High Var - Over Fitting

10:41 AM Wednesday, January 20, 2021

0:00:00

Over fitting

- אנחנו מותאמים יותר מדי ל training שלנו
- המודל גמיש מדי
- יש לנו מעט מדי מידע, והמודל לא מזהה את התבנית הכוללת

נפתור בעזרת:

- **Ensembles** - אנסמבל: הרבה מספר מודלים, חיזוי משותף
- הפחתת ממד הקלט **(Clustering \ PCA)**
- **רגולריזציה**: מגבלות על המשקולות
- צמצום מספר הפיצ'רים **Feature Selection**

איך נפתור?

Feature Selection
1. נתחיל מקבוצה ריקה
2. נעבור בלולאה חמדנית, בכל איטרציה נבחר פיצ'ר אחד

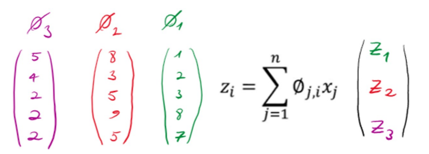
→ בכל איטרציה בלולאה נבדוק את כל הפיצ'רים נבחר את הפיצ'ר שהכי משפר לנו את המודל → נוסיף את הפיצ'ר הזה לקבוצת הפיצ'רים הנבחרה → נעצור כשנגיע למכסה שהוקצתה לנו. כמה אימונים מתבצעים במקסימום: $n\binom{n}{2}$

רגולריזציה
1. נוסיף לפונקציית השגיאה עונש למשקולות של האיברים בחזקה גבוהה
2. איננו יודעים על איזה משקולת לוותר, לכן נעניש את כל המשקולות חוץ מהביאסים

→ **L1 - לאסו**: ככל שהעונש גדל, יותר **משתנים מתאפסים** ונעלמים (בRidge L2 לא נעלמים)

PCA
נמצא ערכים עצמיים וקטורים עצמיים (טרנספורמציה ליניארית)

→ הווקטורים העצמיים הראשונים הם (מתמטית) הוקטורים המשפיעים ביותר. נקח את n הוקטורים העצמיים שאנחנו רוצים (n הוא המימד שאליו אנחנו רוצים לרדת)

→ 

$$z_i = \sum_{j=1}^{n} \phi_{j,i} x_j$$

Ensembles
חכמת ההמונים - צירוף של **הרבה מודלים שונים**, לא קורלטיביים

→ **שימוש ברגרסיה:** נמצע את התחזיות של כל המודלים → **שימוש בקלסיפיקציה:**
1. נמצא הסתברויות
2. הצבעת הרוב

→ **הגיון:** ע"י מיצוע של מודלים שאינם קשורים, Overfitted נעלים את הVariance שנובעים מהתאמת היתר

Bagging
אימון מודלים שונים על תת קבוצות שונות של data

→ **Bootstrap** - יצירת קבוצות אימון שונות ע"י דגימה אקראית **עם חזרות** מתוך קבוצת אימון יחידה. לכל קבוצה מותאמת קבוצת ולידציה משאר הדגימות → **K-fold** מחלקים את הדאטה לk קבוצות, כל פעם אחת מהן ולידציה, והשאר אימון → הברה: יש Ensembles שעובדים על אותו מודל (KNN למשל) אבל

בפועל ניתן "לנצל" k-fold cross validation - שמטרתו לשערך שגיאה ולכוונן היפר-פרמטרים, כדי לבנות Bagging בעזרת k המודלים שנוצרו Ensembles

feature selection- Hybrid

1. Begin with Null model
2. K=1…n, until (stop criteria)
   - First add a feature as in Forward selection
   - Then, remove a feature that no longer provide improvement
   - Save M_k
3. Among the n models, Choose the best M_k using CV

שילוב 2 השיטות:
התחל מקבוצת מאפינים ריקה
לכל k=1…n בדוק מבין המאפינים החסרים, איזה מאפיין להוסיף, הוסף את זה שמשפר הכי הרבה
אח"כ, בדוק מבין המאפיינים שבקבוצה, איזה מאפיין כדאי להוריד (נוכחותו אינה הכרחית)
בסוף התהליך בחר ע"י CV מבין k המודלים שנבנו

הבעיה בשיטה: עלולה להסיר feature שאינו משמעותי כשלעצמו, אבל בצירוף עם feature אחר הוא דווקא כן משמעותי להפחתת השגיאה

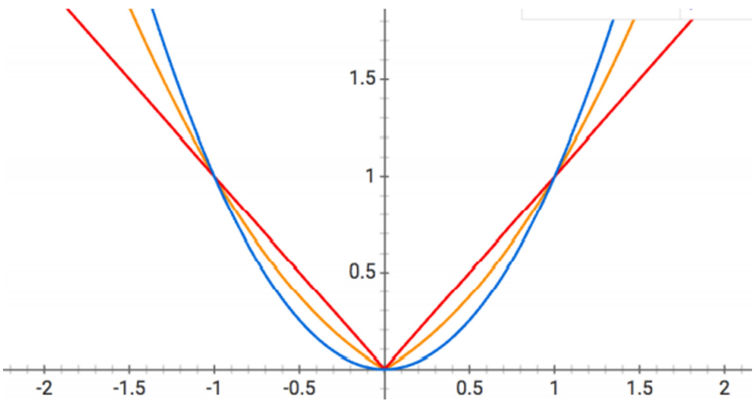
regularization

משתמשים בדר"כ לקלספיקציה

נוסיף לפונקציית ה-loss שלנו "עונש" למשקולות של האיברים בעלי חזקה גבוהה, נקבל פונקציית היפותזה "חלקה" יותר.

