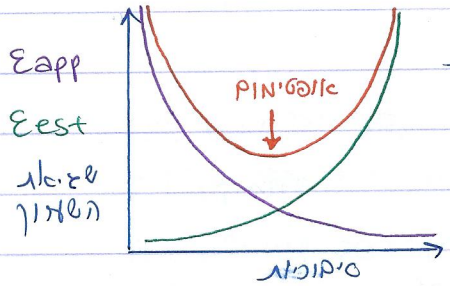
תרגול # 9 — ב' - סמ 1/61

רגולריזציה:



מוטיבציה: רוצים למצוא את הנק' האופטימלית ב- bias-variance tradeoff.
אם נבחר סיבוכיות קטנה מדי נעשה underfitting, וגדולה מדי נעשה overfitting.

⇐ אפשר להתקרב לאופטימום בעזרת רגולריזציה.

עד כה במציאת אופטימיזציה ניסינו למצוא: $h^* = \underset{h\in\mathcal{H}}{\arg\min}\, F_S(h)$

(כאשר F_S הוא פונק' המחיר). החידוש הוא שמעכשיו ננסה למצוא:

$\boxed{h^* = \underset{h\in\mathcal{H}}{\arg\min}\,(F_S(h) + \lambda R(h))}$, כלומר נוסיף איבר שלא תלוי בדאטה

אלא רק בהיפותזה עצמה: בגסות שלה. שמאפיינת כמה היא מסובכת.

איבר זה נקרא איבר הרגולריזציה, ותפקידו ליצור יחס כוחות שעוזר

על רמת הסיבוכיות של ההיפותזות שאנו מקבלים כך שנהיה קרובים לאופטימום.

(*) עבור ערכים גדולים של λ: נכפיל את $R(h)$ במקדם גדול, ועבור

h מסובכות נקבל עונשה גבוהה שתחיק אותנו מלהיות ה- $\arg\min$.

לכן למעשה נעדיף פונק' עם סיבוכיות נמוכה שם נשלם עליה פחות.

אז נקבל שונות נמוכה! - bias גבוה.

(*) עבור ערכי λ קטנים: האיבר השישון יהיה הדומיננטי (ההבדל על

סיבוכיות h יהיה קטן משמעותית) ונעדיף היפותזות מסיבוכיות גבוהה.

אז נקבל שונות גבוהה! - bias נמוך.

דרך שנייה לאיבר הרגולריזציה: נראה איך ניתן להגדיר פונק' שונות:

ה- Setting יהיה רגרסיה לינארית: $F_S(h)$ יהיה Mean Least Squares.

1. Ridge - ℓ_2 regularization: $\hat{w}_\lambda^{ridge} = \underset{w}{\arg\min}\, \underbrace{RSS(S;w)}_{\text{Residual Sum of Squares}} + \lambda\|w\|_2^2$

★ כמו שלא- w הפתרון יש נורמה גבוהה יותר, כך ההיפותזה מסובכת יותר.

הוכחה סרוגלית ⇐ למה: תהי $X \in \mathbb{R}^{n\times n}$ מטריצה אורתוגונלית, $D \in \mathbb{R}^{n\times n}$ מטריצה

אלכסונית. אזי: $XDX^T + \lambda I = X(D+\lambda I)X^T$

טענה: תהי X design matrix! - y response vector כך ש:

$X\Sigma V^T$ פירוק ה- SVD של X. אזי ה- ridge estimator

הוא: $\hat{w}_\lambda^{ridge} = X\Sigma_\lambda V^T y$ כאשר $[\Sigma_\lambda]_{i,i} = \frac{\sigma_i}{\sigma_i^2+\lambda}$.

הוכחה: נסמן: $f_{\ell_2}(w) = RSS(S;w) + \lambda\|w\|_2^2$, עבור $\lambda \geq 0$.

