פרק ז'

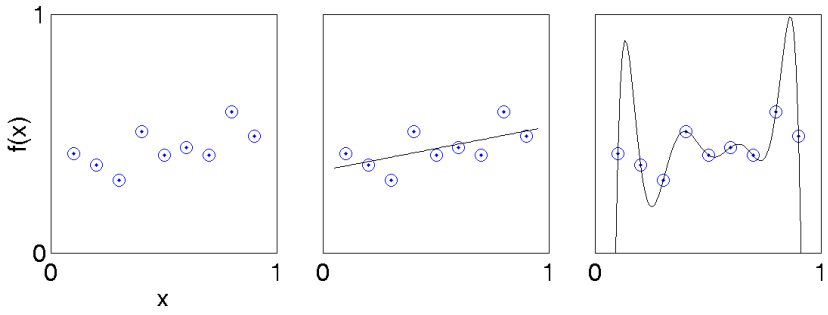
בחירת מודל

בפרקים קודמים עסקנו במקרה שבו, בנוסף לנתונים שאותם רצינו ללמוד, היה לנו מידע מוקדם על הצורה הכללית של התפלגות הנתונים. לרוב, מידע זה היה בצורה של התפלגות פרמטרית, ולנו נותר להעריך את הפרמטרים של ההתפלגות אשר מתאימים לתצפיות הנתונות. פעמים רבות עם זאת, ניתנות לנו תצפיות, אך אין לנו ידע מוקדם על המודל (פרמטרי או אחר). הפרק הנוכחי עוסק בדרכים לבחירה מושכלת של המודל.

7.1 מודלים מורכבים ופשוטים

כשהמידע שבידינו מגיע ממקור רועש (כפי שקורה בדרך כלל בחיים) קשה להגדיר באופן חד משמעי מהו המודל הנכון: מצד אחד ניתן לבנות מודל פשוט שיתאר את הנתונים בקירוב גס. מצד שני ניתן לבנות מודל מפורט ומורכב, התואם בדיוק רב את הנתונים שבידינו, אך היות והם רועשים כנראה אינו תואם את המציאות.

לדוגמא, נניח שנתונות לנו אוסף של n תצפיות שאנחנו מעוניינים להעריך את ההתפלגות שלהן. מודל שנותן הסתברות $1/n$ לכל אחת מהתצפיות במדגם והסתברות אפס לכל ערך אחר, ממקסם את הנראות של המדגם תחת המודל אך משאיר אותנו בתחושה חזקה כי הוא שגוי. מצד שני מודל שמניח התפלגות אחידה על כל טווח הערכים של התצפיות עשוי להיות הפשטה מיותרת שלא מתארת את הנתונים בדיוק מספיק.



דוגמא נפוצה נוספת היא המקרה של קירוב אוסף נקודות על ידי פונקציה פולינומיאלית. האיור שלעיל מראה דגימה של פונקציה בתשע נקודות, ואת הקירוב המתקבל על ידי ישר (פולינום ממעלה אחת, ולכן פשוט יחסית) לעומת הקירוב

המתקבל על ידי פולינום ממעלה 8. (נזכיר כי כל n נקודות ניתן לקרב במדויק על ידי פולינום ממעלה n-1). קל לראות כי הקירוב ממעלה 8, עובר בדיוק דרך כל הנקודות, אך לצורך כך הוא מקבל ערכים קיצוניים בנקודות שאינן נקודות המדגם. קשה יותר להחליט מהי המעלה הנכונה של הפולינום שיתאר את הנקודות בצורה המיטבית.