נגדיר את γ (gamma) כקבוע הרגולריזציה, העונש למשקולת שערכה 1.

$$loss_{reg}(w) = loss(W) + \gamma\left|\left|W\right|\right|_n$$



רגולריזציית Ridge – L2 – מרחק אוקלידי

$$\|W\|_2 = \frac{1}{2}\sqrt{\sum_{i=1}^{n} w_i^2}$$

$$w_i = w_i - \lambda\frac{\partial loss}{\partial w_i} - (\gamma \cdot w_i) =$$

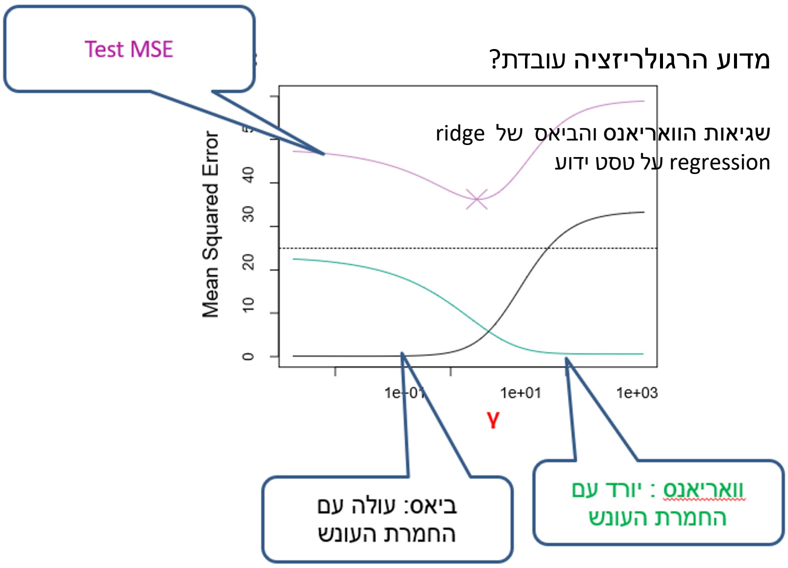
$$= w_i(1-\gamma) - \lambda\frac{\partial loss}{\partial w_i}$$

רגולריזציית Lasso – L1 – מרחק מנהטן

$$\|W\|_1 = \sum_{i=1}^{n} |w_i|$$

$$w_i = w_i - \lambda\frac{\partial loss}{\partial w_i} - (\gamma \cdot sign(w_i))$$

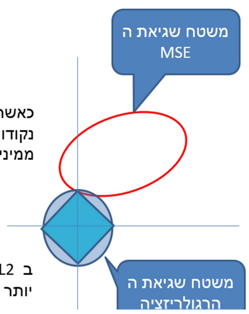
מדוע הרגולריזציה עובדת?



L1+L2 – Elastic Net

שילוב של Ridge ושל Lasso, מן ממוצע משוקלל.
במשקולות קטנות ה-Lasso ישפיע יותר.
במשקולות גדולות ה-Ridge ישפיע יותר.

בגלל שהשגיאה תלויה בריבוע הביאס (פרבולה)
העליה בשגיאת הביאס היא איטית בהתחלה ואח"כ
מתדרדרת בקצב גדול יותר



כאשר הפינה של יהלום L1 פוגעת המשטח ה MSE
נקודות אחרות על ביהלום (כאלו בעלי אותו עונש) מתרחקות
ממינימום ה MSE עוד יותר.

38

ב L2 נקודות (שאינן פינות) על מעגל ׳האילוץ, יתקרבו עוד
יותר למרכז האליפסה (ולכן לא תיבחר פינה)

PCA _ Principle Component Analysis

טרנספורמציות לינאריות ממימד n למימד p<n המנסות לשמר כמה שיותר אינפורמציה באמצעות **קומבינציות לינאריות של המאפיינים המקוריים**

הוקטור העצמי הראשון שיבחר הוא הוקטור שממקסם את השונות של הנקודות (במקרה שלנו, הקו הירוק)

