

# מבוא לבינה מלאכותית – תרגול 1

## נושאים:

- הקדמה – עניינים טכניים
  - הקדמה – על הקורס
  - סטטיסטיקה בייסיאנית, אומדים והסקה סטטיסטית
  - מושגים דרך דוגמה – רגרסיה פולינומיאלית
- 

## הקדמה טכנית:

מייל שלי – [levinai@biu.ac.il](mailto:levinai@biu.ac.il)

ציון בקורס – 100% עבודה סופית

יעלו מספר תרגילים במהלך הסמסטר (ל-math wiki), ללא משקל לציון וללא חובת הגשה. הם יוכלו להיות תאורטיים או תכנותיים, כאשר שפת התכנות בה נשתמש היא python. לא נלמד לתכנת ב-python במסגרת הקורס, לכן מי שלא יודע כדאי שילמד, לפחות ברמה בסיסית.

## הקדמה על הקורס:

הקורס נועד לתת רקע בסיסי לתחומים שמאוגדים תחת השם "בינה מלאכותית".

למה אנחנו מתכוונים במושג "בינה מלאכותית"?

בני האדם פיתחו מחשבים, שהם כלים בעלי יכולות חישוב חזקות ומהירות בהרבה מבני האדם, אבל הם רק מכונות שממלאות פקודות בסיסיות. כדי לנצל את המחשבים לפתרון בעיות מסובכות יותר, היינו רוצים לתכנת מחשבים כדי שיוכלו לפעול בצורה "חכמה", להבין ולקבל את ההחלטה הנכונה ביותר על סמך מידע שמועבר אליהם.

זו משימה קשה. נכון להיום (לפחות עד כמה שידוע לי), אין בינה מלאכותית חזקה שמסוגלת לדמות פעולות רבות שמבצע האדם, כמו לפתח שיחה הגיונית ומלאה. בכל זאת, ניתן כיום להשתמש במחשב לביצוע פעולות חכמות רבות בתחומים ספציפיים, ובחלקם אפילו טוב יותר מבני אדם, כמו לשחק שחמט.

במציאות של היום, שבה נאספת כמות חסרת תקדים של נתונים (data) על כל דבר שאפשר, פיתוח של כלים שמסוגלים לנצל את הנתונים לצורך קבלת החלטות באופן אוטומטי או בהסתמך על המחשב הוא דבר נדרש, ופה אנחנו נכנסים. בקורס הזה נלמד כלים בסיסיים בתחום, ובפירוט בתחומי למידת מכונה מפוקחת ולא מפוקחת וקצת בנושא עיבוד מוקדם (preprocessing) של נתונים.

## סטטיסטיקה בייסיאנית:

הבעיה הכללית שאנחנו נתקלים בה למעשה בכמעט כל הקורס היא הבעיה הבאה:

נתון לנו אוסף תצפיות  $z$ . אנחנו רוצים לבנות פונקציה  $f_w$  התלויה בפרמטרים כלשהם  $w$ , שמתארת בצורה הטובה ביותר את המודל שממנו דגמנו את התצפיות. איך נמצא את  $w$ , שמגדירים לנו את הפונקציה  $f_w$  בצורה המיטבית? בגישה הבייסיאנית, אנחנו רוצים בעצם למצוא את

$$\hat{w} = \arg \max_w P(w|z)$$

כלומר, את אוסף הפרמטרים בעלי הסיכוי הטוב ביותר להתאים, בהינתן אוסף התצפיות. נשתמש בנוסחת Bayes:

$$P(w|z) = \frac{P(z|w)P(w)}{P(z)}$$

ננתח מה היא אומרת:

- $P(z)$  – לא משהו משמעותי במיוחד עבורנו, בתור מי שמנסה למצוא את  $w$ . זה בעצם סתם גורם נרמול,  $P(z) = \int P(z|w)P(w)dw$ .
- $P(z|w)$  – הסיכוי לקבל את התצפיות שקיבלתי, בהינתן פרמטרים  $w$  של המודל. פונקציה זו נקראת **פונקציית הנראות (likelihood)** של המודל.
- $P(w)$  – התפלגות הפרמטרים  $w$ . לזה קוראים **prior**, והוא מתאר את ההנחה ("האמונה") שלנו לגבי איך הפרמטרים אמורים להתפלג, מבלי לדעת איך נראות התצפיות של המודל. זאת התפלגות שרירותית שאנחנו קובעים עבור המודל שלנו.

