9.1.19 בס״ד

חן [illegible]

203571278

מטלה מסכמת - מבחן סופי:

1. א. במהלך [illegible] (אותה [illegible] רק [illegible] (נק׳) את התוצאה.

הטוב ביותר שלי שהגשתי היא 0.75468.

כל השינויים והשיפורים אותם ניסיתי מופיע במחברת (html file)

המצורפת והנה:

- Feature engineering:

- [illegible] הערכים [illegible] ביחד [illegible] בהתאמה

ב [illegible] [illegible] רק ט, [illegible] [illegible]

[illegible] [illegible] ה׳ [illegible] ה-15 עד ה-14

[illegible] [illegible] וכך הלאה.

- [illegible] [illegible] ה [illegible]

- [illegible] [illegible] באותו יום / שבוע [illegible] ל- y

- [illegible] " " " " [illegible] ל- y

- [illegible] [illegible] [illegible] 1/5 [illegible] יום [illegible] ל- y

- [illegible] של 0/1 [illegible] [illegible] [illegible] או לא

[illegible] אותם [illegible] [illegible] [illegible] [illegible]

- linear model:

- הרצנו את המודלים עם רגולריזציה L1, L2

וללא רגולריזציה.

- כאשר השתמשנו עם L1/L2 היה צורך למצוא את

α/λ לכן ביצענו 10 Fold C.V על ה- train

data כדי למצוא ערך אופטימלי של α/λ

(כאשר ניתן לראות את הגרפים במחברת)

KNN:

- ביצענו KNN באמצעות שני מרחקים שונים האחד

הינו cosine similarity והשני הינו מרחק אוקלידי.

- כמו כן על מנת למצוא את ה- k

9.1.19 המשך

האופטימלי מצאנו ע"י C.V. 10-Fold (התוצאות במחברת)

- Boosting:

- ביצענו ניסויים בשלושה מודלים של Boosting, כאשר הראשון הינו gradient boosting הרגיל ואחר למפקנו והשני הינו מודל בשם light gradient boosting

d. <u>ניתוח התוצאות ותיאוריה סטטיסטית:</u>

- linear models:

- ניתן לראות כי התוצאות של המודלים בעלי הרגולריזציה אכן משפרים את התוצאות כאשר ניתן לראות כי מודל Lasso הינו המודל הטוב ביותר (גם מבין כל שאר המודלים)

- ניתן להבין זאת ע"י האינטואיציה שמאחורי מודל - ה- 200 פיצ'רים בקירוב כאשר בוודאי חלקם הם אינפורמטיביים וחלקם פחות. לכן רגולריזציה ע"י שימוש כאשר יש ה- L1 וגם ה- L2 מורידים הרבה של ה- variance

- KNN:

- במודל ה.ל. גרוע מאוד שמאמר זה. נראה הגיוני מכיוון שבגלל כי הן קללת המימד תפגע ביכולת המודל ושם המודל עצמו ה.ל. אינו מודל כל מספיק, שמעיד שאנו קשרים מורכבים בתוצאה. במיוחד ה Data שיש לנו.

- הפיצ'רים של הניסיון של Cosine אינו הרבה יותר לו. במצב כי מכיוון שכנראה יוצר שתמיד חצה 5 ויצר שתמיד חצה. 1 יותר בגלל Similarity 1 כמעט יותר הדומים ע. של זה.

- Boosting:

- ניתן לראות כי בהמון מקרים (לפחות אצלי כל המנייר) המודלים הללו מנצחים רבים, אפילו אלה. שוים

המודלים הנבחרים.

- ניתן לראות כי lgB בעל תוצאות טובות יותר מאשר grad boosting, כמו המודל המשמעותי ביניהם היא שבה הם נעצות ה-Boosting מתבצע קיבוץ של התכונות בעלות הערכים הנמוך ביותר בוודאי יחד, כך שמתמקד בתכונות בעלות ערכים גבוהים.

מסקנות ודיון כללי:

- הגישה שהוצגה ע״י הרצת המודל הינה גישה פשוטה. פחות הסבירה דומה, שהיא ״דגלו״ המון פיצ׳רים בחשבון יחד והעריכה מוקדם את הפרדיקציה הנכונה. בסופו של דבר, שבה ה.י. דומה לפני נוע יותר.
- חוץ מזה, נוספת שהגענו היא הרבה מהירות של הרצת המודל, כמו הוני בעיות מצד לבצע LOOCV של מנת לשדרג את הדיוק. בפחות מגבלות יותר, דרוש אתה, דרש המנבאים בגודל מקסימלי.
- בסופו של דבר הצלחנו את המילוי ערכים חסרים בצורה נכונה, היא כי עוזרים בכל המטריצה עם אפסים במקום ערכים חסרים, ולכן להשתמש במידע הגלוי ברמת Y.
- ניסיון כושל שלנו להגדיל מיצוי פיצ׳רים היה לנסות להעביר את ה-set של לפני (אותו dev set שהשתמשנו עבור ה-lgB) בתוך הפונקציה של ה-test. כאשר כך מתברר שאנו ליצור של הוא יש לעבור "בפועל" ל-dev-set במקום ה-train. אך גם בשלב זה ניתן לראות מסר אמיתי כיצד ודאי ההשוואה של קישור בין התוצאות ה-test ל-dev ניתנה ללא מסר אמיתי.
- בסופו אני חושב שלא הצלחנו מספיק, ניתוח של המודלים, כמו אני ע״י בחינה שהשתמשנו כיתה ע״י בסיס היינו מגיעים למסקנות מורכבות שהמודלים לפני שאנו דרש פיצ׳רים מדויקים על תרומות יותר נכונים של המודלים ב-Boosting.

9.1.19 בס״ד

הצעה למודל:

במודל אני חושב שלא הצלחנו (אולי גם המנצחים) לתפוס את
המידע המלא הגלום בתווים ובייחוד העובדה שהיומיים
לפני קרובים מ"סגמנט" של הניקוד, כמו שלפי ידיעתי שנו
במדויקות, לכן אני מציע שיטת המקבל בין באמצעות
מתוך Bidirectional RNN, כאשר נוכל להבין קרובים לרשת
הן מהתחילה והן מהסוף ובכך לתפוס את ההתייחסות
בזמן. במודל הבא, כי שימוש עבודת תכונות ובה מכיון
שנועד יותר אחרת באמצעות Back Prop גם
כדי מעורבות מאוד (יותר שקוף את הסדר Y ראשון
מתוך יותר שנה הידע הקודם ה-99).
כאשר אפשר ייצוג מוכן להציג embedding מוכן
של שימוש (סוכנים בעזרת מוכרים מאומנים Word2Vec
ובכך לנסות לראות האם ניתן להבין קשיין בין
סוכנים (או ויזואליזציה של סוכנים) בעזרת אשר כללים.

12.1.19

בס"ד

1. (1) Option:

בשלב ראשון את בדיקת MNIST / Zipcode Data ואת

תוצאות בשלושה מודלים:

- svm
- NN בעל שכבה אחת בלבד ללא רגולריזציה.
- CNN בעל מספר שכבות ורגולריזציה.

* כל התוצאות והקוד נמצאים בתיקיות המצורפות

__התוצאות העיקריות:__

- SVM: את המודל הזה הרצנו עם הפרמטרים הבאים
  - גרעין rbf
  - tol / ε = 0.001
  - one vs rest
  - $gamma = \frac{1}{n\_features}$

כפי שניתן לראות בבירור, ה-svm נתן תוצאה טובה מאוד עם דיוק של 94.2% על ה-test ודיוק של 99.2% על ה-train, מה שמצביע על בדיוק overfitting.
ניתן לראות מה-confusion matrix כי הספרה 5 היא הכי קשה, אך גם ניתן לראות הרבה טעויות בין הספרה 2 ו-3 והטעויות בין הספרה 3.
בנוסף כפי הנראה שלי ניתן לשפר את התוצאות ע"י כיוונון הפרמטרים בצורה טובה יותר.

- NN: בתוצאות ה-NN השכבה היחידה ללא רגולריזציה.
כי הבעיה מאוד ... שלמדו תוצאות ... ויש סיכוי מוגבלות ולרגולריזציה ... תוצאות טובות יותר. ... שכבה אחת עם 3500 נוירונים:

input	hidden layer	output
256	3500	3

כאשר השתמשתי בטופולוגיה. עם relu ופונקציית output של softmax,

ניתן לראות בבירור יפה כי הדיוק ב train הגיע ל 100% לעומת ב- test שהדיוק היה 93.7%. התוצאות לא רעות אך ניכר שאנחנו ב- overfitting.

- CNN: בשלב הבא במקום לעבוד רק עם רשת מחוברת ולא מנצלת את המבנה שמתקבל בתמונות, בחרתי בארכיטקטורה הבאה:

input → conv2D → relu → conv2D → relu → max pool →

→ Drop out (0.25) → conv2D → relu → conv2D → relu → max pool

→ Drop Out (0.25) → flatten → Dense → Drop out (0.5)

→ Dense → softmax

• כמובן שבאופטימיזציה נוכל לשפר תוצאות מ הארכיטקטורות הקודמות בקורס.