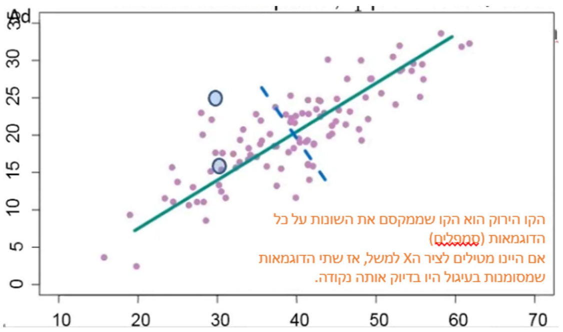
מכיוון שהוקטורים אורתוגונליים אחד לשני (ב90 מעלות \ בלתי תלויים לינארית)
הוקטור הבא חייב להיות הכחול (כי אנחנו ב2 ממדים, אין עוד דרגות חופש)

אינטואיציה:

1. אנחנו כל פעם מוצאים שמייצג את המידע בצורה הכי טובה (ימזער את ה MSE\ימקסם את השונות)
2. אנחנו מטילים עליו את הנקודות
3. מתייחסים אליו ולנקודות עליו כציר חדש, ועוד פעם חוזרים ל1

הסתייגות:

אנחנו מוצאים את כל הוקטורים העצמיים בפעם אחת.

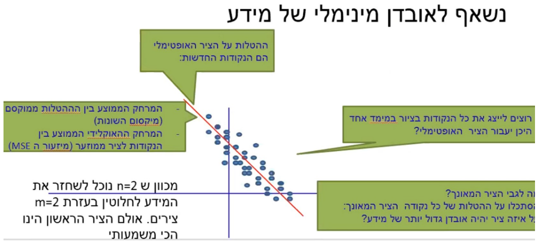


$$z_1 = \phi_1 x = 0.839\overbrace{(pop - \overline{pop})}^{\text{פיצ'ר 1}} + 0.54\overbrace{(ad - \overline{ad})}^{\text{פיצ'ר 2}}$$

$$\phi_1 = \begin{bmatrix} 0.839 \\ 0.54 \end{bmatrix}$$

הפחתת מימדים

בהינתן דוגמאות D={x} ממימד n
נסב דוגמאות אלו לנקודות במרחב ממימד m<<n
נשאף לאובדן מינימלי של מידע

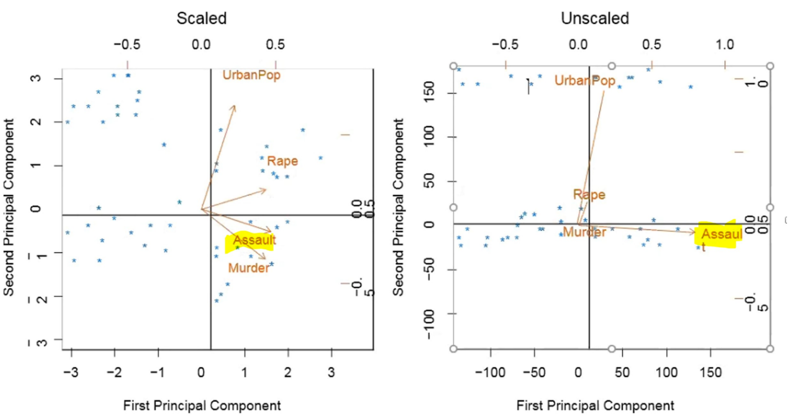


לדוגמה המרצה מ5 ממדים ל3 ממדים:

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} \qquad z_i = \sum_{j=1}^{n} \phi_{j,i} x_j \qquad \phi_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 8 \\ 7 \end{pmatrix} \quad \phi_2 = \begin{pmatrix} 8 \\ 3 \\ 5 \\ 9 \\ 5 \end{pmatrix} \quad \phi_3 = \begin{pmatrix} 5 \\ 4 \\ 2 \\ 2 \\ 2 \end{pmatrix}$$

1. מספר הטרנספורמציות הן כגודל הממד שאליו אנחנו רוצים להגיע
2. כל טרנספורמציה כזו (של כל טרנספורמציה Φ_n היא וקטור מהמימד הישן של הסמפלים n

PCA נותן תוצאות שונות בתלות ב Scaling (אשר משפיע על השונות)



לAssault יש שונות גדולה ולכן ישוקלל יותר מאחרים ב PC1
חלוקה של כל מאפיין בסטית התקן מונעת זאת
לחילופין, אם היינו מודדים Assault פר אוכלוסיה של 100 אנשים, השונות היתה קטנה משמעותית

כלל עדכון המשקולות ב GD:

$$MSE_h(w) = \frac{1}{2m}\sum_{p\in D}\left(t_p - y_p\right)^2$$

$$y_p = h_w(x_p) = \sum_i w_i x_{ip} + b$$

$$\Delta w_i = -\lambda \frac{\partial loss_D(w)}{\partial w_i}$$

$$\Delta w_i = -\lambda \sum_{p\in D} \frac{\partial MSE_D(w)}{\partial y_{ip}} \frac{\partial y_{ip}}{\partial w_i}$$

$$= \frac{\lambda}{m}\sum_{p\in D}\left(t_p - y_p\right)\frac{\partial y_p}{\partial w_i}$$

$$= \frac{\lambda}{m}\sum_{p\in D}\left(t_p - y_p\right)x_{ip}$$