המאמר היומי של מייק - 01.02.25  
Classical Statistical (In-Sample) Intuitions Don’t GeneralizeWell: A Note on Bias-Variance Tradeoffs, Overfitting and Moving from Fixed to Random Designs

מבוא:

שיטות ML מודרניות מציגות התנהגויות שסותרות באופן בולט אינטואיציות סטטיסטיות מסורתיות, במיוחד בנוגע לאימון-יתר (over-training), לאיזון בין הטיה לשונות, וליכולת הכללה. הסטטיסטיקה הקלאסית טוענת לעתים קרובות שככל שמורכבות המודל עולה, ההטיה יורדת, אך השונות עולה - איזון ידוע בין הטיה לשונות. עם זאת, תופעות כמו Double Descent או DD בקצרה ו- benign overfitting מאתגרות השקפה זו. המאמר המסוקרה טוען שתופעות אלה אינן נובעות באופן בלעדי ממודלים מורכבים, פרמטריזציית-יתר, או דאטה רבי-ממד, אלא דווקא ממעבר יסודי בין שני סוגי הבעיה הסטטיסטית: fixed and random design. המאמר מספק חקירה מתמטית של האופן שבו מעבר זה משנה באופן משמעותי עקרונות סטטיסטיים.

הגדרת הבעיה: משטרי random design - D\_r vs fixed design D\_f

ההבחנה בין D\_f ל- D\_r היא התובנה המהותית של המאמר:

משטר D\_f: הנקודות בטסט סט נותרות זהות לאלו שבאימון, כאשר רק התוויות שלהן נדגמות מחדש. ניתוח סטטיסטי קלאסי מניח את זה לעתים קרובות ועבורו אנו מנסים למזער את שגיאת השערוך in-sample.

משטר D\_r: גם הנקודות וגם התוויות במהלך הבדיקה נדגמים באופן בלתי תלוי מהתפלגות הדאה. משטר זה מתיישר עם האופן שבו מודלי ML משוערכים כיום, תוך התמקדות בשגיאת הכללה או שגיאת חיזוי מחוץ למדגם (out-of-distribution).

המעבר D\_f ל-D\_r גורם לשינויים עמוקים בהתנהגות של הטיה, שונות, ושגיאת החיזוי הכוללת. שינוי עדין אך משפיע זה הוא הסיבה המרכזית לכך שתופעות ML מודרניות נראות כמפרות את האינטואיציה הסטטיסטית הקלאסית.

מתמטית, השגיאות בשני המשטרים מוגדרות כך. שגיאת D\_f (שהיא in-sample) כאשר הן תוצאות שנדגמו מחדש בקלטים קבועים.

כאשר y\_i הפלטים שנדגמו מחדש עבור הפלטים מהטריין סט. שגיאת D\_r (מחוץ למדגם או out-of-distribution) מוגדרת באופן הבא:

כאשר גם x\_0 וגם y\_0 הם דגימות חדשות מהתפלגות הדאטה. שינוי זה מוביל להשלכות מרחיקות לכת עבור איזון ההטיה-שונות ותכונות ההכללה של מודלים. הטיה ושונות ב-D\_f מקבל צורה שמוכרת לנו היטב:

כאשר σ^2 הינו הרעש שלא ניתן לצמצום, (f\*(x היא הפונקציה ground-truth הנלמדת ואילו (f^(x הוא המשערך. עבור אומדנים פשוטים כמו k-NN. השונות יורדת מונוטונית עם עליית k כאשר יותר שכנים ממוצעים וההטיה עולה מונוטונית מכיוון שהממוצע כולל שכנים פחות דומים. איזון זה יוצר את העקומה בצורת U המוכרת מספרי הלימוד עבור שגיאת החיזוי כפונקציה של מורכבות המודל.