דוגמה ל-prior: נניח ואנחנו מנסים לאמוד את הפרמטר  $p$ , שהוא הסיכוי שמטבע נתון לנו נופל על "עץ".

אם אנחנו מאמינים מראש שהמטבע די הוגן, נתאים את התפלגות ה-prior שלנו להיות עם ממוצע ב-0.5 וסטיית תקן די קטנה (התפלגות צרה יחסית). לעומת זאת, אם אנחנו בספק רב לגבי כך שהמטבע הוגן, ה-prior שלנו יהיה התפלגות רחבה הרבה יותר (ואם נלך למקרה הקיצוני ביותר, אז התפלגות אחידה).

אנחנו מנסים למקסם את הפונקציה  $P(w|z)$ . שקול לנסות למקסם את הפונקציה  $\log(P(w|z))$ , ואחרי פיתוח מתמטי קצר שכנראה ראיתם בהרצאה, מקבלים ביטוי לפונקציה הסופית שאותה אנחנו מנסים למקסם, **פונקציית ה-loss**.

אותה פונקציה מורכבת משני מחוברים עיקריים: פונקציית שגיאה (מדד שמתאר כמה אנחנו באמת טועים) ופונקציית **רגולריזציה** (פונקציה אחרת, המתארת כמה אנחנו סומכים על ה-prior שלנו).

## אומדים:

**אומד** הוא מונח מתאוריה סטטיסטית. בצורה פשטנית, אומד הוא הערכה של פרמטר על סמך מודל. למעשה, גם קודם ניסינו למצוא אומד לפרמטרים של המודל. יש סוגים שונים של אומדים, ואחד מהן

נקרא **אומד נראות מרבית (maximum likelihood estimator)**:  $\hat{w}_{MLE} = \arg \max_w P(z|w)$ .

הדוגמה הקלאסית לאמידה היא במקרה של התפלגות גאוסיאנית, עם פרמטרים  $\mu, \sigma^2$  ותצפיות  $\{x_i\}_{i=1}^n \sim N(\mu, \sigma^2)$ . נמצא אומדי נראות מרבית ל-  $\mu, \sigma^2$  על סמך התצפיות.

לאחר פיתוח קצר, מקבלים שלוג הנראות היא הפונקציה הבאה:

$$\ln(P(\{x_i\}|\mu, \sigma^2)) = \text{const} - n \ln \sigma - \frac{1}{2} \sum_i \frac{(x_i - \mu)^2}{\sigma^2}$$

וכאשר גוזרים לפי כל אחד מהמשתנים ומשווים ל-0, מקבלים את האומדים:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

האומד עבור  $\mu$  הוא **אומד חסר הטיה**, כלומר אומד כזה שמקיים  $E[\hat{\mu}] = \mu$ . מבלי להיכנס לפרטים, חוסר הטיה הוא תכונה שברוב המקרים נרצה עבור האומדים שלנו (אומד טוב הוא אומד שסטיית התקן שלו הוא נמוכה, ויש קשר בין סטיית התקן לבין ההטיה של האומד).

לעומת זאת, האומד עבור  $\sigma^2$  הוא אומד מוטה, ויש לו תיקון מפורסם שהופך אותו לחסר הטיה – שינוי המכנה מ- $n$  ל- $(n-1)$ .

## אמידת טווח ובדיקת השערות:

זהו מעבר קצר על נושא ששווה יהיה להיזכר בו לקראת סוף הקורס, כשנלמד על מדדים סטטיסטיים להערכת איכות (ובכל מקרה כדאי לדעת אותו).

דיברנו קודם על אומדים נקודתיים, כלומר כאלה המספקים ערך אחד לכל פרמטר. ניתן גם לדבר על אמידת טווח – מתן טווח ערכים שפרמטר נמצא בהם במובהקות גבוהה (לא "סיכוי", כי יש לפרמטר יש ערך אחד מתאים, אבל דומה ברעיון).