בגלל האופטימייזר השתמשתי ב- RMSprop, כמו האדם סוג עדכון, שיפור קטן מאוד.

בנוסף לכך היו ניסויים נוספים על כמות הלייארים כמות נוירונים, שינוי augmentation ו- Padding שונים.

כעת ניתן לראות כי הדיוק על ה-train הגיע ל-92.47%
ודומות דיוק על ה-test של 98.09%. כמו כן ניתן
לראות מהגרף שהפעלת הכניסות הרגולריזציה עזרה מאוד
משמעותית (ניתן לראות כיצד שיפור הגרף ב- קיבלנו עזר גבוהה יפה!)
ניתן לראות ב-confusion כי מרבית הטעויות נעשו
שהספר 5 נחשב ה-3.
כמו כן, מטריצת הטעויות הפשוטות כמו ספרות שדומות זו לזו
בוקה כאשר הדגמתי בין ה-one hot vec לספרות הן:
5:2 , 3:1 , 2:0
ניתן לראות שיש פה טעות שכל' מובנת שאני ע"י אעשה!

<u>דיון סופי:</u>

- המודל ולא הגדרות בערך היו MNIST !!!
- ניתן בוודאות פרמטרים לשפר את התוצאות אך כמובן
  צריך לשים לב שמאוד שמגבלת השפעת את הגנרליזציה.
  במקביל ל- test (מה שהיה צריך בהתחלה) נראה
  כמו overfitting ל- test.
- ניתן לראות כי באמצעות מספר קטן של איטרציות מוקדמות
  בשלב (כאשר הדיוק תחת רשת קטנה יחסית) ניתן לשפר את
  התוצאות בצורה מאוד משמעותית.
- אני מניח שאם הייתי מגדיל את מספר ה- epoch
  וניתן רואים שהאימון עוד היה ממשיך לשפר את התוצאות
  עוד.
- מעניין היה לבחון "כמה קשה" את אותה משימה עם רשת
  את השגיאה תמיד!
  - שינוי הפרמטרים; פרמטרים ל- Adam / GD
  - שינוי הלמידה; מומנטום (פרמטר מתאים שינוי)
  - שינוי Augmentation: RF ועוד
  - שינוי ה- Drop Out

10.1.19 נוב

2.

a.

i. X גדול מאפיינים קטגוריאליים רבים, בעל הרבה ערכים חסרים ועדין, הוא מאוזן יחסית:

התייחסות כללית:

- ערכים קטגוריאליים רבים; במידה ויש הרבה, כיכרים של ערכי קטגוריות ויכולה להיות שבהם מאפיין. ככל יש הרבה ערכים שונים זה ואנחנו מסכנת variance/overfitting גבוהה. (כגון: קוראה של מיון מחוזים ע"י הוון שבינה בנימה.)
- ערכים חסרים; נניח: מינים לחשוב על מודל שיודע להתמודד עם ערכים חסרים
- קיום הסובה: מניח, שקיום, חשוב ננא בגיית סייג ו- Balance Data. ואם נבצע model selection נשתמש בקיום, כמטריקה להחלטה על הערכת המודל

הערה: נניח כי לכל מאפיין קטגוריאלי אין הרבה ערכים אפשריים כי אחרת נקבל גודל סטטיסטית גבוהה, מאוד קשה מאוד (במידה וכן ניתן לטפל בזה ע"י bucketing של ערכים לפי מעמה של Domain knowlege לפי)

מודל מתאים:

- gradient boosting; נשתמש במודל זה עם עצי החלטה מסוג גרדיאנט ו- E-Boosting. מודל זה מתמודד עם עצים ואף יודעים להתמודד עם מאפיינים קטגוריאליים ועם עם ערכים חסרים.
- ערכים חסרים שימוש ב- surrogate splits או בשיטות עזר נוסף שבו הם בהתאם לשימוש,
- ערכים קטגוריאליים, שימוש 2^{Q-1} חלוקות אפשריות

השימוש במודל כזה יאפשר לנו התמודדות עם הבעיה בנתונים והוגנים לבין זמה ולא יאפשר לנו להתמודד עם בעיית ה- Variance מכיוון שבאימפלמנטציות התאמת מודל שונה. לכל ג'ורג'ינג נוסף יטויצ' את ה-Variance) ולא רק את ה- Bias.

- cat Boost (categorical Boosting); זהו מודל שלא למדנו במסגרת הקורס ומוצג לפני כחצי שנה ע"י yandex. מודל זה מתמודד בכלל מיני ווריאציות קטגוריות המשתנים הקטגוריאליים אשר מביא תוצאות טובות בהשוואה למודלים: ומאפשר "ירוד" של זמן הריצה.

מודל שאינו מתאים;

- KNN; בניגוד לכל המודלים שהוצגו עד כה, יתקשו להתמודד עם ערכים חסרים (למרות שיש שיטות להתמודד עם זה, כמו ע"י השמה ושונות של הערכים החסרים) וערכים קטגוריאליים, כאשר ניתן לטעון כי אין באמת משמעות למרחק אוקלידי (או כל מרחק אחר) בין קטגוריות שונות.

ii. ג בגלל מספרם הקטגוריאליים רבים, בגלל העובדה ערכים חסרים ומאפיינות בניית המודל היו קשות.

התאמה כללית:

- סבב: ערכים קטגוריאליים וחסרים ג.ה. לסוגי ויקים נכל לצאת ג.ו.נחה.

- מורכבות בניית מודל: נדרש אופן ח"ל לאת"ם לבנך מכיוון שהמשתנים הקטגוריאליים יקרים שמשל עמוד רבים חסויות ולכן אשר מודפות על זמן הריצה.

מודל שאינו מתאים;

- gradient Boosting - מכיוון שבגלל מורכבות בניית המודל חשובה לנו מודל כזה שלא יתאמן במשך זמן

10.1.14 ג.ב.

בממוצע, ומאחר שזו בעיית רגרסיה, אז עבור מינימום את ה
האיבר בעלות. כי ניתן לחלוקות שטחוניות המיון לכל
רק. הזמן $O(Q\log Q+n)$ כאשר יש לחשב אנו את
התחזיות של הערכים האחרים אשר יניב זמן
ריצה תקין בכל הגבלה לפרק זה. אם ידעו לנו
במיצה ומחירות ריצה ושאלה זהו.

מודל מתאים (בהינתן רגרסיה):

- linear model with Lasso:
- ערכים קטגוריאליים: נשתמש ביצוג באמצעות וקטור one hot encoding, כלומר נגדיר משתנה dummy עבור כל ערך של קטגוריה אם לא משתנה.
- ערכים חסרים: אם כן, ערך חסר לא משנה קטגוריה אז נגדיר קטגוריה של חסר. אם כן ערך חסר בממוצע, ונציב נוכל להשתמש בהסתברות בממוצע מה שלא מתחרות (למשל כמו השונה שהוצגה ע"י המנצחים)

כאשר בגלל כי נציב כאן להשתמש בממוצע גם את המקדמים ע.י. חיזוי לינארי עם רגולריזציה מתאימה בזמנות של Lasso, בנוסף נוכל להתמודד עם בעיית ה-variance (של מקדמים גבוהים מאוד). כאשר מודל זה יתאים זמן ריצה. והזמן ידוע כאשר זמן הריצה של Lasso תוך שימוש ב- LARS הינו $O(p^3+p^2n)$

- CatBoost: כי מקבל קטגוריות, ניתן להתמודד במודל זה. וגם מבניין שאינו לשמור מקום רק אשר צריך.

iii. n גדול, p קטן ומניחים שיש תלות מורכבת בין x ל-y:
ננייח התלויות סבוכות עדיין רבות.

מודל מתאים:

- NN (Deep learning): ברור כי מודל מספיק לאת כריונים יותפקו Gol

10.1.19 חוג

ברשתות משל תבניות גדול מאוד ישנן מעגלים בחלקן
כאשר נוכל לשנות את הארכיטקטורה של הרשת בהתאם
לבעיה. הפונקציות, ע"י כמות השכבות, סוג השכבות
(מלאות, CNN, RNN: Fully connected, CNN, RNN), סוג האקטיבציה, שכבת ה-output
(sigmoid עבור $k=2$ ו-softmax עבור $k>2$)
וכמובן גם ההפסד שמתאימה לבעיה (למשל cross entropy).
בנוסף נוכל להשתמש ברגולריזציה. על מנת לאזן
את הבייאס וה-variance שנוצרת לנו
ב L2 ו- dropout. מודל זה יותאם לנו
ללמוד תלויות מורכבות בין x לבין y מתוך
למידה של הרשת עצמה (האופטימיזציה).