כלשהו. ננסה למצוא את w המאפס אותה - נגזור:

$$\frac{1}{2}\nabla_w f_{\ell_2}(w) = \underbrace{XX^Tw - Xy}_{\text{least Squares-מ}} + \underbrace{\lambda w}_{\text{מאיבר הרגולריזציה.}}$$

ונשווה לאפס: $XX^Tw - Xy + \lambda w = 0 \iff (XX^T+\lambda I)w = Xy$

ונקבל: $\hat{w}_\lambda^{ridge} = (XX^T+\lambda I)^{-1}Xy$

נשתמש בפירוק ה-SVD:

$$\hat{w}_\lambda^{ridge} = (XX^T+\lambda I)^{-1}Xy = (U\Sigma V^TV\Sigma^TU^T+\lambda I)^{-1}Xy =$$
$$= (U\Sigma\Sigma^TU^T+\lambda I)^{-1}Xy = (U\Sigma^2U^T+U\lambda U^T)^{-1}Xy =$$
$$= U(\Sigma^2+\lambda I)^{-1}U^TXy = U(\Sigma^2+\lambda I)^{-1}U^TU\Sigma V^Ty =$$
$$= U(\Sigma^2+\lambda I)^{-1}\Sigma V^Ty = U\Sigma_\lambda V^Ty$$

עבור: $[\Sigma_\lambda]_{i,i} = \frac{\sigma_i}{\sigma_i^2+\lambda}$, כאשר σ_i הערכים הסינגולריים של X. ■

תרגיל: דרך נוספת לתשובה על Ridge Regression היא לתשובה על בעיית OLS (Ordinary Least Squares) על דאטה

שעבר שינוי: $X_\lambda = [X \mid \sqrt{\lambda I_d}] \in \mathbb{R}^{d\times(m+d)}$, $y_\lambda = [y \mid 0_d]^T \in \mathbb{R}^{m\times d}$

מטריצות "הדבקה" וקטור ומטריצה "הדבקה"

אז פתרון ה-OLS שמאפס ל-X_λ, y_λ הוא: (כפי שלמדנו בריגרסיה לינארית)

$$\hat{w}_\lambda^{OLS} = (X_\lambda X_\lambda^T)^{-1}X_\lambda y_\lambda = (X_\lambda^{\ddagger})^T y_\lambda$$

כאשר להזכיר $X^{\ddagger} = V\Sigma^{\ddagger}U^T$ ה-pseudo-inverse של X:

$\Sigma^{\ddagger}_{i,i} = \begin{cases} 1/\Sigma_{i,i} & : \Sigma_{i,i} \neq 0 \\ 0 & : \Sigma_{i,i} = 0 \end{cases}$. כמו כן הראינו ש- $(XX^T)^{-1}X = (X^{\ddagger})^T$.

וכעת ה- pseudo-inverse שמאפס ל- X_λ, y_λ

הוא בדיוק הפתרון ל: Ridged-Regression של X, y.

(★) תכונה חשובה של ה-ridge estimator היא שהוא מאמוד על הטעיה

שלדוגמה X הוא מטריצה (כמעט) סינגולרית ולכן בחישוב $(XX^T)^{-1}$. "מכן

ונתקל בבעיה נומרית, אי כך שאנו מעדיפים לחשב את $(XX^T+\lambda I)^{-1}$

והוספת λI (ל-λ גדול מספיק) תהפוך את XX^T למטריצה סינגולרית,

ולכן לא נסיים את התהליך בשגיאה נומרית קבועה.

(★) תכונה נוספת (נכונה בתרגיל): ה-estimator הוא biased,

אך במקרה הזה זה תורם לשפור הביצועים (לא כמו בשיטות עדיפה ☺).

2. Lasso - ℓ_1 Regularization

מתאים למקרה שבו מעבר לפעולת הרגולריזציה (למנוע מהפרמטרים שלנו להיות מסיבוכיות גבוהה מדי ולכן overfit) הוא מיצר פתרונות sparse: הוא שופט חלק מהפצ'רים.

★ זה מועיל לנו, כי זה מאפשר את העולם האמיתי שבו ישנם פיצ'רים חשובים פחות מאחרים, אינם לא נכונה ללמוד. ⇐ דרך לעשות feature selection.