בהינתן מדגם  $\{x_i\}_{i=1}^n$  מהתפלגות התלויה בפרמטר  $\theta$ , **רווח בר-סמך** (לעתים רווח סמך) **בעל רמת מובהקות  $\alpha$**  עבור  $\theta$  הוא קטע  $(V_1, V_2)$  (שקצותיו הם פונקציות תלויות במדגם ולא בפרמטר), כך שהסתברות למאורע  $\theta \in (V_1, V_2)$  היא  $1 - \alpha$ .

למשל, במקרה של התפלגות נורמלית עם שונות ידועה, אפשר לבנות רווח סמך עבור  $\mu$  במובהקות  $\alpha$  להיות הקטע  $(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$ , כאשר  $\bar{X}$  ממוצע הדגימות הנתונות ו- $z_{\frac{\alpha}{2}}$  הוא המספר המקיים  $P(\Phi > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$  כש- $\Phi$  משתנה נורמלי סטנדרטי.

הנושא של אמידת טווח מתקשר לנושא של **בדיקת השערות**:

נניח שיש לנו השערה על סמך המדגם (למשל, שערך פרמטר כלשהו שווה למספר מסוים), המכונה **השערת האפס ( $H_0$ )**. נניח שיש לנו **השערה אלטרנטיבית ( $H_1$ )** (ההשערה האלטרנטיבית יכולה להיות מסוגים שונים, למשל השערה שערך הפרמטר שונה מהמספר הקודם, השערה שהוא גדול ממנו, השערה שהוא שווה למספר אחר ועוד).

אנחנו מנסים להכריע האם ניתן לדחות את השערת האפס לטובת ההשערה האלטרנטיבית. נקבע משתנה מקרי (**סטטיסטי מבחן**, למשל ממוצע המדגם) ורמת מובהקות  $\alpha$  (בדרך כלל לוקחים 0.05 או 0.01). מחפשים רווח סמך של הפרמטר עם המובהקות הזאת, ואז דוחים את השערת האפס אם הפרמטר לא נמצא בתוך רווח הסמך.

פרט לתהליך זה, ישנה אפשרות נפוצה אחרת לבדיקת השערות:

עבור סטטיסטי המבחן, נחשב את ה-**p-value** של הסטטיסטי הזה, ואם הוא חורג מרמת מובהקות שנקבעה מראש אז דוחים את ההשערה.

p-value מוגדר להיות הסיכוי, בהנחת השערת האפס, שמתקבל מדגם לפחות קיצוני כמו המדגם הנתון. "קיצונות" יכולה להתפרש בכל מיני אופנים, למשל שסטטיסטי המבחן כמשתנה מקרי גדול (או קטן, או שערכו המוחלט גדול וכד') לפחות כמו הערך של סטטיסטי המבחן שהתקבל מהמדגם הנתון.

## דוגמה מנחה – רגרסיה פולינומיאלית:

לצורך הדגמה של מושגים שונים, שחלקם כבר ראינו וחלקם נראה בהמשך, נסתכל על הדוגמה הבאה:

נתון המדגם הבא של נקודות:  $\{(x_i, y_i)\}_{i=1}^n$ , כאשר  $y_i = \sin(2\pi x_i) + \eta$  ו- $\eta$  רעש אקראי מהתפלגות עם תוחלת 0 ושונות לא ידועה. הדגימות שלנו הן למעשה **דגימות מתויגות**, כלומר קבוצה  $\{x_1, \dots, x_n\}$  כך שלכל  $x_i$  מתאים תיוג  $t_i$ . אנחנו ננסה למצוא פולינום שיהווה קירוב למודל ממנו נלקחו הדגימות הבאות:

$$y(w, x_i) = \sum_{j=0}^m w_j x_i^j$$

כך שלמעשה הפרמטרים שלנו הם  $\vec{w} = (w_1, \dots, w_m)$ .

אנחנו נעשה את זה על ידי מזעור של **פונקציית השגיאה**  $E(w) = \frac{1}{2} \sum_{i=1}^n (y(w, x_i) - t_i)^2$ . הקבוצה  $\{x_1, \dots, x_n\}$  נקראת **קבוצת האימון (training set)** של המודל, בעזרתה אנחנו **מאמינים את המודל**, כלומר מוצאים את הפרמטרים המתאימים ביותר.