- SVM עם kernel machines: כמו האלגוריתם במודל זה
הוא מאפשר את מימד המאפיינים הגבוהים
במימד גבוה מאוד (ובכך ללמוד תלויות מורכבות)
ואז נבחר מודל ליניארי. כמובן שבחירת מימד גבוה
מאוד עלול להביא להתאמת יתר, ולכן נשתמש ברגולריזציה, למשל כמו שראינו
ברשת L2 וכמובן שנבחר גם הפסד שמתאים, למשל
כמו הפסד לוגיסטי.

מודל שאינו מתאים:
- מודל פשוט: בדומה לכל מודל ליניארי שאינו
מסוגל ללמוד תלויות מורכבות בין x
לבין y מכיוון שהוא מודל ליניארי במימד p
המקורי.
לכן, n גדול, p גדול ויחס מורכבויות חשובים.
מודל מתאים (ננסה בעיית רגרסיה):
- linear model with subset selection: בדומה לבעיה הקודמת אנו מאמינים כי יש
תלות מורכבת בין התכונות לבין y שגם ניתן לבחון
מודל בגודל subset selection, כמו ובנוסף מודל
המש

10.1.14 (המשך)

בשכנויות K שכנים מעטים יחסית. כאן ננסה לנבא את השכנים שהכי קרובים לחיזוי y. וכאן יקטין את ה- variance.. אשר הפתרון של שאלה זו הוא תלוי בגודל המשתנה את K.

- linear model with lasso: ראשון! רגולריזציה L1 תאפשר לנו לבחור מודל ליניארי עם פחות מקדמים שווים ל-0 מודל. וכאן ניתן להמיר מודל כמו כן ה משתנים מספר. כלל החיזוי (אלו שהמקדמים שלהם לא שווים ל-0) וכאן ננסה להתמקד במשתנים שהם חשובים. נמוך שכן המטרה הזו מציאת S.

* נמוך שהוא מציאת K ו-S נבצר בשכנויות כ.ל.

מודל שהוא מתאים:

- linear model: השני! כי שאלה בשאלה 4 מודל ליניארי ללא רגולריזציה, יש לנו כאן בעיה interpolation וכאן נקבל מודל שה variance שלו גבוה. כמו כן המשתנים יהיו חלק בחיזוי של y למרות שחלקם לא חשובים.

(b) נתון הנחות בכל אחד מהמודלים הבאים:

- LDA:

נניח כי ה-X שלה נורמלית ולכן מניחים לה שונויות:

$$P(y=g_k \mid \underline{x}=\underline{x}) = \frac{Pr(\underline{x}=\underline{x} \mid y=g_k)\cdot Pr(y=g_k)}{\sum_{l=1}^{K} Pr(\underline{x}=\underline{x} \mid y=g_l)\cdot Pr(y=g_l)} = \frac{f_k(\underline{x})\cdot \pi_k}{\sum_{l=1}^{K} f_l(\underline{x})\cdot \pi_l}$$

מוגבלת.

כאשר נניח כי צריך לחזות, צריך להעריך את הצפיפות f_k וכן את π_k כאשר Pr הנכון ה-K

10.1.19 בס"ד

כאשר ב- LDA הנחת הנורמליות מוצדקת אזי הוא נכון

$$\underline{x} \mid y = g_k \sim N(\underline{\mu}_k, \Sigma)$$

כאשר נרצה לסווג לקטגוריות k את ... נקודה $\underline{x}$

נשאל אם

$$Pr(y=g_k \mid \underline{x}) > Pr(y=g_1 \mid \underline{x})$$

כאשר מתקיימות ההנחות הללו מקבלים כי גבול ההחלטה הוא:

$$\underline{x}^T \Sigma^{-1}[\underline{\mu}_k - \underline{\mu}_1] - 0.5\,\underline{\mu}_k^T \Sigma^{-1} \underline{\mu}_k + 0.5\,\underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 + \log\left(\frac{\pi_k}{\pi_1}\right) = 0$$

- linear regression:

הנחת המודל היא: $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$

$E(\underline{\varepsilon}) = 0$, $Var(\underline{\varepsilon}) = \sigma^2 \cdot I$

כאשר מתקיימות הנחות אלו לפי משפט גאוס מרקוב נקבל שהאומד OLS הינו האומד BLUE.

אם אנו נוסיף את הנחת הנורמליות עבור ε:

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

נקבל שוויון בין אומדי OLS לאומדי נראות מירבית ונקבל שהאומד שלנו הוא UMVUE.

כאשר מתקיימות הנחות אלו, ניתן לומר מהי התפלגות האומדים:

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

ובעזרת זה ניתן לבנות מבחנים על $\hat{\beta}$ (מבחני השערות)

- linear regression with ridge:

בתרגיל 2 (שיעורי הבית) שאלה 2 סעיף (a) הראינו כי אומד ridge שקול לאומד בייסיאני עם ההנחות הבאות:

$$y_i \sim N(x_i^T \beta, \sigma^2)$$

אם ניקח את ה prior:

$$\beta \sim N(0, \tau^2 \cdot I)$$

נקבל $\lambda_{ridge} = \frac{\sigma^2}{\tau^2}$

כלומר הראינו כי $\beta_{ridge}(\lambda)$ הוא ה - mode

אם נוסיף הנחות נוספות על ההנחות שהנחנו, נקבל

שמהנחות שונות מתקבלות התפלגויות ה mode הוא

המתאים.

ואלו הן ההנחות הנוספות שמתקבלות ב- ridge.

10.1.19 פתר'

?. ridge Lemma LOO:

(a) יהיו n דגימות וקטוריות עם רגולריזציה L2 כלומר ridge
מקדם λ

$$\hat{\beta}_{ridge}(\lambda) = (\underline{X}^T \cdot \underline{X} + \lambda \underline{I})^{-1} \underline{X}^T \underline{y}$$

כלומר

$$\hat{\underline{y}} = \underline{X}\,\hat{\beta}_{ridge}(\lambda) = \underbrace{\underline{X}(\underline{X}^T\underline{X}+\lambda \underline{I})^{-1}\underline{X}^T}_{S(\underline{X},\lambda)} \cdot \underline{y}$$

נסמן $S(\underline{X},\lambda)$

לא תלוי
תלות
fixed $\underline{X}$

ולכן

$$\hat{\underline{y}} = S(\underline{X},\lambda)\cdot \underline{y}$$

(b) נראה כי את ה-LOO לפני שם המקדם מנ"מים,

פתרון ridge לבעיה כאשר התצפית ה- i לא מוגדרת
להיות

$$\hat{\underline{\beta}}^{(-i)}(\lambda) = \underset{\beta}{\operatorname{argmin}} \sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - \underline{x}_j^T \underline{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

נסמן:

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - \underline{x}_j^T \hat{\underline{\beta}}^{(-i)}(\lambda))^2 + \lambda \sum_{j=1}^{p} \hat{\underline{\beta}}^{(-i)2}(\lambda) = RSS(\lambda)$$

נניח בשלילה כי $\hat{\underline{\beta}}^{(-i)}(\lambda)$ אינו הפתרון ridge עבור המודל
$\tilde{T}$ אלא $\hat{\tilde{\beta}}(\lambda)$ הוא הפתרון של הבעיה מינימום
עבור $\tilde{T} = (\underline{X}, \tilde{y})$ כי

$$\underset{\text{מינימום על } \tilde{T}}{RSS(\hat{\tilde{\underline{\beta}}}(\lambda))} \le \underset{\text{מינימום על } \tilde{T}}{RSS(\hat{\underline{\beta}}^{(-i)}(\lambda))}$$

10.1.14

המשך

כעת ננסה לחשב את $RSS(\hat{\beta}^{(-i)}(\lambda))$ על $\tilde{T}$

$$RSS_{\tilde{T}}(\hat{\beta}^{(-i)}(\lambda)) = \sum_{\substack{j=1 \\ j\neq i}}^{n} [\underbrace{y_j}_{\tilde{y}_j = y_j} - x_j^T \hat{\beta}^{(-i)}(\lambda)]^2 + [\underbrace{y_i}_{\tilde{y}_i = \hat{y}_i^{(-i)}} - \underbrace{x_i^T \cdot \hat{\beta}^{(-i)}(\lambda)}_{= \hat{y}_i^{(-i)}}]^2$$

משום השיוויון

$$+ \lambda \sum_{j=1}^{p} \hat{\beta}^{(-i)}(\lambda)^2$$

אז:

$$RSS_{\tilde{T}}(\hat{\beta}^{(-i)}(\lambda)) = RSS^{(-i)} + [\hat{y}_i^{(-i)} - \hat{y}_i^{(-i)}]^2 = RSS^{(-i)}$$

כעת תחת הנחת השלילה, נקבל כי:

$$RSS_{\tilde{T}}(\tilde{\beta}(\lambda)) \leq RSS_{\tilde{T}}(\hat{\beta}^{(-i)}(\lambda)) = RSS^{(-i)}$$

אבל!

$$RSS^{(-i)} \geq \underbrace{RSS_{\tilde{T}}(\tilde{\beta}(\lambda))}_{\text{על } n \text{ תצפיות}} \geq \underbrace{\sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - x_j^T \tilde{\beta}(\lambda))^2 + \lambda \sum_{j=1}^{p} \tilde{\beta}_j^2(\lambda)}_{\text{הנוסף על } n-1 \text{ תצפיות}}$$

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - x_j^T \hat{\beta}^{(-i)}(\lambda))^2 + \lambda \sum_{j=1}^{p} \hat{\beta}^{(-i)}(\lambda)^2 \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - x_j^T \tilde{\beta}(\lambda))^2 + \lambda \sum_{j=1}^{p} \tilde{\beta}_j^2(\lambda)$$

אבל זה בדיוק אומר ש- $\hat{\beta}^{(-i)}(\lambda)$ אינו הפתרון שממזער

את פונ׳ ההפסד כאשר התצפית i לא במדגם ה-train

בסתירה שהנחנו ש- $\hat{\beta}^{(-i)}(\lambda)$ הוא הפתרון שממזער את.