מה שהביא אותנו ל- Lasso הוא מקרה ה- Best-Subset Selection:
במקרה זה איבר הרגולריזציה הוא עם "נורמה" אפס:

$$\underset{w}{\arg\min}\, f_{\ell_0}(w) = \underset{w}{\arg\min}\, \underbrace{\|y - X^T w\|_2^2}_{\text{(Least Squares)}} + \lambda \|w\|_0 \quad (\star)$$

$\|\cdot\|_0$: סופרת את מס' הנכיבים שאינם אפס.

כך ה best-subset Selection בוחר את הפיצ'רים החשובים ע"י איפוס הפצ'רים הנגיחים. כך הקטנו את המימד d.

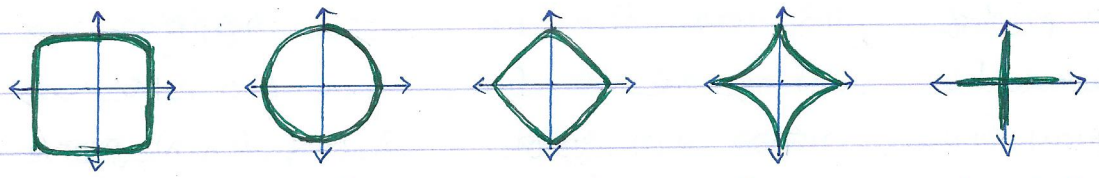
הבעיה: מציאת Best subset of features היא בעיית אופטימיזציה NP-H והיא אינה קמורה. (כי $\|\cdot\|_0$ אינה נורמה, כל נורמה אכן קמורה).
⇐ זה מבקש אותנו לפ' ז פתרונות משקלים בהתאם נורמה ℓ_q:

1. Sparsity (גזירות): עבור $q \le 1$, אם נקח ב- (★) נורמת ℓ_q, הפתרונות שנקבל יהיו גזירים כמו שהגדרנו מקודם. ונקבל active set של פיצ'רים מימד קטן ל- d.

2. Convexity (קמירות): עבור $q \ge 1$ בעיית האופטימיזציה המוגדרת ע"י התלפת נורמת $\|\cdot\|_0$ ב- ℓ_q תהיה בעיית אופטימיזציה קמורה (כי כלפי LS קמורה ואיבר הרגולריזציה קמ) שעבורה נראה אלגו' הפתרון אותם בעלות (בזמן פול').

בעיית ה- Lasso Regression עוקבת $q=1$ ומרוויחה את ז הפתרונות הגזירים
אז: $\underset{w}{\arg\min}\, f_{\ell_0}(w) = \underset{w}{\arg\min}\, \|y - X^T w\|_2^2 + \lambda \|w\|_1$

למה מקבלים פתרונות גזירים? נבחן את בגזר היחידה עבור נורמות שונות:



$q=4$ $q=2$ $q=1$ $q=0.5$ $q=0.1$

ℓ_2: קמורה וגזירה ⇐ לכל גזירה. ℓ_1: ★ קמורה ולא גזירה ⇐ גזירה. לכל קמורה ואי גזירה ⇐ גזירה

אינטואיציה לבר

שהפונקציה הסיבתית לגילוי: $f_{\ell_0}(w)$ הוא בעיית אופטימיזציה תחת אילוצים (ראינו ברגרסיה רגשית לבר).

הפתרון האופטימלי של הבעיה המלאה (עם איבר הרגולריזציה) מתקבל בחיתוך בין משטח האילוצים (אצלנו = תלורה) עם קו הגובה ה- contour במרחב w של הפונקציה $\|y - X^T w\|_2^2$ שהוא למעשה אינפס. הלוגה האינטגרלית נובעת מפירוק ה-SVD של X) החיצוני ביותר.

לצורה הסרת פינות, אין סיבה שהחיתוך יתרחש על אחד הצירים, אך לצורה עם פינות על הצירים (שמוגדרות ע"י ℓ_q ל- $q \geq 1$ כפי שראינו) החיתוך בד יתקיים על אחד הצירים ולכן הפתרון יהיה על הצירים, כלומר חלק מהקואו' שלו יהיו 0.