הדוגמה האחרונה ממחישה מספר מושגים בסיסיים בתורת הלמידה. נשים לב כי למרות שנתון לנו מדגם סופי, אנחנו מעוניינים למעשה למצוא פונקציה שתתאר את ההתפלגות העומדת בבסיס התהליך שיצר את התצפיות, ולא את התצפיות המסויימות שניתנו לנו. לכל מודל שנבנה, נרצה אם כן להבדיל בין השגיאה על התצפיות שנתונות לנו (שגיאה זו נקראת **שגיאת האימון**), לעומת השגיאת הממוצעת על כל התצפיות שעשויות היו להנתן לנו (שגיאה זו נקראת **שגיאת ההכללה**). מסתבר שלעתים קיים יחס חליפין בין היכולת לתאר את התצפיות הנתונות בדיוק רב אל מול היכולת להכליל ולתאר תצפיות שאותן לא ראינו. המקרה בו משיגים שגיאת אימון קטנה על חשבון שגיאת ההכללה קרויה Over-fitting, והיא פוגעת ביכולת ההכללה או החיזוי של המודל, כך שהמודל טועה יותר כשיקבל דוגמאות חדשות מההתפלגות המקורית. ההוכחה של תכונות אלו עומדת במרכזה של תורת הלמידה (ואי אפשר שלא להזכיר כאן את תורת VC Vapnik,95) אך לא נוכל לכסות נושאים אלו כאן. בפרק הנוכחי נעסוק על קצה המזלג בנסיון למצוא את שביל הזהב בין מודלים מסובכים המתארים את המידע הנתון בדיוק רב יותר לבין מודלים פשוטים שלהם יש לעתים יכולת הכללה גדולה יותר.

7.1.1 התער של אוקהם וסיבוכיות קולמוגורוב

הגישה הידועה כסימפליציזם קובעת כי כאשר ניתן להסביר מידע שמגיע מתצפיות במספר דרכים אפשריות - נעדיף תמיד את הדרך הפשוטה ביותר המתארת את המידע. נהוג לומר כי גישה זו הוצעה במאה השתים עשרה על ידי הנזיר הפרנציסקני ויליאם מאוקהם, שטען "Pluralites non est ponenda sine necessitate" (לריבוי אסור להיות "מוצב" אלא מתוך הכרח). גישה זו מכונה לכן התער של אוקהם, ומשמשת לגילוח חלקים מיותרים בתיאוריה.

מספר סיבות עומדות מאחורי גישה זו: ראשית מודל פשוט יותר קל יותר להבנה ולמימוש חישובי. שנית, כפי שרמזנו לעיל, מודלים פשוטים יותר מספקים יכולת הכללה טובה יותר, דהיינו מצליחים לזהות טוב יותר מבנים סטטיסטיים משמעותיים הנמצאים בנתונים, ולהמנע מהתחשבות במבנים סטטיסטים הנובעים מסופיות המדגם שברשותנו.

כיצד אם כן מודדים פשטות של מודל ? מדד חשוב בהקשר זה הוא **סיבוכיות קולמוגורוב** (Kolmogorov Complexity). סיבוכיות קולמוגורוב של מחרוזת מוגדרת כאורך התכנית המינימלית (על מכונת טיורינג אוניברסלית) המתארת את המחרוזת. סיבוכיות קולמוגורוב חסומה מלעיל כמובן על ידי אורך המחרוזת, וכן על ידי כל תכנית אחרת שאנו יכולים לחשוב עליה, שתתאר את המחרוזת. ניתן לקבל חסמים על התוחלת של סיבוכיות קולמוגורוב תוך שימוש בחישובי אנטרופיה, אולם במקרה הכללי סיבוכיות קולמוגורוב אינה ניתנת לחישוב, ובמובן זה אינה קונסטרוקטיבית.

גישה נפוצה למדידת סיבוכיות המודל היא לספור את מספר הפרמטרים החופשיים במודל. לדוגמא, בספרות של רשתות נוירונים מופיעות הצעות רבות ושונות לכימות סיבוכיות רשתות Feed-Forward, החל בכמות הנוירונים בשכבת הביניים, דרך כמות המשקולות הכוללת, וכלה בנורמה של וקטור המשקולות. שיטות אלו מבוססות לעתים קרובות על הוספה של "**איבר עונש**" (**penalty term**) לפונקציית המטרה, כך שהמטרה של אלגוריתמי הלמידה הופכת להיות הקטנת שגיאת האימון תוך כדי שמירה על סיבוכיות נמוכה. לעתים קרובות איבר העונש נבחר באופן שרירותי, או באופן שמקל על המימוש החישובי. לעומת אלו, הסעיף הבא מתאר עקרון מובנה יותר למדידת סיבוכיות המודל כפונקציה של הפרמטרים.

