{N}(1-\frac{D}{N})\frac{N-n}{N-1}$$

התפלגות בינומית שלילית: $X \sim NB(r,p)$

חוזרים על אותו ניסוי ברנולי בזה אחר זה באופן בלתי תלוי עד אשר מצליחים בפעם ה-r. כלומר, מבצעים את הניסוי עד שמצליחים r פעמים. מגדירים את X להיות מספר החזרות עד שהתקבלו r הצלחות.

$$P(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r}, k=r,r+1,\dots,\infty\ \mathbb{E}[X]=\frac{r}{p}, Var[X]=\frac{r(1-p)}{p^2}$$

<u>התפלגויות מיוחדות (בדיד):</u>

התפלגות מעריכית: $X \sim exp(\lambda)$

התפלגות רציפה המאפיינת את הזמן עד להתרחשות מאורע מסוים. λ הוא ממוצע מספר האירועים המתרחשים ביחידת זמן (אותו פרמטר מההתפלגות הפואסונית). $X \sim exp(\lambda)$, $\lambda < 0$.

גם בהתפלגות זו יש את תכונות חוסר הזיכרון: $P(X>(a+b)|X>a]=P(x>b)$.

התפלגות אחידה: $X \sim U(a,b)$.

זו התפלגות שפונקציית הצפיפות שלה קבועה בין a ל-b.

פונקציית הצפיפות: $f(x)=\frac{1}{b-a}, a<x<b$. פונקציית ההתפלגות המצטברת: $F(t)=\frac{t-a}{b-a}$.

$$\mathbb{E}[X]=\frac{a+b}{2}, Var[X]=\frac{(b-a)^2}{12}$$

התפלגות נורמלית: $X \sim \mathcal{N}(\mu,\sigma^2)$

התפלגות נורמלית היא התפלגות חשובה מאוד כיוון שהיא מופיעה בהמון מקרים. פונקציית הצפיפות של ההתפלגות הנורמלית נראית כמו פעמון, כאשר לעקומה קוראים גם עקומת גאוס. ההתפלגויות הנורמליות נבדלות אחת מהשנייה באמצעות הממוצע וסטיית התקן (-הפרמטרים שמאפיינים את ההתפלגות). התפלגות נורמלית סטנדרטית היא התפלגות נורמלית בעלת תוחלת 0 ושונות 1:

$$X \sim \mathcal{N}(0,1^2)$$

עבור תוחלת ושונות μ, σ, פונקציית הצפיפות של משתנה נורמלי הינה:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ניתן להשתמש במומנטים בכדי למצוא קשרים בין התפלגויות. למשל עבור שני משתנים המתפלגים נורמלית:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2), Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

המומנטים מקיימים:

$$M_X(t) \cdot M_Y(t) = e^{\mu_x t + \frac{1}{2}\sigma_x^2 t^2} \cdot e^{\mu_y t + \frac{1}{2}\sigma_y^2 t^2} = e^{(\mu_x + \mu_y)t + \frac{1}{2}(\sigma_x^2 + \sigma_y^2)t^2} = M_{X+Y}(t)$$

ולכן ניתן לחשב את ההתפלגות של $X + Y$:

$$X + Y \sim \mathcal{N}\left(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2\right)$$

<u>אי שיוויוניים:</u>

<u>מרקוב:</u>

בהינתן $X \geq 0$, התוחלת $\mathbb{E}[X]$, עבור פרמטר $a > 0$ מתקיים:

$$p(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

<u>צ'בישב:</u>

בהינתן התוחלת $\mathbb{E}[X]$ והשונות $Var[X]$, עבור פרמטר $a > 0$ מתקיים:

$$p(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var[X]}{a^2}$$

<u>צ'רנוף:</u>

בהינתן התוחלת $\mathbb{E}[X]$, עבור שני פרמטרים $a, t > 0$ מתקיים:

$$p(X \geq a) \leq \frac{\mathbb{E}[e^{tx}]}{e^{ta}}$$

<u>ינסן:</u>

עבור משתנה מקרי X בעל תוחלת, עבור פונקציה קמורה $g: \mathbb{R} \to \mathbb{R}$ מתקיים:

$$g(E[x]) \leq \mathbb{E}[g(x)]$$

<u>התפלגות דו ממדית:</u>

$$F_{x,y}(a, b) = P(x \leq a, y \leq b)$$

תכונות:

$$\lim_{a,b \to \infty} F_{x,y}(a, b) = 1$$

$$\lim_{a \to -\infty} F_{x,y}(a, b) = \lim_{b \to -\infty} F_{x,y}(a, b) = 0$$

$$P(c < x < a, d < y < b) = P(x < a, y < b) - P(x < a, y < d) - P(x < c, y < b) + P(x < c, y < d)$$
$$= F_{x,y}(a, b) - F_{x,y}(a, d) - F_{x,y}(c, b) + F_{x,y}(c, d)$$