מקרה פרטי: Orthogonal design: $XX^T = I_d$, X ל- design matrix ה-.

במקרה זה לפתרון יש צורה סגורה (כמו שהיתה ב- Ridge), בלונ' אין.

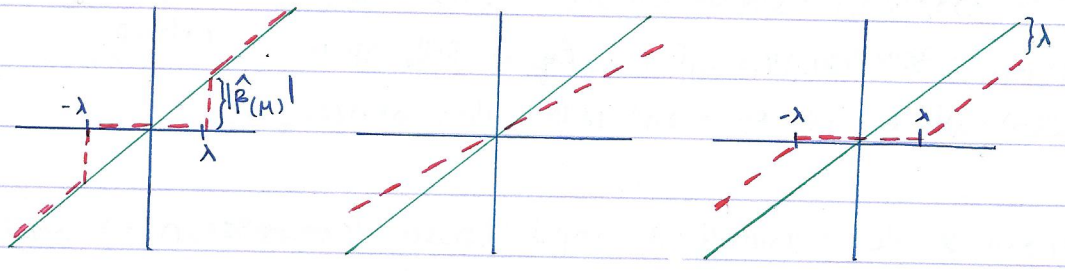
נכתוב את 3 הצורות שעבדנו עליהן:

במקרה הזה, הכול אל ימר פשוט מהכלי ⇐

(i) $\hat{w}_\lambda^{ridge} = \frac{\hat{w}^{OLS}}{1+\lambda}$

(ii) $\hat{w}_\lambda^{lasso} = \eta_\lambda^{soft}(\hat{w}^{OLS})$ where $\eta_\lambda^{soft}(x) = \text{sign}(x)\,[|x| - \lambda]_+ = \begin{cases} x-\lambda & : x \geq \lambda \\ 0 & : |x| < \lambda \\ x+\lambda & : x \leq -\lambda \end{cases}$

(iii) $\hat{w}_\lambda^{subset} = \eta_\lambda^{hard}(\hat{w}^{OLS})$ where $\eta_\lambda^{hard}(x) = \mathbb{1}_{[|x| \geq \lambda]} \cdot x$



הפתרון הרגולרי

הפתרון לפי OLS

Best Subset	Ridge	Lasso
⇓	⇓	⇓
עושה רק ספסיפיקציה: שולח ערכים שאינם 0 ל-0. מתוך לגמרם לא משנה כלום.	לא עושה ספסיפיקציה, לא שולח שום ערך ל-0. הערכים קטנים בפקטור $\frac{1}{1+\lambda}$.	עושה ספסיפיקציה: שולח ערכים שאינם 0 ל-0. שאר הערכים, מוקטנים ב-λ. כלומר מ כולם את הקבוצה.
ספסיפיקציה	shrinkage	shrinkage + ספסיפיקציה

טענה: תהא X design matrix אורתוגונלית, ו-y response vector.
נסמן את פתרון ה- OLS ב- $\hat{\omega}$. אזי פתרון לבעיית רגרסיבית לאסו הוא מתקבל:
$$\hat{\omega}^{lasso}(\lambda) = \eta_\lambda^{soft}(X)$$

הוכחה: נציב בדרך ש: $\hat{\omega} = (XX^T)^{-1}Xy \underset{\downarrow}{=} Xy$
(אותנו במקרה האורתוגונלי)

מתקיים:
$$f_{\ell_1}(\omega) = \tfrac{1}{2}\|y - X^T\omega\|_2^2 + \lambda\|\omega\|_1 = \frac{1}{2}\left(\|y\|^2 - 2\underbrace{y^TX^T}_{\hat{\omega}^T}\omega + \omega^T\underbrace{XX^T}_{I_d}\omega\right) +$$
$$+\lambda\|\omega\|_1 = \frac{1}{2}\left(\|y\|^2 + (\omega^T - 2\hat{\omega}^T)\omega\right) + \lambda\|\omega\|_1 =$$
(נפתח בכפל, קיבוץ, טלסקופ) $\omega^T\omega - \hat{\omega}\omega + \lambda\|\omega\|_1$
$$= \frac{1}{2}\|y\|^2 + \sum_{j=1}^{d}\left(\tfrac{1}{2}\omega_j^2 - \hat{\omega}_j\omega_j + \lambda|\omega_j|\right) =$$
$(|\omega| = \text{sign}(\omega)\cdot\omega)$
$$= \frac{1}{2}\|y\|^2 + \sum_{j=1}^{d}\left(\tfrac{1}{2}\omega_j - \hat{\omega}_j + \lambda\,\text{sign}(\omega_j)\right)\omega_j$$