אולם במשטר D\_r מוליד התנהגות חדשה- האינטואיציה של הטרייד-אוף בין הטיה לשונות כבר לא עובדת בצורה כה פשוטה. ההטיה אינה יורדת מונוטונית עם המורכבות: השכן הקרוב ביותר עשוי שלא להתאים באופן מושלם לנקודת הבדיקה, מה שמוביל להטיית התאמת שכנים שאינה אפס. ההטיה יכולה להציג דפוס בצורת U, כאשר מודלים בעלי מורכבות בינונית ממזערים את ההטיה. התנהגות זו ניתן לבטא על ידי פירוק ההטיה ל:

שני הרכיבים הם:  
הטיית התאמת שכנים: נוצרת כאשר הממוצע המשוקלל של נקודות האימון אינו משחזר באופן מושלם את נקודת הבדיקה.  
הטיית מיצוע: נובעת אי-לינאריות של פונקציה האמיתית (כלומר המיפוי מנקודה ללייבל).

פירוק זה חושף שגם במצבים פשוטים ונמוכי-ממד, תכנון אקראי מכניס מורכבויות שמשבשות אינטואיציות קלאסיות.

תופעת Double Descent:

תופעת DD מתייחסת להתנהגות הלא-מונוטונית של שגיאת החיזוי כפונקציה של מורכבות המודל. היא מורכבת מעקומה בצורת U במשטר under-parametrization (מספר פרמטרי מודל קטן ממספר הדוגמאות) וירידה שנייה במשטר over-parameterization (מספר פרמטרי מודל גדול ממספר הדוגמאות). המחברת מדגישה כי DD אינו יכול להתרחש במצבי D\_f מכיוון שאינטרפולציה תמיד מובילה לשגיאת in-sample קבועה ERR\_fixed = σ^2. זאת מכיוון שמודלים במשטר זה חוזים באופן מושלם בנקודות האימון, מה שמוביל להטיה ושונות אפס בתכנון קבוע. עם זאת, במשטר D\_r, תופעת DD מופיעה באופן טבעי בגלל שינויים במורכבות המודל האפקטיבית (שלא נמדדת במספר הפרמטרים) ותכונות ההכללה בעת המעבר מעבר לאינטרפולציה.

Benign Overfitting(BO) vs. Benign Interpolation(BI)

המחבר מבקר את המונח BO, ומציע במקומו מונח BO. הגדרות קלאסיות של אוברפיט מרמזות על ביצועי הכללה ירודים, מה שסותר את הרעיון שביצועים מושלמים בטריין סט יכולה לעתים להניב ביצועים טובים גם על הטסט. במשטר R\_f, אינטרפולציה אינה יכולה להיות benign בגלל הדומיננטיות של שונות הרעש ERR\_fixed = σ^2.

במשטר R\_d, לעומת זאת, מודלים כמו רשתות נוירונים ויערות אקראיים(random forests) יכולים להציג התנהגות חדה-חלקה, בה הם מבצעים אינטרפולציה חדה בנקודות האימון אך מכלילים בצורה חלקה לקלטים שלא נראו. התנהגות זו ניתנת לכימות באמצעות מדדי מורכבות אפקטיבית. זאת אומרת מודלים שמפחיתים מורכבות אפקטיבית על טסט סט נוטים להציג אינטרפולציה שפירה.

השלכות:

חשיבה מחדש על חינוך סטטיסטי: קורסי מבוא צריכים להבהיר את ההבחנה בין R\_f ל-R\_d.

הסקה סיבתית ו-ML: בתחומים בהם נקודות מסט אימון עשויים לחזור (למשל, הסקה סיבתית), הנחות R\_d עשויות עדיין להיות רלוונטיות.

בחירת מודל ML: הבנה מתי אינטרפולציה היא benign דורשת מדידת מורכבות בזמן בדיקה, לא רק ביצועי אימון.

סיכום

עבודה זו מציעה פרספקטיבה מאוד מעניינת על מדוע אינטואיציות סטטיסטיות קלאסיות לא תמיד עובדת טוב בבעיות ב-ML מודרני. על ידי הדגשת השוואה בין R\_f ל- R\_d, המאמר מספק מסגרת מאחדת להבנת DD, Benign Interpolation, והתפקיד המתפתח של טרייד-אוף ההטיה-שונות.

https://arxiv.org/pdf/2409.18842