Minimum Description Length (MDL) 7.1.2

אורך תיאור מינימלי

נניח שנתון מדגם $X^{(n)}=\left(x_1,x_2,...,x_n\right)$ ואנו רוצים לתאר אותו על יד מספר מיטבי, k , של פרמטרים $\theta_1,\theta_2,...,\theta_k$. אם הנתונים באים ממקור רועש - הדיוק המרבי שבו אנו יכולים לדעת את ערכו של θ_i יהיה $\delta_i=c_i/\sqrt{n}$ (סדר-גודל של סטית תקן בהתפלגות המדגם של הפרמטר) מספר הביטים הנדרש לתאור בדיוק כזה יהיה $\log\left(1/\delta_i\right)$. כעת, בהינתן המודל - מוגדרת לנו פונקצית התפלגות ואנו יכולים לחשב את הנראות (Likelihood) המותנית

$$L\left(X^{(n)}\mid\theta_1,...,\theta_k\right)=P\left(x_1,...,x_n\mid\theta_1,...,\theta_k\right)$$

לוגריתם הנראות המותנית יקיים, על-פי משפט Shannon-McMillan (AEP) כי

$$\log L\left(X^{(n)}\mid\theta_1,...,\theta_k\right)=$$
$$\log P\left(x_1,...,x_n\mid\theta_1,...,\theta_k\right)\xrightarrow[n\to\infty]{}-n\cdot H\left[P\left(x\mid\theta_1,...,\theta_k\right)\right]+Const$$

ועל-פי משפט קידוד המקור של שנון, אם נמצא סט של פרמטרים $\theta_1,...,\theta_k$ שיביא למקסימום את לוג הנראות (כלומר נביא את האנטרופיה המותנית למינימום) נוכל לתאר את המידגם באמצעות קוד בעל אורך מינימאלי. זאת היות ואורך התיאור הממוצע של המדגם בהינתן המודל עבור קוד אופטימלי, תלוי לינארית באנטרופיה המותנית.

בסך הכל, התיאור הכולל של המדגם יכיל שני מרכיבים: התיאור של המודל האופטימלי (התיאור של סט הפרמטרים $\theta_1,...,\theta_k$ המביאים את הנראות למקסימום ומתוארים בדיוק הרלוונטי), והתיאור של המדגם בהנתן המודל, באמצעות קוד אופטימלי. האורך של התיאור הכולל מורכב לכן משני גורמים: אורך תיאור המודל האופטימלי ואורך תיאור התצפיות בהנתן המודל.

$$DL\left(X^{(n)} \mid \theta_1,...,\theta_k\right) = -\log_2 P\left(X^{(n)} \mid \theta_1,...,\theta_k\right) + \sum_{i=1}^{k} \log_2 \frac{1}{\delta_i}$$
$$= -\log_2 P\left(X^{(n)} \mid \theta_1,...,\theta_k\right) + \frac{k}{2}\log_2 n - \sum_{i=1}^{k} \log_2 (c_i)$$
$$= -\log_2 P\left(X^{(n)} \mid \theta_1,...,\theta_k\right) + \frac{k}{2}\log_2 n + O(k)$$

אנו מוצאים כי כדי להקטין את אורך התיאור הכולל, עלינו להגדיל את הנראות או להקטין את סדר המודל k. ניתן להביא לנראות מקסימלית אם נשתמש ב- k=n פרמטרים ונבחר ב- $\theta_i = x_i$ (לדוגמא, בבעיית החלוקה לצברים אותה פיתחנו בהרחבה בפרק שלוש, נוכל לבחור כל דגימה כמרכז של צבר שמכיל בדיוק נקודת מדגם אחת) אבל במקרה כזה נשלם בהגדלת האבר השני $\frac{k}{2}\log_2 n$. כפי שהסברנו לעיל, ניתן לראות את האיבר השני כ- **איבר עונש** שמעניש אותנו על הגדלה של סדר המודל. איבר זה הוא משמעותי יחסית לאיבר הנראות רק כאשר מדובר במדגמים קטנים יחסית, במדגמים גדולים מספיק איבר הנראות משמעותי הרבה יותר והרגישות לשינויים ב- k היא קטנה. עם זאת, MDL מהווה שיטה קונסיסטנטית לאמידת סדר המודל: ניתן להראות כי כאשר $n \to \infty$ הסדר אותו אומדים יהיה הסדר הנכון.

מהבחינה הפרקטית, על מנת למצוא סדר אופטימלי למודל על פי עקרון MDL, יש לחשב את אורך התיאור האופטימלי עבור ערכים שונים של סדר המודל k, ולבחור במודל שעבורו ה- DL הוא הקצר ביותר.