הנוסף לכן הוכחנו שמ״ל. לכן

$\hat{\beta}^{(-i)}(\lambda)$ הוא גם הפתרון עבור $\tilde{T}$ ולכן מתקיים כי $\tilde{y}_i = \hat{y}_i^{(-i)}$

10.1.19

ומכאן הוכחנו ש פתרון Penalized ridge מקיים את שני התנאים של הלמה, ומכאן ניתן להשתמש בה עבורו.

(ג) מכיוון שהמטריצה $S(\underline{X},\lambda)_{ii}$ (האיבר במיקום) הוא מספר בין 1 ל- 0 בגלל כי נתון ש- $S(\underline{X},\lambda)_{ii}$ יקטן וכן $1-S(\underline{X},\lambda)_{ii}$ יגדל וכך עבור ה

$$\left(y_i-\hat{y}_i^{(-i)}\right)=\frac{\left(y_i-\hat{y}_i\right)}{1-S_{ii}(\underline{X},\lambda)}$$

יקטן בערך מוחלט, כלומר:

$$LOOCV=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i-\hat{y}_i\right)^2}{\left(1-S_{ii}(\underline{X},\lambda)\right)^2}$$

יקטן.

כלומר ככל ש- $S(\underline{X},\lambda)_{ii}$ יקטן וכך גם ה- LOOCV יקטן.

מכיוון שה- LOOCV יקטן וכך יקטין הפער ה המרחק, בין שגיאת האימון לבין ה- ...

LOOCV squared error for ridge

יקטן מכיוון ששגיאה זו היא מעין "ניפוח" של שגיאת האימון.

$$W=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i-\hat{y}_i\right)^2}{\left(1-S_{ii}(\underline{X},\lambda)\right)^2}-\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

תלוי ב- $S_{ii}(\underline{X},\lambda)$

ומכאן ש W שהוא המרחק בין שגיאת האימון לבין שגיאת ה- LOOCV יקטן.

מש"ל

17.1.19

כעת נסביר מדוע אנו מצפים כי $S_{ii}(\lambda,X)$ ירד.
באופן כללי, ידוע של λ תפקיד רגולרי ישירות
אינטואיטיבית, ה- λ של $S_{ii}(\lambda,X)$ משפיע על ה optimism כמעט
ישירות ה- λ יכולה להקטין את $\hat{y}_i$. ולכן ככל שגדל
λ משמעה גדל הערך רגולריזציה ולכן משמעותה גם
של הרגולריזציה. ה- λ מקטין את ה optimism. כלומר
כאשר $\lambda \to \infty$. הפרמטרים שואפים ל-0 $\hat{\beta}_{ridge} \to 0$
ועל כן כי ה- optimism גדל כלומר, ולכן
מכאן ה- optimism. ולכן ככל שגדל, $\lambda \to 0$ optimism מוגבל.
משמע ש- $S_{ii}(\lambda,X)$ הינו פונ' יורדת של λ.
מכיוון שה- optimism הוא פונ' עולה ב- $S_{ii}(\lambda,X)$
לכן כשמגדילים λ יורדים את $S_{ii}(\lambda,X)$ וכך
ה- optimism.

(ד) נראה מתמטית ש- $S_{ii}(\lambda,X)$ היא פונ' יורדת ב-λ:

$$S(X,\lambda) = X(X^TX+\lambda I)^{-1}X^T = S(\lambda,X)$$

נעזר בפירוק ה- SVD:

$$X_{n\times p} = U_{n\times p} D_{p\times p} V^T_{p\times p}$$

$$* \quad X^TX = V\cdot D^T\cdot U^T\cdot U\cdot D\cdot V^T = VD^2V^T$$

$$* \quad X^TX+\lambda I = VD^2V^T + \lambda V\cdot V^T = V(D^2+\lambda I)V^T$$

$$* \quad (X^TX+\lambda I)^{-1} = V\,\mathrm{diag}\left(\frac{1}{d_j^2+\lambda}\right)V^T$$

$$S(X,\lambda) = UDV^T\cdot V\,\mathrm{diag}\left(\frac{1}{d_j^2+\lambda}\right)V^TVDU^T$$

11.1.18

$$= U D \cdot \operatorname{diag}\left(\frac{1}{d_j^2+\lambda}\right) D \cdot U^T$$

$$= \underset{n\times p}{U} \cdot \underset{p\times p}{\operatorname{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)} \cdot \underset{p\times n}{U^T}$$

$$= \begin{pmatrix} | & & | \\ u_1 & \cdots & u_p \\ | & & | \end{pmatrix} \cdot \begin{pmatrix} \frac{d_1^2}{d_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{d_p^2}{d_p^2+\lambda} \end{pmatrix} \begin{pmatrix} - u_1 - \\ \vdots \\ - u_p - \end{pmatrix}$$

אז: ניתן לראות כי

$$S(\bar{X}, \lambda)_{ii} = \sum_{j=1}^{p} U_{ij}^2 \cdot \frac{d_j^2}{d_j^2+\lambda}$$

וכמובן שניתן לראות כי ככל ש- λ גדל, [illegible]
[illegible] . [illegible] ש- λ [illegible]
[illegible].

פורמלית ניתן לכתוב:

$$\frac{\partial S(X, \lambda)_{ii}}{\partial \lambda} = \sum_{j=1}^{p} \frac{-U_{ij}^2 \cdot d_j^2}{(d_j^2+\lambda)^2} < 0$$

כלומר אכן [illegible] ככל λ.

[illegible]

[illegible] $S(\bar{X}, \lambda)_{ii}$;

$$w = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i-\hat{y}_i)^2}{(1-S_{ii}(X,\lambda))^2} - \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

1-ל 0 בין

אם $S_{ii}(X, \lambda)$ [illegible] $(1-S_{ii}(X,\lambda))^2$

(191)

11.1.14 גית

יעיל ונוסחה

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{(1-S_{ii}(\underline{x},\lambda))^2}$$

יקר: ב- $W_{ii}(\underline{x})$ (מניין שנמצאו במטבע.) תוכן לא W

יקר: ב- $S_{ii}(\lambda,\underline{x})$

* בזה ניתן לבחור סדק סימולציה להשוות בעזרת $\hat{y}$:

- יצירת רשימת ערכים של λ; range(0, 1000000) seq
- בניית מודל רגע לכל מתוך λ
- חישוב השגיאה של גבי ה- train לכל ערך λ

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

- חישוב שגיאות ה- LOOCV לכל ערך λ

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{(1-S_{ii}(\underline{x},\lambda))^2}$$

- חישוב הגרף של ערך של λ; $w(\lambda)$
- בחירת הערך $w(\lambda)$
- בשורה. סימולציה. ניתן לחזור גם את הקירוב של ה- GCV:

$$GCV(\hat{y})=\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{(1-tr(S)/n)^2}$$

(ק) נגדיר נרמול סולוציצית Lasso:

$$\hat{\beta}_{Lasso}=\underset{\beta}{\arg\min}\ \|y-X\beta\|^2+\lambda\|\beta\|_1$$

נוכל לבחון את הפתרון הממשי בתחום Lasso, והם גורם כי הם תת-חוקים מנוסים שאינם גאוסיים

(כ)

לא מתקיים עבור Lasso, כלומר לא ניתן להציג
רגרסיה לינארית כך ש-

$$\hat{y}_{Lasso} = S(\lambda) y$$

(ד) אלגוריתם את הגרעין, כלומר אופטימלי שגיאת KNN
לא תלוי את הנתוני אימון של הפונק' ולכן גם לא
מתקיים.
נחשב KNN:

$$\hat{y}_i = \sum_{\substack{j=1 \\ j:\, x_j \in N_k(x_i)}}^{N} \frac{1}{k} \cdot y_j = S(\Sigma) \cdot Y$$

$$S(\Sigma)_{ij} = \begin{cases} \frac{1}{k} & \text{if } x_j \in N_k(x_i) \\ 0 & \text{else} \end{cases}$$