אם x, y בלתי תלויים אז מתקיים:

$$\forall a,b\ F_{x,y}(a,b) = F_x(a)\cdot F_y(b)$$

זוג משתנים נקרא דו-מימדי רציף אם קיימת פונקציית צפיפות דו-מימדית $f_{x,y}(s,t)$, כך שמתקיים:

$$P(x,y\in A) = \int f_{x,y}(s,t)dsdt$$

באופן שקול מתקיים:

$$f_{x,y}(s,t) = \frac{\partial^2}{\partial s\partial t}F_{x,y}(s,t) = \frac{\partial^2}{\partial t\partial s}F_{x,y}(s,t)$$

התפלגות שולית:

$$F_x(s) = P(x\le s) = P(x\le s, y\le\infty) = \int_{-\infty}^{s}\int_{-\infty}^{\infty} f_{x,y}(x,y)dxdy$$

נוסחת ההסתברות השלמה לצפיפות (באופן שקול גם ל-$f_y(t)$):

$$f_x(s) = \frac{d}{ds}F(x_s) = \int_{-\infty}^{\infty} f_{x,y}(s,y)dy$$

כעת ניתן גם לכתוב תנאי שקול למשתנים בלתי תלויים – x,y בלתי תלויים אם מתקיים:

$$\forall \mathrm{x,y}\ f_{x,y}(X,Y) = f_x(X)\cdot f_y(Y)$$

<u>סטטיסטיקה היסקית</u>

אם יודעים את סוג ההתפלגות אבל לא יודעים את מרכיביה, ניתן לאמוד את המרכיבים בעזרת מדגם. המדגם מאפשר לנו להשתמש באומדן עבור מספר מאורעות שווי התפלגות. דוגמא – נניח רוצם למדוד גובה של קבוצה מסוימת – כלל התלמידים בבית ספר מסוים. ידוע שגובה מתפלג נורמלית, אבל לא יודעים כאן את התוחלת והשונות. לשם כך ניתן להשתמש באומד – פונקציה שמנסה לנתח את המאורעות ומתוך כך להסיק את התוחלת והשונות.

בניתוח נצא מנקודת הנחה שידוע כי הערכים במדגם נלקחים כולם מתוך התפלגות X, השייכת למשפחה של התפלגויות שתלויות בפרמטר אחד או יותר שאינם ידועים. (למשל בדוגמה: $X\sim N(\mu,\sigma^2)$ כאשר σ,μ לא ידועים). בפועל נתונות n דגימות בלתי תלויות מתוך ההתפלגות: $X_1,X_2,\dots,X_n$, ורוצים לאמוד את הפרמטרים הלא ידועים (כפונקציה של הערכים שדגמנו).

אומד בלתי מוטה: אומד מוגדר להיות בלתי מוטה אם התוחלת של האומד שווה לפרמטר אותו אנו מנסים לאמוד, כלומר, אם $\mathbb{E}(\hat{\theta}) = \theta$, אז האומד הוא חסר הטיה. במילים אחרות – אומד יהיה חסר הטיה אם התוחלת של המשתנה המקרי המחושב לפי θ שווה ל-θ עבור כל θ.

דוגמאות לאומדים בלתי מוטים:

אומד בלתי מוטה לתוחלת – ממוצע חשבוני:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \mathbb{E}[x_i] = \theta$$

אומד בלתי מוטה לשונות:

$$\mathbb{E}[s^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

הוכחה:

$$\mathbb{E}[s^2] = \mathbb{E}\left[\frac{1}{n-1}\cdot\sum_i (x_i-\bar{x})^2\right] = \frac{1}{n-1}\sum_i \mathbb{E}(x_i-\bar{x})^2$$

$$\frac{1}{n-1}\sum_i \mathbb{E}[((x_i-\mu)-(\bar{x}-\mu))^2]$$

$$\frac{1}{n-1}\sum_i \mathbb{E}[(x_i-\mu)^2-2(x_i-\mu)(\bar{x}-\mu)+(\bar{x}-\mu)^2]$$

$$\frac{1}{n-1}\sum_i \mathbb{E}[(x_i-\mu)^2]-\mathbb{E}[2(x_i-\mu)(\bar{x}-\mu)]+\mathbb{E}[(\bar{x}-\mu)^2]$$

$$\frac{1}{n-1}\sum_i \sigma^2 - 2\,(\frac{1}{n}\sum_j \mathbb{E}\big[(x_i-\mu)(x_j-\mu)\big]+\frac{1}{n^2}\sum_j\sum_k \mathbb{E}[(x_j-\mu)(x_k-\mu)]$$

$$\frac{1}{n-1}\sum_i\left[\sigma^2-\frac{2\sigma^2}{n}+\frac{\sigma^2}{n}\right]$$

$$\frac{1}{n-1}\sum_i\left[\frac{(n-1)\sigma^2}{n}\right]=\frac{n-1}{n(n-1)}\sum_i \sigma^2 \;=\sigma^2 \blacksquare$$

<u>אומד נראות מרבית – Maximum likelihood estimator (MLE):</u>

בהינתן סדרת דגימות מתוך התפלגות עם פרמטר לא ידוע, נגדיר את פונקציית הנראות שלהן כמכפלת ההסתברויות של כל הדגימות, או "הנראות של המדגם":

$$L(x_1,x_2\dots x_n|p(\theta)) = \prod_i P_\theta(x_i)$$

זוהי פונקציה הן של הדגימות והן של הפרמטר.