7.2 עקרון מקסימום האנטרופיה

בפרק ג' עסקנו בבעיות בהן הצורה הכללית של ההתפלגות ידועה, והיה עלינו להעריך את הפרמטרים של ההתפלגות המתאימים לתצפיות הנתונות. קיימים כמובן מקרים בהם אין לנו ידע מוקדם על הצורה של ההתפלגות, אך יש לנו מידע חלקי (למשל ידועה לנו התוחלת) ועלינו לבחור צורה פרמטרית "טובה". עקרון מקסימום האנטרופיה (Janynes 1957, Good 1963), מציע כי ההתפלגות הרצויה היא ההתפלגות שלה יש אנטרופיה מקסימלית מבין כל ההתפלגויות המקיימות את האילוצים הנתונים לנו.

7.2.1 מקסימיזציה של אנטרופיה תחת אילוצים

עקרון מקסימום האנטרופיה נועד לטפל במקרה בו אנחנו מעונינים לבנות מודל פרמטרי של הנתונים אבל אין לנו ידע מוקדם על הצורה הפרמטרית של ההתפלגות. כדי לבנות מודל פשוט ככל האפשר אשר מתאים לנתונים, נבחר מספר תכונות של המדגם (למשל הממוצע) ונבחר התפלגות המקיימת את התכונות (למשל התפלגות שהתוחלת שלה מתאימה לממוצע של המדגם). קיימות כמובן התפלגויות רבות המקיימות את האילוצים, אך עקרון מקסימום האנטרופיה קובע כי עלינו לבחור את ההתפלגות שלה האנטרופיה המקסימלית מבין כל ההתפלגויות המקיימות את האילוצים.

דוגמה

בניסוי אלקטרופיזיולוגי, עוקבים אחר מספר פוטנציאלי הפעולה של תא עצב כתגובה לגירוי. זהו משתנה מקרי X המקבל את הערכים 3,2,1,0, ... בהסתברויות לא ידועות $P(X=1)=p_1, P(X=2)=p_2, ...$. היות וכמות הנתונים שלנו קטנה, אנחנו מנסים לבנות מודל פרמטרי פשוט, אך כמובן עליו לחזות היטב את התפלגות מספר הספייקים בתגובה לגירוי. נחשב את ממוצע מספר הספייקים בתגובה לגירוי (אותו נסמן ב- m_1), וכעת נמצא את ההתפלגות בעלת מקסימום האנטרופיה תחת האילוץ $E(p) = m_1$. נרשום במפורש את הלגרנג'אן שאותו נרצה להביא לנקודת אקסטרמום

$$L(c_0, c_1, p_1, p_2, ...) = H(X) - c_0(\sum_i p_i - 1) - c_1(\sum_i x_i p_i - m_1)$$

וקיבלנו מערכת משוואות שאותה ניתן לפתור נומרית ולמצוא את c_1. כאשר נמצא את c_1 נקבל הצגה מפורשת של ההתפלגות p(X).

נשים לב כי גם במקרה הכללי בו יש לנו n אילוצים שונים $F_i(X) = m_i$ אזי נגזרת הלגרנג'יאן תיתן משוואה מהצורה

$$\log(p_i) = \sum_i c_i F_i(X) + \log(Z),$$

דהיינו ההתפלגות היא בעלת צורה אקספוננציאלית

$$p_i \propto \exp(\sum_i c_i F_i(X)).$$

כאשר האילוצים הם על תוחלות, אזי התחום הוא קמור ולכן לאנטרופיה, שהיא פונקציה קעורה, יש מקסימום יחיד בתחום. במקרה זה קיימים אלגוריתמים איטרטיביים יעילים המבטיחים התכנסות למקסימום הגלובלי.

7.2.2 הצדקות לשימוש בעקרון מקסימום האנטרופיה

שני טיעונים עיקריים עומדים מאחורי השימוש בעקרון מקסימום האנטרופיה. ראשית, האנטרופיה היא מדד לאי ודאות, ומכאן שאם נתונות לנו שתי התפלגויות המקיימות את האילוצים, אז ההתפלגות שלה אנטרופיה נמוכה יותר מכילה לכאורה אילוצים נסתרים שהוכנסו מבלי משים ומקטינים את אי הודאות שבה. נראה אם כן כי מי שמאמין כי אנטרופיה היא המדד הנכון למדידת אי ודאות, יעדיף אנטרופיה מקסימלית.