נגזר את הביטוי ונשווה ל-0 כדי למצוא מינימום:
$$0 = \frac{\partial\left(\frac{1}{2}\|y - X^T\omega\|_2^2 + \lambda\|\omega\|_1\right)}{\partial\omega_j} = \frac{\partial\left(\underbrace{\frac{1}{2}\|y\|^2}_{\text{קבוע ב- }\omega_j} + \underbrace{\sum_{j=1}^{d}\left(\tfrac{1}{2}\omega_j - \hat{\omega}_j + \lambda\,\text{sign}(\omega_j)\right)\omega_j}_{\text{לכל } j\neq i\text{: קבוע ב- }\omega_j}\right)}{\partial\omega_j} =$$
$$= \frac{\partial\left(\tfrac{1}{2}\omega_j - \hat{\omega}_j + \lambda\,\text{sign}(\omega_j)\right)\omega_j}{\partial\omega_j} = \frac{\partial\left(\tfrac{1}{2}\omega_j^2 - \hat{\omega}_j\omega_j + \lambda\,\text{sign}(\omega_j)\omega_j\right)}{\omega_j} =$$
$$= \omega_j - \hat{\omega}_j + \lambda\,\text{sign}(\omega_j) \iff \omega_j = \hat{\omega}_j - \lambda\,\text{sign}(\omega_j)$$

נחלק למקרים:

(*) אם $|\hat{\omega}_j| < \lambda$:
- נניח בשלילה ש- $\omega_j > 0$: אזי $\hat{\omega}_j - \lambda\,\text{sign}(\omega_j) = \hat{\omega}_j + \lambda$ ($\text{sign}(\omega_j) = -1$), אך לכל $|\hat{\omega}_j| < \lambda$ נקבל ש- $-\lambda < \hat{\omega}_j < \lambda$ ולכן $\hat{\omega}_j + \lambda > -\lambda + \lambda = 0$ בסתירה.
- נניח בשלילה ש- $\omega_j < 0$: אזי $\hat{\omega}_j - \lambda\,\text{sign}(\omega_j) = \hat{\omega}_j - \lambda$
אך לכל $|\hat{\omega}_j| < \lambda$ נקבל ש: $\hat{\omega}_j < \lambda$. ולכן $\hat{\omega}_j - \lambda < \lambda - \lambda = 0$ בסתירה.
ולכן $\boxed{\omega_j = 0}$.

(*) אם $\hat{\omega}_j \geq \lambda$ אז השוויון $\omega_j = \hat{\omega}_j - \lambda\,\text{sign}(\omega_j)$ נכון רק עבור $\omega_j \geq 0$
(אחרת נקבל שלכל ימין תיהפך, ולכל שמאל שלילי) כלומר: $\boxed{\omega_j = \hat{\omega}_j - \lambda}$

(*) אם $\hat{\omega}_j \leq \lambda$ אז השוויון $\omega_j = \hat{\omega}_j - \lambda\,\text{sign}(\omega_j)$ נכון רק עבור $\omega_j \leq 0$
(אחרת נקבל שלכל ימין שווה ל: $\hat{\omega}_j - \lambda \geq -\lambda = 0$ מצד אחד
שמאל שווה ל- $\omega_j \geq 0$ בסתירה) כלומר: $\boxed{\omega_j = \hat{\omega}_j + \lambda}$

$\Leftarrow$ וקיבלנו: $\hat{\omega}^{lasso}(\lambda) = \eta_\lambda^{soft}(X)$ כנדרש.

∎