# פרק 13: רגולריזציה, פשטות וידע מוקדם

## Ridge Regression ו- Lasso

**מבוא**

לפני שנדון ברגרסיה נתבונן בבעיית שערוך ממוצע תחת אילוצים:

$$\text{Minimize} \quad \sum_{i=1}^{n}(Y_i-\mu)^2 \ \text{ subject to } \ \mu^2 \le C$$

בעזרת כופלי לגרנז הבעייה שקולה ל:

$$\text{Minimize} \quad \sum_{i=1}^{n}(Y_i-\mu)^2 + \lambda_C \mu^2$$

נגזור:

$$-2\sum_{i=1}^{n}(Y_i-\hat{\mu}_C) + 2\lambda_C\hat{\mu}_C = 0$$

ונקבל:

$$\hat{\mu}_C = \frac{\sum_{i=1}^{n} Y_i}{n+\lambda_C} = K_C\bar{Y}, \qquad K_C = \frac{n}{n+\lambda_C} \qquad (*)$$

ורואים שהאפקט של C קריטי:

$$C \to 0, \quad \hat{\mu}_C \to \bar{Y}$$
$$C \to \infty \quad \hat{\mu}_C \to 0$$

מסקנה: האפקט של רגולריזציה הוא אפקט מקווץ (shrinking).

הערה: ניתן להוסיף באופן שקול מס׳ כלשהוא $n_C$ של נקודות מלאכותיות $Y_i=0$ שהרי נקבל את הבעיה

$$\text{Minimize} \quad \sum_{i=1}^{n}(Y_i-\mu)^2 + \mathrm{n}_C(0-\mu)^2$$

בתרגול נראה ש:

$$\text{Minimize} \quad \sum_{i=1}^{n}(Y_i-\mu)^2 + \lambda_C \mu^2$$

שקול לשערוך MAP (maximum a-posteriori) של $\mu$ עם פריור גאוסי.

הסתכלות שקולה לפיכך היא להטות את הפתרון (bias) בכיוון רצוי. השאלה היא כיצד למצוא הטייה מועילה. האינטרפרטציות הן:

1. בייסיאנית: מציאת פריור מתאים
2. סיבוכיות: פתרון פשוט (נורמה קטנה) לעומת פתרון מסובך (נורמה גדולה)

## Ridge Regression

נתבונן כעת ברגרסיה רב מימדית ונניח:

1. ל- X ממוצע 0, וקטור $p$ מימדי.
2. ל- Y ממוצע 0

נחפש רגרסור מהצורה $Y \approx \beta^T X$. למודל כזה קוראים ה- "מודל הסטנדרטי."

נגדיר את הפונקציה הבאה:

$$\mathrm{SSE}_\lambda(\beta) = \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 + \lambda \|\beta\|_2^2$$

ה- $\beta$ שמביא למינימום את $SSE_\lambda(\beta)$ הוא הפתרון של בעיית ה- ridge regression:

$$\text{Minimize} \quad \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 + \lambda \|\beta\|_2^2$$

קל לראות שהבעייה שקולה לבעיית אופטימיזציה עם אילוצים על הנורמה של הרגרסור $\beta$ (מכופלי לגרנז') ולפיכך בעיית האפטימיזציה היא בעיה קמורה.

שאלה: מה אוסף הדוגמאות המלאכויות עבורן נקבל את הבעיה דלעיל ללא איבר הרגולריזציה?

נגזור לפי $\beta$ ונקבל:

$$\frac{\partial}{\partial \beta(k)} \mathrm{SSE}_\lambda(\beta) = 2\sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right) X_i^T(k) + 2\lambda \beta(k)$$

וברישום מטריצי לאחר השוואה ל-0 נקבל:

$$-Y^T X + \hat{\beta}_\lambda^T \left(X^T X + \lambda I\right) = 0$$

(נשים לב: $\beta_\lambda$ היא וקטור $p$ מימדי, $Y$ וקטור $n$ מימדי ו- $X$ מטריצה $n\ x\ p$.)

ולכן:

$$\hat{\beta}_\lambda^T = Y^T X \left( X^T X + \lambda I \right)^{-1}$$

או
$$\hat{\beta}_\lambda = \left( X^T X + \lambda I \right)^{-1} X^T Y$$

(היפוך המטריצה מותר תמיד כי $\lambda > 0$ .)

זו הגרסה הוקטורית של משערך הממוצע (*). נעיר שהאינטרפטציה הבייסיאנית עובדת גם כאן (ראה תרגול)

## LASSO

LASSO היא בעיית האופטימיזציה בה נורמת 2 של איבר הרגולריזציה מוחלף ע"י נורמת 1, ז"א :

$$\text{Minimize} \quad \sum_{i=1}^{n} \left( Y_i - X_i^T \beta \right)^2 + \lambda \sum_{k=1}^{p} \left| \beta(k) \right|$$

קל לראות שזוהי בעייה קמורה עם אילוץ על נורמת 1 של $\beta$ . למרבה הצער, הערך המוחלט מונע מאיתנו גזירה נוחה של פונקציית המטרה. למרות זאת ניתן לחשב את הפסאודו-גרדיאנט של פונקציית המטרה וקיימים אלגוריתמים יעילים לפיתרון בעיית האופטימיזציה.

אלגוריתם ה- LASSO עובד היטב אם קיימים פתרונות דלילים (ז"א רוב הקאורדינטות של $\beta$ הם 0).

## בחירת $\lambda$

אם היינו יודעים את ה- MSE של מסווג מסויים היינו בוחרים את המסווג עם ה- MSE המינימלי. ניתן לשערך את ה- MSE ע"י אימות צולב (Cross Validation).

שיטה אחרת היא לרשום חסם הכללה התלוי ב- $\lambda$ , כפי שנעשה בהרצאה הדנה בתיאוריה.

נסיים בהבחנה מעניינת. ניתן להתבונן במטריצה הבאה :

$$S_\lambda = \left( X^T X + \lambda I \right)^{-1} X^T$$

המקיימת (עבור Ridge Regression) : $\hat{\beta}_\lambda = S_\lambda Y$ . ניתן לחשוב על $S_\lambda$ כהטלה ממרחב $n$ מימדי למרחב $p$ מימדי. עבור $\lambda = 0$ מספר דרגות החופש הוא $p$, וככל שנגדיל את $\lambda$ יקטן מספר דרגות החופש האפקטיביות שניתן להעריכו ע"י $n - TR\left(S_\lambda\right)$ .