אם ההתפלגות רציפה מגדירים במקום זאת את פונקציית הנראות להיות מכפלת הצפיפויות:

$$L(x_1,x_2\dots x_n|p(\theta)) = \prod_i f_\theta(x_i)$$

אומדן הנראות המקסימלית הוא פשוט הערך של הפרמטר שממקסם את פונקציית הנראות. כלומר, $\hat{\theta}$ הוא אומד נראות מקסימלי עבור θ אם $\hat{\theta} = arg\max_\theta L(x_1,x_2\dots x_n|p(\theta))$.

מכוון ש-log הינה מונוטונית, למקסם את L שקול למקסם את $\log(L)$ (log likelihood), וזה לרוב יותר קל, מכיוון שהמכפלה הופכת לסכום:

$$\log L\big(x_1,x_2\dots x_n\big|p(\theta)\big) = \sum_{i=1}^{n}\log f_\theta(x_i)$$

נראה מספר דוגמאות לחישוב ה-MLE:

א. מציאת הפרמטר λ בהתפלגות פואסונית:

$$X\sim poi(\lambda)$$

שלב א' – נגדיר את אומדן הנראות – $L=(x_1,x_2\dots x_n|p_\lambda)=\prod_i P_\lambda(x_i)$. בהתפלגות פואסונית מקיימת: $P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$, לכן בעצם צריך למצוא מקסימום לביטוי:

$$\prod_i P_\lambda(x_i) = \prod_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

מסובך למצוא לזה מקסימום, לכן נוציא לוג:

$$\ln\left(\prod_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) = \sum_i \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

$$= \sum_i \ln\left(e^{-\lambda}\lambda^{x_i}\right) - \ln(x_i!) = \sum_i \ln\left(e^{-\lambda}\right) + \ln\left(\lambda^{x_i}\right) - \ln(x_i!)$$

$$= \sum_i x_i \ln(\lambda) - \lambda - \ln\left(x_i!\right)$$

כעת נגזור:

$$\frac{\partial L}{\partial \lambda} = \sum_i \frac{x_i}{\lambda} - 1 = \sum_i \frac{x_i}{\lambda} - \sum_i 1 = \sum_i \frac{x_i}{\lambda} - n$$

וכשנשווה ל-0 נקבל:

$$\sum \frac{x_i}{\lambda} = n$$

נבודד את הפרמטר אותו מנסים לאמוד:

$$\lambda = \frac{\sum x_i}{n}$$

וקיבלנו אומד עבור הפרמטר λ, וכאשר נתון סט התוצאות, פשוט נציב אותן, ונמצא מפורשות את הערך של האומד.

זה בעצם תהליך מציאת ה-MLE. כעת נבדוק האם האומד הוא מוטה או לא, כאשר נשתמש בעובדה שעבור התפלגות פואסונית $\mathbb{E}(x) = \lambda$:

$$\mathbb{E}(\lambda) = \mathbb{E}\left(\frac{\sum_i x_i}{n}\right) = \frac{1}{n}\sum_i \mathbb{E}[x_i] = \frac{n\lambda}{n} = \lambda$$

קיבלנו שתוחלת האומד שווה לפרמטר, ולכן הוא בלתי מוטה.

ב. התפלגות נורמלית:

$$X \sim (\mu, \sigma^2)$$

פה יש שני פרמטרים לאמוד – התוחלת והשונות. ראשית נגדיר את הנראות:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

לכן הנראות תהיה (נשים לב שהמכפלה תעבור לסכום במעריך של האקספוננט):

$$\prod_i f(x) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2}$$

נוציא לוג:

$$\ln(L) = \ln\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\right) + \ln\left(e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}\right)$$

$$= nln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - ln\left(e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}\right)$$

נשים לב שבביטוי הראשון מה שיש בתוך ה– ln זה בעצם $(2\pi)^{-\frac{1}{2}} + (\sigma^2)^{-\frac{1}{2}}$, ואז המעריך יכול לרדת מחוץ ל– ln:

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

כעת בשביל לאמוד את התוחלת יש לגזור לפי μ, וכדי לאמוד את השונות ישKDZUR לפי σ^2:

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_i(x_i-\mu) = \frac{1}{\sigma^2}\sum_i(x_i-\mu)$$

וכשנשווה ל-0 נקבל:

$$\hat{\mu} = \frac{\sum_i x_i}{n}$$

וכנעשה אותו תהליך על השונות נקבל:

$$\hat{\sigma}^2 = \frac{\sum_i(x_i-\hat{\mu})^2}{n}$$

נשים לב שעבור התוחלת קיבלנו שהאומד הוא בעצם הממוצע של המדגם.