מבוא להנדסת נתונים

1.

בלמידה פרמטרית, אנו מנסים לשערך פונקציה מתוך מדגם . ישנן שתי גישות עיקריות ללמידה פרמטרית.

בשיטות פרמטריות אנו עושים הנחות על הצורה של הפונקציה ומנסים לשערך את כמות סופית של פרמטרים שמרכיבים אותה.

יתרונות:

* מפשט את הבעיה ומקל מאוד על שיערוך הפונקציה.
* מוריד את הסיכון לoverfit.

חסרונות:

* מוסיף שגיאת שערוך, שכן יתכן הפונקציה האמיתית לא נמצאת במרחב החיפוש החדש.

בשיטות לא פרמטריות, לא מבצעים הנחות על הצורה של והיא יכולה להיות מורכבת מכמות לא סופית של פרמטרים.

יתרונות:

* מעריך את הפונקציה בצורה יותר מדוייקת.

חסרונות:

* דורש יותר דאטא בכדי לשערך בצורה טובה.
* נוטה יותר לoverfit.

2.

בדומה ל1, מודלים גמישים מאפשרים טווח יותר גדול של אפשריות לשערוך הפונקציה, במחיר של overfit ופגיעה בפשטות וב Interpretability.

3.

Bias – שגיאה שנוצר על ידי הפשטה של הבעיה (מרחב החיפוש) לצורה מצומצמת יותר של פונקציות. (underfitting).

Variance – שגיאה שנוצרת מהרגישות של המודל לתנודות בסט האימון. (overfitting).

הטרייד אוף בינהם מתבטא בכך שאם נגביל את עצנו מאוד וננסה להמנע מVariance על ידי הקטנת מרחב החיפוש (לדוגמא לפונקציות לינאריות). אם הפונקציה אינה לינארית במשתנים יהיה לנו bias גבוהה.  
לעומת זאת, אם אם ננסה להמנע מbias על ידי לקחת מוודל מאוד גמיש ולא להגביל בכלל את הפונקציה שמחפשים, אנו עשויים לקבל variance גבוה, על ידי לקחת פונקציה שמתאימה מאוד לסט האימון אבל לא מייצגת נכון את הפונקציה האמיתית.