אפשר (אפילו אנליטית) למצוא את  $w^* = \arg \min_w E(w)$ , אבל האם זה אומר שמצאנו פתרון טוב לבעיה שלנו? לא בהכרח! ייתכן שאנחנו מתאימים משקלים טוב מדי לסט האימון שלנו, ועבור דגימות חדשות מאותו מודל (**test set**) נקבל שהפונקציה תחזה עבורן ערך רחוק מאוד מהמציאות. למשל, אם מספר הפרמטרים שלנו (כלומר מעלת הפולינום) הוא  $m = n - 1$ , נוכל לעשות אינטרפולציה ולקבל פולינום שעובר דרך כל הנקודות, וכך השגיאה תהיה 0, אבל זה ממש לא מה שאנחנו רוצים.

למצב שבו הפונקציה שאנחנו בונים מתאימה את עצמה טוב מדי לסט האימון אבל לא מוכללת טוב מספיק לטסט המבחן קוראים **overfitting**.

כדי למנוע זאת, אנחנו בונים פונקציית loss שתהיה סכום של פונקציית השגיאה מקודם ופונקציית רגולריזציה, למשל:

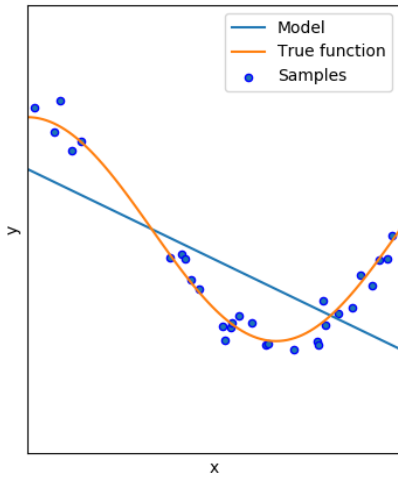
$$\tilde{E}(w) = E(w) + \frac{\lambda}{2} \|w\|_2^2$$

לפונקציית הרגולריזציה הנ"ל קוראים  **$L_2$  regularization / ridge regularization**, וניתן גם להשתמש בפונקציות אחרות (הנפוצה השנייה היא  $L_1$ , שידועה גם כ-**LASSO**, והיא  $\|w\|_1$ ). למקדם  $\lambda$  קוראים **מקדם רגולריזציה**, והוא שולט במידת האמונה שלנו לפונקציית הרגולריזציה. אם המקדם גבוה, אז אנחנו מאמינים מאוד שערכי המשקולות  $w$  צריכים להיות נמוכים והמודל צריך להיות פשוט יותר, ולהפך.

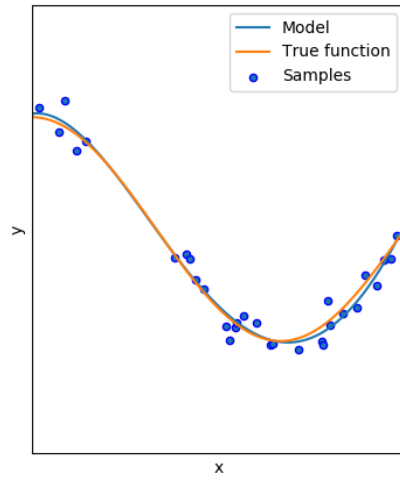
צריך לשים לב שערך רגולריזציה גבוה מדי, וכן מספר פרמטרים נמוך מדי, עלול להביא את המודל שלא להתאים לדגימות מכיוון שהוא פשטני מדי, מצב שנקרא **underfitting**. בעמוד הבא מופיעים 3 ציורים של רגרסיה פולינומיאלית שבהן מוצגות התופעות.

**הערה** –  $L_1$  שונה מ- $L_2$ . בהשוואה ל- $L_1, L_2$  היא פונקציה ש"מענישה" בחומרה יותר חריגה של הפרמטרים מ-0, גם אם היא מתרחשת בפרמטר אחד בלבד. לכן, היא יכולה לגרום למשקלים להיות אחידים יותר בגודליהם בהשוואה ל- $L_1$ .

Degree 1  
MSE =  $4.08e-01$  (+/-  $4.25e-01$ )



Degree 4  
MSE =  $4.32e-02$  (+/-  $7.08e-02$ )



Degree 15  
MSE =  $1.83e+08$  (+/-  $5.48e+08$ )