כמו כן KNN מוגדר את הנתוני הראשון של הפונק',
אם נחזה, שמא אינו מוגדר את הנתוני אימון, בתחילה.
הפונק' נפח לא לפונקציה סטנדרטית;

$$\hat{y}_i^{(-i)} = \sum_{\substack{j=1 \\ j \neq i \\ j:\, x_j \in N_k(x_i)}}^{N} \frac{1}{k} \cdot y_j$$

כלומר $\hat{y}_i^{(-i)}$ הוא הממוצע של ה-k שכנים של x_i
הי' התצפית i נרצה את $\tilde{y}_i$ של מציגים
$\tilde{T}$

$$\hat{\tilde{y}}_i = \sum_{\substack{j=1 \\ j:\, x_j \in N_k(x_i)}}^{N} \frac{1}{k} \tilde{y}_j = \sum_{\substack{j=1 \\ j \neq i \\ j:\, x_j \in N_{k-1}(x_i)}}^{N} \frac{1}{k} \cdot y_j + \frac{1}{k} \hat{y}_i^{(-i)}$$

10.1.19

ג׳נ

נראה דוגמה כי לא בהכרח מתקיים שוויון בין

$$\hat{\tilde{y}}_i = \hat{y}_i^{(-i)}$$

דוגמה.

נתונה מערכת הבאה:

$$x = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} \qquad y = \begin{pmatrix} 0 \\ 50 \\ 100 \end{pmatrix} \qquad k=2$$

אזי נחשב עבור $i=2$;

$$\hat{y}_2^{(-i)} = \sum_{\substack{j=1 \\ j\neq 2 \\ j\,;\,x_j \in N_2(x_i)}}^{3} \frac{1}{2} y_j = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 100 = 50$$

נגדיר עכשיו כי

$$\tilde{x} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} \qquad \tilde{y} = \begin{pmatrix} 0 \\ 50 \\ 100 \end{pmatrix}$$

$$\hat{\tilde{y}}_2 = \sum_{\substack{j=1 \\ j\,;\,x_j \in N_2(x_i)}}^{3} \frac{1}{2}\cdot \tilde{y}_j = \frac{1}{2}\cdot 50 + \frac{1}{2}\cdot 0 = 25$$

ולכן

$$\hat{\tilde{y}}_i \neq \hat{y}_i^{(-i)}$$

לכן הטענה אינה מתקיימת באופן כללי.

11.1.19

גיליון

4.

$p \geq n$, $\underset{n\times p}{X}$

(a) X הינה מטריצה full rank.

מכיוון ש X שורותיה סטוכסטיות אזי full rank אזי
כל אחת מהשורות בלתי תלויות לינארית ולכן

$$rank(X) = \min(n,p) = n$$

כאשר הנתונים שהמטריצה מכילה דומות ממשיות לינאריות.

במצב כזה מכיוון שהמטריצה הינה היא

$$y = X\beta \qquad / \qquad L(y_i, \hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

מתקיים שיוכל מהמשוואות הנתונות n משוואות רק
p נעלמים לכל היותר.

לכן המודל שיש ומספר פתרונות שיחשב את המשוואות.
מצב כזה יגרום לכך שהמודל יחזה $\hat{y}$ מושלם
זהו כל תצפית ב training set.

כלומר מכיוון שיש לנו יותר נעלמים מאשר תצפיות
נוכל לייצר כל y_i בהתאמה ליניארית לינארית
של x_i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi} = y_i$$

כלומר נקבל $\hat{y}_i = y_i$ עבור כל תצפית במקרה
המיוחד שלנו ולכן בהכרח הפתרון במודל כזה
הוא האופטימלי, וכו'.

$$\sum_{i=1}^{n}(y_i - x_i^t\hat{\beta} - \hat{\beta}_0)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$$

(i) כמו שציינתי קודם הפתרון הוא יחיד מכיוון
שיש לנו מערכת רק n משוואות ו- p נעלמים
אך בגלל שיש אינסוף פתרונות למערכת המשוואות

הפתרון של n משוואות רק p נעלמים שהכל הוא מיון שהוא יחיד נכון. לינארית

11.1.19 המשך

לגבי דיון יש לנו מערכת של n משוואות עם $p+1$ נעלמים נוסף $n \geq p$ כלומר נקבל ולמעשה פתרונות.

(ii) הוכח כי המשוואה לא תשתנה עבור הפסד מוחלט או הפסד אחוזונים.

הוכחה: כי ניתן לסדר שמניין ש $n \geq p$ ניתן למצוא את y_i כקומבינציה לינארית של x_i לכן כי שגיאות מעבר עם הפסד מוחלט או הפסד אחוזונים לא תשנה את המשוואה הנ״ל, וכו׳.

הפסד מוחלט: $\sum_{i=1}^{n} |y_i - x_i^T \beta - \beta_0| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = 0$

הפסד אחוזונים: $\sum_{i=1 : y_i \geq \hat{y}_i}^{n} \tau (y - \hat{y}_i) + \sum_{i=1 : y_i < \hat{y}_i}^{n} -(1-\tau)(y_i - \hat{y}_i) = 0$

(b) רגרסיית [illegible] שגיאה;

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} (y_i - x_i^T \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$ [illegible]

נסמן את פתרון הבעיה $\hat{\beta}(\lambda)$ ונניח בשלילה כי הפתרון $\hat{\beta}(\lambda)$ הוא כזה שמוריד ... הנגזרת לפי β_j שלו;

$$\left. \frac{\partial \left(\sum_{i=1}^{n} (y_i - x_i^T \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)}{\partial \beta_j} \right|_{\hat{\beta}_j(\lambda)} = 0$$ אבל כי !!!

מכיוון שזוהי הנקודה בה יש מינימום, ולכן:

$$\left. 2 \sum_{i=1}^{n} (y_i - x_i^T \beta - \beta_0) \cdot x_{ij} + 2\lambda \beta_j \right|_{\hat{\beta}_j(\lambda)} = 0$$

11.1.14 גזור

כעת נגזור שוב נקבל את אותה משוואה אינטרפולציה.
המינימום עובר:

$$2\sum_{i=1}^{n}\left(y_i - x_i^T\hat{\beta}_{inter} - \hat{\beta}_{0_{inter}}\right)\cdot x_{ij} + 2\lambda\hat{\beta}_{j_{inter}}$$

$$\Downarrow$$

$$\hat{y}_i = y_i$$

$$= 2\cdot\sum_{i=1}^{n}\left(y_i - \hat{y}_{i_{inter}}\right)x_{ij} + 2\lambda\hat{\beta}_{inter} = 2\lambda\underbrace{\hat{\beta}_{j_{inter}}}_{>0} \neq 0$$

כאשר גוזרים לפי מתקיים $\hat{\beta}_{j_{inter}}=0$ לכל j (מנוגד)

שוב כי לפי ה.ו. יכול להיות נקודת אינטרפולציה (כי לפי יחיד עבור $\hat{\beta}$ אלפים רבים של נקודות.)
לכן ובפרט סתירה, כו',

$$\hat{\beta}(\lambda) \neq \hat{\beta}_{inter}$$

ב. נשים:

$$\hat{\beta}^{(l2)} = \arg\min\sum_{j=1}^{p}\beta_j^2$$

$$s.t$$

$$\sum_i\left(y_i - x_i^T\beta - \beta_0\right)^2 = 0$$

כעת נחזור לבעיה המקורית עבור כל למבדא שהוא:

$$L(\lambda) = \sum_{i=1}^{n}\left(y_i - x_i^T\beta - \beta_0\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

כעת נבדוק את עצמנו על למבדא הוא כולל אינטרפולציה:

$$L(\lambda) = \sum_{i=1}^{n}\left(y_i - x_i^T\beta - \beta_0\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \quad s.t \quad \sum_{i=1}^{n}\left(y_i - x_i^T\beta - \beta_0\right)^2 = 0$$

כעת נראה במבחן המקרה שמתקיים בתרגיל אינטרפולציה. ערש

11.7.19 נכונה

אין לנו מה שיבחין במינ' הנעץ כי אנחנו במצב האינטרפ'

$$L(\lambda) = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum\beta_j^2 = \lambda\sum\beta_j^2$$

$$\Downarrow$$

$$y_i=\hat{y}_i \qquad \text{s.t.} \quad \sum_{i=1}^{n}(y_i-x_i^T\beta-\beta_0)^2=0$$

אז רק ממזערים את גודל הנעץ עבור תוצאות

$$\min_{\beta}\lambda\sum_{j=1}^{p}\beta_j^2 \underset{0<\lambda}{\Leftrightarrow} \min_{\beta}\sum_{j=1}^{p}\beta_j^2 \qquad \text{s.t.} \quad \sum_{i=1}^{n}(y_i-x_i^T\beta-\beta_0)^2=0$$

← $\hat{\beta}^{(מ)}$

ולכן יתרון שמבחין בין הפתרונות של האינטרפולציה. הפתרון שמביא את המינ' גודל למינימום הוא $\hat{\beta}^{(l2)}$

נשים לב כי עבור $0<\lambda$ פתרון האינטרפולציה שמוביל למינימום של ridge הנמוך ביותר הוא $\hat{\beta}^{(l2)}$

(ב) נראה כי:

$$\hat{\beta}(\lambda) \xrightarrow[\lambda\to 0]{} \hat{\beta}^{(l2)}$$

בסעיף (a) ראינו כי $\hat{\beta}(\lambda)$ אינו פתרון אינטרפולציה, מכיוון שנ'בחנו ש- $2\lambda\beta_j\neq 0$ לכל j אולם אם $\lambda\to 0$ אז נקבל כי $2\lambda\beta_j \xrightarrow[\lambda\to 0]{} 0$ לכל j ולכן $\hat{\beta}(\lambda)$ בגבול מתכנס לפתרון אינטרפולציה.