סוג הטיעונים השני לטובת שימוש בעקרון מקסימום אנטרופיה מסתמך על כך שאוספי תצפיות שלהן אנטרופיה גדולה יותר הם נפוצים יותר (ראה משפט AEP בפרק הקודם), ובהקבלה, התפלגויות שהאנטרופיה שלהן גדולה יותר הן נפוצות יותר. ניסוח בייסיאני לטיעון זה, יקבע כי הסיכוי האפריורי לפגוש התפלגות עולה באופן אקספוננציאלי עם האנטרופיה של ההתפלגות.

Cross Validation 7.3

נסכם את הפרק הנוכחי בתיאור קצר של השיטה הנפוצה ביותר לבחירה של סדר המודל. בניגוד לסעיפים הקודמים בהם תיארנו עקרונות תיאורטיים לבחירת סדר המודל, שיטת ה- cross validation מציעה פתרון אמפירי פשוט. מחלקים את הנתונים לשני חלקים: **סט אימון** (training set), ו**סט בחינה** (testing set). לכל סדר מודל k בונים מודל אופטימלי על סמך התצפיות בסט האימון, ובוחנים את טיב המודל שאימנו על סמך התצפיות בסט המבחן. בסופו של דבר, בוחרים בסדר המודל שנתן את התוצאות הטובות ביותר על סט המבחן. בפרקטיקה קיימות וריאציות רבות לשיטה זו (bootstrap, jack-knife), אך לא נפרט לגביהן כאן.

תרגילים

1. מהי ההתפלגות בעלת האנטרופיה המקסימלית המקיימת
   א. את האילוצים $E(X)=m_1 \;;\; E(X^2)=m_2$
   ב. את האילוץ $E(X)=m_1$ עבור משתנה מקרי X המקבל ערכים חיוביים בלבד ?

2. נתונות מדידות ממקור פולינומי "רועש", כלומר לכל נקודה x_i נקבל ערך מדידה $y_i=\sum_{k=0}^{N} a_k x_i^k + \eta_i$ כאשר a_k הם מקדמי הפולינום, N הוא סדר הפולינום x_i הן הנקודות בהן בוצעה מדידה, y_i הוא ערך המדידה שהתקבל ו- η_i הוא ערך רעש המדידה של הנקודה (נניח כי הרעש בכל מדידה הוא בלתי תלוי והוא מוגרל מהתפלגות גאוסית). נחלק את הנקודות לקבוצת אימון training set ולקבוצת הכללה testing set .

   א. עבור הנקודות שבקובץ האימון, לכל ערך של N (כאשר N=0,1,2,…,5) מצאו את מקדמי הפולינום $\{b_1,b_2,...,b_N\}$ שעבורם $X^2=\sum_{x_i\in tset.data}\left(y_i-\sum_{k=0}^{N} b_k x_i^k\right)^2$ הוא מינימאלי.

   ב. לכל N ולכל פולינום שמצאתם, חשבו את השגיאה הממוצעת עבור הנקודות שבקובץ האימון (שגיאת האימון), המגדרת באופן הבא (|trainset| הוא גודל קבוצת האימון.
   $$E_{training}=\frac{1}{|trainset|}\sum_{x_i\in tset.data}\left(y_i-\sum_{k=0}^{N} b_k x_i^k\right)^2$$

   ג. לכל N ולכל פולינום כנ"ל, חשבו גם את שגיאת ההכללה, כלומר עד כמה מוצלח הפולינום שחושב עבור סט האימון לנקודות של סט ההכללה
   $$E_{generalization}=\frac{1}{|testset|}\sum_{x_i\in tset.data}\left[y_i-\sum_{k=0}^{N} b_k x_i^k\right]^2$$

   ד. מהו אם כן הפולינום (סדר המודל) העדיף, ומדוע?

   ה. נניח כי ידוע שהרעש מפולג גאוסית עם ממוצע $\mu=0$ וסטיית תקן, $\sigma=0.1$. עבור N שונים כנ"ל חשבו את ה- Description Length של הנתונים שבקבוצת ההכללה ומצאו מהו ה- Minimum Description Length ומתוך כך מהו סדר המודל האופטימלי. השוו לתוצאות עבור חישוב דומה לקבוצת האימון.