# Bayesian Hyper Parameter Optimizationאופטימיזציה בייסיאנית של היפר פרמטרים

*תקציר: במהלך ההכשרה, אסכם מספר נושאים "רוחביים" שאינם מתקשרים לפרק ספציפי. הנושא הראשון הוא אופטימיזציה בייסיאנית של היפר-פרמטרים. במסמך זה אסביר מה הבעיה שאופטימיזציה בייסיאנית מנסה לפתור, כיצד זה פותר ואציג ניסוי מצומצם שמראה את היתרון של אופטימיזציה בייסיאנית לעומת שיטות אחרות.*

## רקע

מציאת היפר-פרמטרים הינה בעיה מוכרת בעולם ה DS. רוב האלגוריתמים מתאפיינים במספר רב של קונפיגורציות הרצה, להן השפעה רבה על התוצאות של האלגוריתם. מציאת היפר פרמטרים אופטימליים הינה בעיה קשה שדורשת מספר רב של הרצות. גישות מוכרות לחיפוש הינן grid search ו random search אשר מנסות למצוא את הפתרון האופטימלי על ידי הרצות ערכים שלא מסתמכים על הרצות קודמות. ב Bayesian optimization מנסים לנצל תוצאות של הרצות קודמות על מנת לבחור ערכים חדשים לפרמטרים, מה שמקל על חיפוש במרחבי חיפוש גדולים (scalability) .

## רעיון

ממדלים את פונקציית המטרה ובכל איטרציה בוחרים את הערכים שנראה (לפי המודל) שישפרו את פונקציית המטרה בצורה המשמעותית ביותר. בצורה פורמלית: נרצה למצוא סט ערכים    כאשר היא פונקציית המטרה (למשל, accuracy במהלך ה cross validation), הם מרחב האופציות לערכים של היפר פרמטרים וערכים מסוימים להיפר פרמטרים בהתאמה.

איך ממדלים את פונקציית המטרה? **בעזרת ניסיונות קודמים**. ביותר מילים, המודל משתמש בתוצאות של ערכים קודמים וכדי לבנות מודל הסתברותי שמנסה לשערך (מודל זה נקרא surrogate). השלבים של האופטימיזציה היא:

1. נבנה מודל surrogate
2. עבור תנאי עצירה מסוים:
   1. נמצא ערכים להיפר-פרמטרים שנותנים תוצאה מקסימלית ל surrogate
   2. ננסה את הפרמטרים עם האלגוריתם עצמו ונקבל ערך מפונקציית המטרה האמתית
   3. נעדכן את מודל ה surrogate

ה tradeoff בשיטה הוא שבחירה של הערכים הבאים בכל איטרציה לוקחת יותר זמן (לעומת grid או random) אבל נריץ פחות איטרציות. כלומר, במקרה שהרצה של פונקצית המטרה היא יקרה (למשל, כוללת אימון של מודל על הרבה מידע) נרצה להשתמש בשיטה הבייסיאנית. לעומת זאת, אם המודל הוא פשוט (בלי הרבה היפר-פרמטרים) ואין הרבה מידע אולי נעדיף gridsearch.

## תכלס

בפועל, צריך לשים לב לפונקציה שמגדירה את ה surrogate ולפונקציה שבוחרת את הערכים הבאים לבדיקה. אופציות מוכרות לפונקציית surrogate הן Gaussian Process, Random Forest Repressor, Tree Structured Parzen Estimator. אופציות לפונקציית בחירה של ערכים הם Expected Improvement, Probability Improvement, GP Upper Confidence Bound.

## מדגים

לשם ההדגמה, עבדתי עם boston dataset מ sklearn ללא כל עיבוד מקדים. השוויתי בין Bayesian, Grid, Random. המודל הינו random forest regressor מ sklearn. בדיקת טיב האלגוריתם התבצעה בעזרת MSE ו 5FCV. מרחב החיפוש שהשתמשתי בו מתואר בטבלה 1. את random, Bayesian הגבתי ל 20 איטרציות.

|  |  |  |
| --- | --- | --- |
| משתנה | מינימום | מקסימום |
| n\_estimators | 50 | 500 |
| max\_depth | 2 | 50 |
| min\_samples\_split | 2 | 9 |

טבלה 1: מרחב החיפוש של הערכים היפר-פרמטרים

מבחינת זמנים, את Grid עצרתי כי לקח לו יותר מידי זמן. Random הצליח לסיים לרוץ לאחר 1:15 דקות ו Bayesian סיים לרוץ לאחר 1:33 דקות. הדבר מסתדר עם העובדה ש Bayesian משקיע יותר זמן בלבחור את האיטרציה הבאה שלו. מבחינת תוצאות: Random הגיע לתוצאה יותר טובה עם MSE של 8.70 לעומת MSE של 8.96 של Bayesian.