ועת נשים לב לבעיית ה- ridge הרגילה:

$$\hat{\beta}(\lambda) = \underset{\beta_0,\underline{\beta}}{\operatorname{argmin}}\sum_{i=1}^{n}(y_i-x_i^T\beta-\beta_0)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

אבל נרצה להראות שמתקיים שכאשר $\lambda\to 0$ מתקבל פתרון אינטרפולציה כלומר:

$$\sum_{i=1}^{n}(y_i-\underline{x}_i^T\underline{\beta}-\beta_0)^2=0$$

אם מניחים ש- $0<\lambda$ בגבול האינטרפולציה משוואה נקבל כי (המשך)

$$\hat\beta(\lambda) = \underset{\beta_0,\beta}{\arg\min}\ \lambda\sum_{j=1}^{p}\beta_j^2 \qquad \lambda\to 0$$

$$\text{s.t}\quad \sum_{i=1}^{n}(y_i - x_i^T\beta - \beta_0)^2 = 0$$

ובמה מוקים שמגדיר את $\lambda\sum_{j=1}^{p}\beta_j^2$ עמו. כאשר λ שו כמו
שמגדיר את $\sum_{j=1}^{p}\beta_j^2$ ולכן ניתן לפתור את הבעיה:

$$\hat\beta(\lambda) = \arg\min \sum_{j=1}^{p}\beta_j^2 \qquad \lambda\to 0$$

$$\text{s.t}\quad \sum_{i=1}^{n}(y_i - x_i^T\beta - \beta_0)^2 = 0$$

וכן, בנוסף, לפתרון של $\hat\beta^{(LS)}$ הכין:

$$\hat\beta(\lambda) \xrightarrow[\lambda\to 0]{} \hat\beta^{(LS)}$$

(9) נסתכל באיזה סבור $\hat\beta^{(LS)}$:

$$\hat\beta^{(LS)} = \arg\min \sum_{j=1}^{p}\beta_j^2$$

$$\text{s.t}\quad \sum_i (y_i - x_i^T\beta - \beta_0)^2 = 0$$

ניזכר בתרגיל kernel representation of ridge regression (שיעורי בית 4 שאלה 2 סעיף ג') של פתרון זהו. אם נעזר ב-SVD להציג תמונות, כי עבור מצב כמו שלנו שבו $X_{n\times p}$ עם $\text{rank}=n$ כלומר $n\le p$ ניתן לפתור את פתרון ridge כך:

$$\hat\beta(\lambda) = X^T(XX^T+\lambda I_n)^{-1}y$$

14.1.19

נושא

כלומר נרצה להראות כי:

$$\hat{\beta}(\lambda) \xrightarrow[\lambda \to 0]{} \hat{\beta}^{(l_2)}$$

כלומר

$$\lim_{\lambda \to 0} \hat{\beta}(\lambda) = \lim_{\lambda \to 0} \hat{\beta}(\lambda) = \lim_{\lambda \to 0} \underline{X}^T (\underline{X}\,\underline{X}^T + \lambda I_n)^{-1} \underline{Y}$$

$$= \underline{X}^T (\underline{X}\,\underline{X}^T)^{-1} \underline{Y} = \hat{\beta}^{(l_2)}$$

$n \times n$

מטריצה $n \times n$ ומדרגה מלאה
מדרגה $=n$ ולכן הפיכה.

(f)

(i) בהתאם, נרצה את המינימום $\hat{\beta}^{(l_1)}$ כפתרון של:

$$\min \sum_{j=1}^{p} |\beta_j|$$

$$\text{s.t.} \quad \sum_i (y_i - x_i^T \beta - \beta_0)^2 = 0$$

כלומר אותם שמוגדרים "אינטרפולציה" בכל הנקודות על הדגימה בדיוק.

נראה על בסיס אותו רעיון שהוכחנו אותו ב-(e) ניתן להראות באותו אופן כי אכן:

$$\hat{\beta}_{lasso}(\lambda) \xrightarrow[\lambda \to 0]{} \hat{\beta}^{(l_1)}$$

(ii) כמו שראינו, ידוע לנו כי בגבול מתקבלת בעיה שקולה לאינטרפולציה. אין הבדל בין הפסד ריבועי לערך המוחלט,
מכיוון ש- $y_i = \hat{y}_i$ ולכן נוכל להציג את $\hat{\beta}^{(l_1)}$:
- הפסד מוחלט:

$$\hat{\beta}^{(l_1)}_{Abs} = \arg\min_{\beta} \sum_{j=1}^{p} |\beta_j|$$

$$\text{s.t.} \quad \sum_i |y_i - x_i^T \beta - \beta_0| = 0$$

(28)

12.1.19 בס״ד

- גרסת מינימיזציה:

$$\hat{\beta}^{(11)}_{qut} = \arg\min \sum_{j=1}^{p} |\beta_j|$$

$$s.t \quad \sum_{\substack{i=1 \\ i: y_i \geq y}}^{n} \tau (y - \hat{y}_i) + \sum_{\substack{i=1 \\ i: y_i < y}}^{n} -(1-\tau)(y_i - \hat{y}_i) = 0$$

וכך ניתן להראות כי אכן:

$$\hat{\beta}_{abs}(\lambda) \xrightarrow[\lambda \to 0]{} \hat{\beta}^{(11)}_{Abs} \quad / \quad \hat{\beta}_{qut}(\lambda) \xrightarrow[\lambda \to 0]{} \hat{\beta}^{(11)}_{qut}$$

(ז)

(i) לפי שניתן לראות בגרפים המצורפים אכן מתקיים כי כאשר שואפת λ (כוונת מינוס):

$$\lambda \to 0$$

אז גם ב ridge וגם ב lasso שגיאות האימון שואפות ל- 0.

ועבור S (כוונת constrain):

$$S \to \infty$$

אז גם ב- ridge וגם ב- lasso שגיאות האימון שואפות ל- 0

כלומר בעת המודלים נעבור לפתרון אינטרפולציה. וכפי שלמדנו פתרון אינטרפולציה בעל הנורמת L1 הנמוכה ביותר במודל של lasso ונורמת L2 הנמוכה ביותר במודל של ridge

(ii)

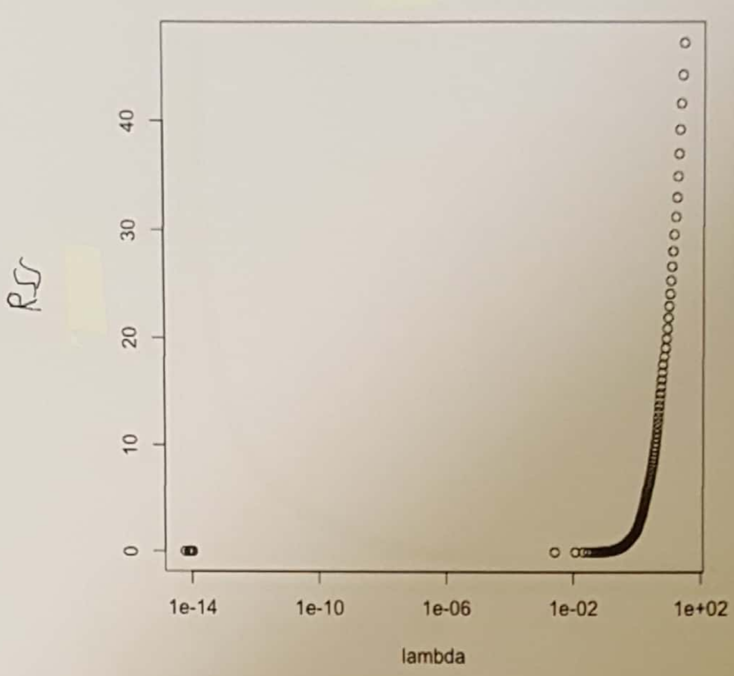
L1 norm for Lasso: 6.5102

L1 norm for ridge: 8.4917

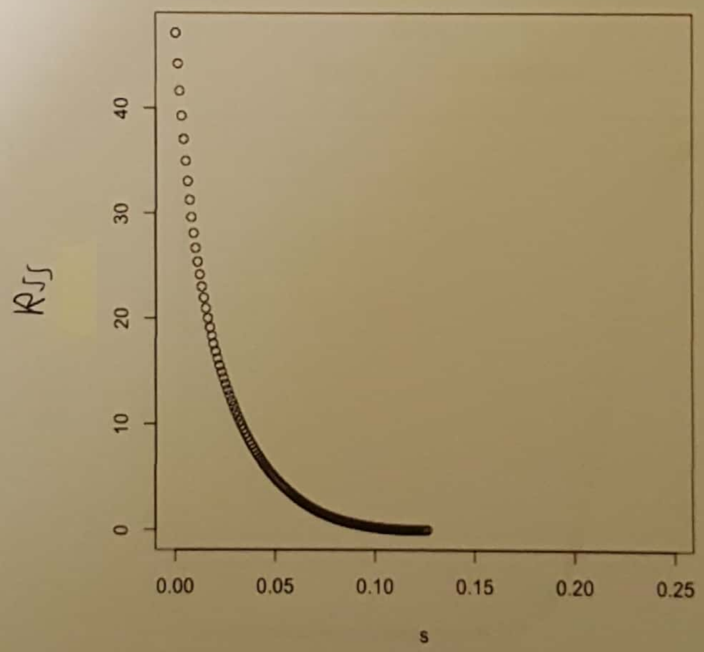
L2 norm for Lasso: 2.6726

L1 norm for ridge: 1.344

RSS
Lasso Vs Lambda

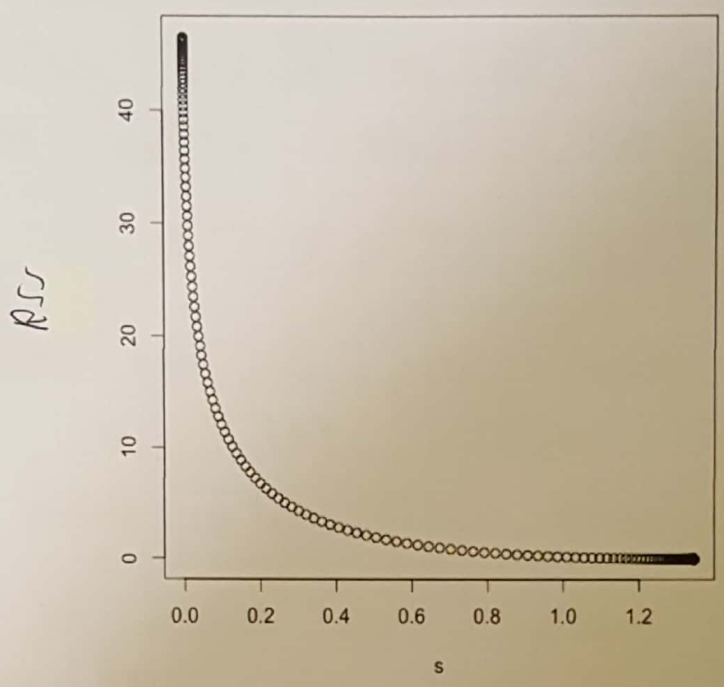


RSS
Lasso Vs s



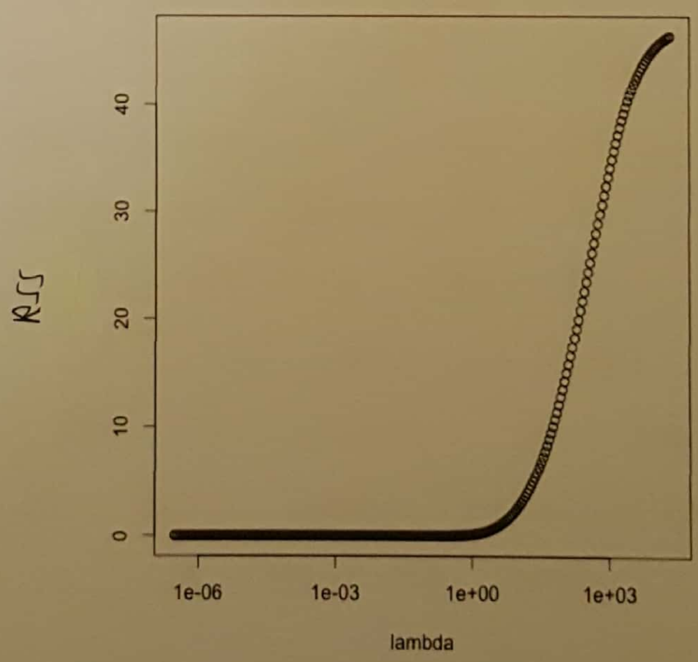
RSS

Ridge Vs s



RSS

Ridge Vs Lambda



כאשר ניתן לראות כמו שציפינו שהנורמה L1 של lasso היא נמוכה, מצד שני ridge והנורמה L2 של ridge נמוכה.
עם lasso נמוך ואמנם במודל פתרונות האינטרפולציה.
כאשר אנו ניקבנו לשפר ממנו לעיל ביותר את הנורמות שציפינו לה. בערך b ו f של הנורמות לו.

$$\hat{\beta}^{(11)} \quad | \quad \hat{\beta}^{(12)}$$

* כאשר הנורמה גדולה של upper bound של, ל- סלג' עבור lasso גבוהה. אבל, כמו מצבות' גדול נצל.
לפי הצלעים מסדר (i), הנורמה ב- norm.val=0.15.

(h) פתרון אינטרפולציה מוגדר על סוגי י.ג. כך שאנו את התצפית i במקום השני. נוכל עבור בניין את ה- class שלו (y_i), כאשר בנתון היא מוסף המדמם י.ג. הנ"ל.
(כלומר עבור בניית סיווג בו חצם נוכל למצוא את מיורים מעליים בין כל התצפיות במקצת הממיכה כך שנוכל לסווג את תצפית במקצת הממיכה לכיוון ה- class שלו, כאשר כל תצפית תהיה בצד המתאים ה- class שלו ה.. ישר.
נראה כאשר כי ההתכנסות ל- $\hat{\beta}^{(11)}$ י.ג. פתרון האינטרפולציה שמצאנו את הנורמה י.ג. בערך פתרון ה- svm (margin maximizing hyper plane). כלומר במקום נקבל:
- את בעיית האופטימיזציה של ה- svm ערשוו כמו שלמדנו בכיתה.

$$\min_{\beta} \|\beta\|_1$$
$$s.t \quad y_i(x_i^T\beta+\beta_0) \ge 1$$

כאשר ויכול שבציוע הגדה שמתירים בין ה... ה..

13.1.19 נט

מישור המפריד בעל המשמעות הרבה ביותר.

כאשר מוצאים מינימום לנורמה של $\|\beta\|_2^2$ תחת האילוץ שכל case

נקודה מסווגת נכון, ז"א

$$y_i(x_i^T\beta+\beta_0)\ge 1$$

כאשר a מקסימום בגבול הנכון של הזה מיוחד.

- במקרה 3 שאלה 5 ראינו גוונות כי ב- Separable case

(כמו עבור נוסף $n\ge p$) אין נוסף עם רגולריזציה

עבור ridge קיים

$$\hat\beta(\lambda)=\arg\min_\beta -l(\beta;x,y)+\lambda\sum_j\beta_j^2$$

where $l(\beta;x,y)$ is logistic log likelihood

אז כאשר $\lambda\to0$ נוקב ל $\frac{\hat\beta(\lambda)}{\|\beta(\lambda)\|_2}$ מתכנס

! מסווג SVM.

כיוון במאמר שהראינו הגאוסיאן ניתן להראות כי

$$\frac{\hat\beta(\lambda)}{\|\hat\beta(\lambda)\|_2}\xrightarrow[\lambda\to0]{}\text{מסווג SVM}=\text{minimum-norm-interpolation}$$

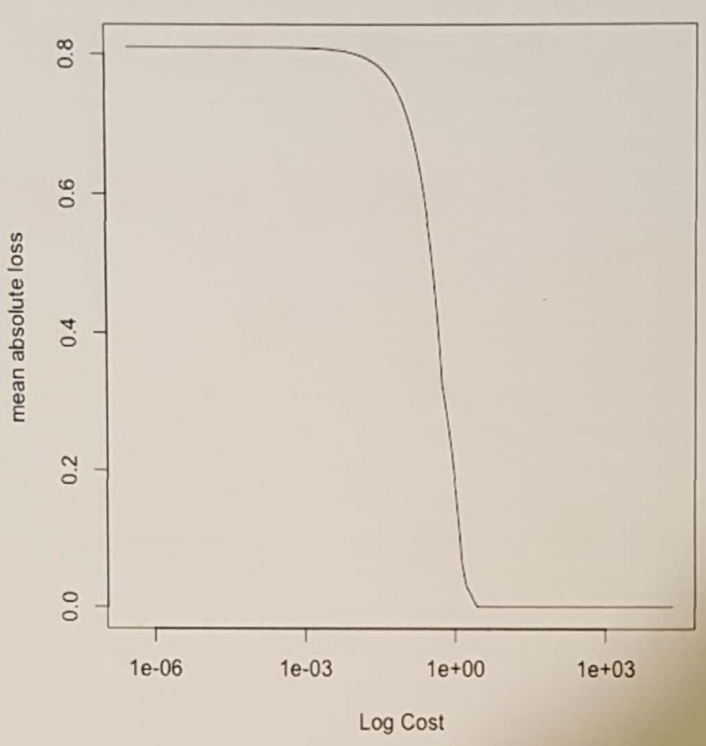
כאשר לא ידוע את הפתרון של $\hat\beta^{(n)}$ לרגרסיה לוגיסטית SVM ועם

הגודל שונות המשתנה בהינתן עבור רגולריזציה, ז"א

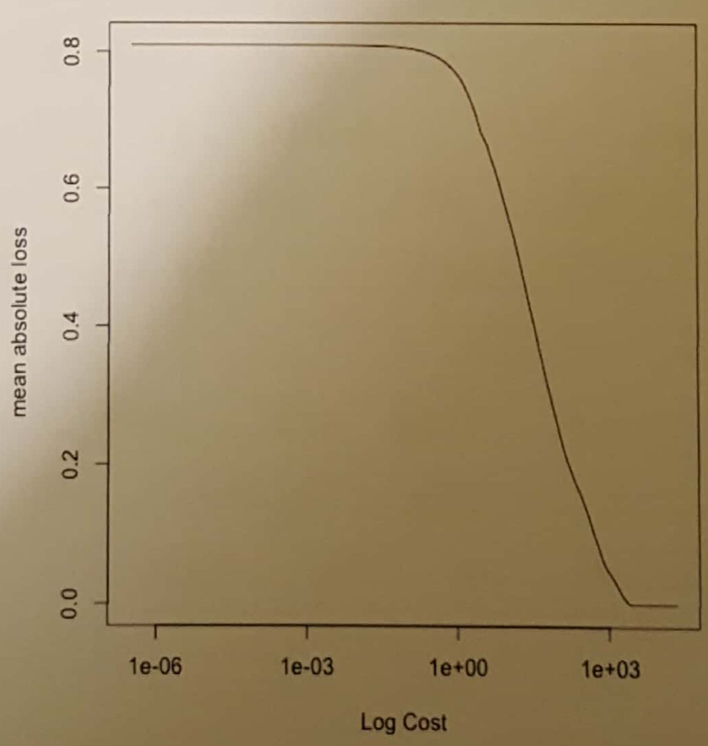
$$\frac{\hat\beta(\lambda)}{\|\beta(\lambda)\|}\xrightarrow[\lambda\to0]{}\text{מסווג SVM,}$$ הפתרון של $\hat\beta^{(n)}$ במובן

Scanned by CamScanner

Support vector regression-RBF kernel Gamma=5.0



Support vector regression-RBF kernel Gamma=0.0001



13.1.19

5. Absulute loss kernel regression:

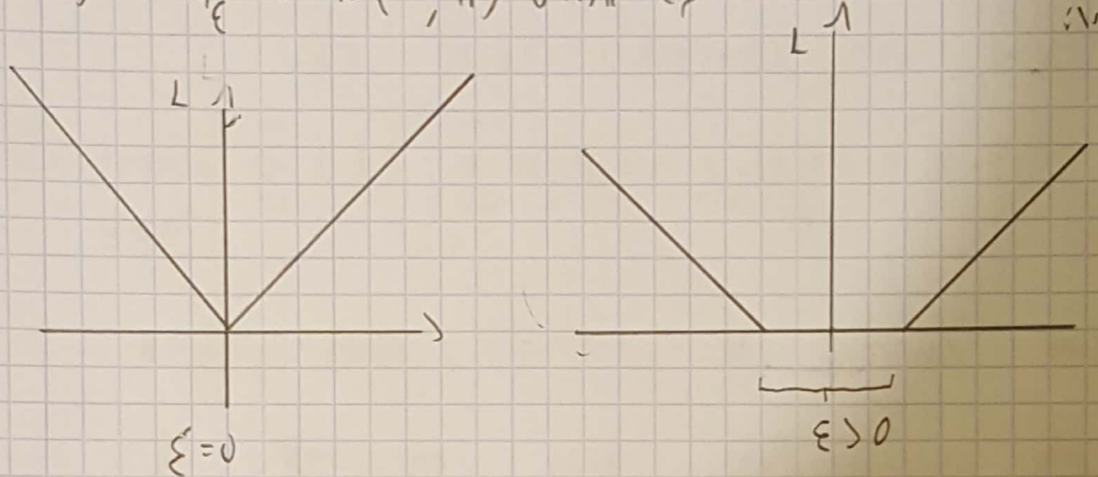
i.

- נבחר $\varepsilon=0$ ו- type = "eps-regression" וגם לבחור

ושאר הפרמטרים כמו קודם ג'י.

תזכורת:

$$\|y-f(x)\|_{\varepsilon} = \max\{0, \|y-f(x)\| - \varepsilon\}$$



ולכן עבור $\varepsilon=0$ נקבל

$$\|y-f(x)\|_{\varepsilon=0} = \max\{0, \|y-f(x)\| - 0\} = \|y-f(x)\|$$

- גם שינינו לאות גרעינים ולמנגנונים ויגן מתבצעת

הרבה פעמים אחרים עם 0 כמו כן.

ג. cost גדול (לא) מינימלי שונה (ברירת מחדל).

כמו כן עבור gamma= מערכת ...

מנגנון יותר.

ii. כדי להשוות לזה את RKHS, נראה כי

כי כדי בג"ח ...

נשים ! H_K מוגדר כ- ...

$$f(\underline{x}) = \sum_{k=1}^{q} \beta_k h_k(\underline{x})$$

ונורמה ...

$$\|f\|_{H_K}^2 = \sum_{k=1}^{q} \beta_k^2$$

ונראה ...

$$\hat{f} = \underset{f \in H_K}{\arg\min} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \sum_{k=1}^{q} \beta_k^2$$

ראינו כי לפי משפט הייצוג הפתרון הוא מהצורה;

$$\hat{f}_\lambda(\underline{x}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\underline{x}, \underline{x}_i)$$

כלומר בעיית המינימיזציה המקורית היא:

$$\hat{\underline{\alpha}} = \underset{\underline{\alpha} \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} L\left(y_i, \sum_{j=1}^{n} \alpha_j K(\underline{x}_i, \underline{x}_j)\right) + \lambda \underline{\alpha}^T K \underline{\alpha}$$

כאשר המטריצה K היא $K = [K_{ij}] = [K(\underline{x}_i, \underline{x}_j)]$

ולכן במקום - לפי n עמודות דרך באינסוף שונים מתקבלת פונ'
אנו שונה ממד עמודות דרך באינסוף שונים מתקבל בגודל
רגולריזציה שונה.

(iii)
- Prediction for small gamma is: 3.365144
- Prediction for Big gamma is: 3.589935

(iv)+(v) ראינו נכון את γ ואת התחזיות המקוריות של
radial kernel. התחזיות המקוריות מתקבלות בגלל שאפשר תצפית
x^* תצפית ה.ד.ק המשקל המתאים ל"סגנונות" הקרובות:

$$K(\underline{x}, \underline{x}^*) = e^{-\gamma \sum_{m=1}^{p} (x_{im} - x^*_m)^2}$$

כאשר $\gamma = \frac{1}{2\sigma^2}$

כלומר אם התצפיות רחוקות $K(\underline{x}, \underline{x}^*)$ יותר קטן
ולכן פחות משקל ה.ד.ק שכן פחות משפיע התחזית:

$$f(\underline{x}^*) = \sum_{i=1}^{n} \hat{\alpha}_i \cdot K(\underline{x}^*, \underline{x}_i)$$

(4)

כעת ניסה ערכים ל-γ, ונראה כמה $\gamma\uparrow$ (גבוה) יותר
גאוסיאנים יותר צרים ויותר קרובים מאוד סביב תצפית ספציפית
ומגיע מותאם למערך האימון (דוגמא לכך היא
ההתכנסות היותר מהירה של $\gamma=0.5$ בעמ' i).
עכשיו נציג תוצאות של הגרפים שקיבלנו, לפי הגרפים ניתן לראות
שככל שעולה ה-γ נקבל תוצאות "שגויות"/"גרועות"
עבור γ נמוך.
כעת מעניין שהדגמתי ניסוי, שלי מעבר למצב התאמה שקיבלנו
ניסוי שבו התאמת כמות הפרמטרים לא רצויה. כלומר אנחנו במצב
של overfit ויצרנו בעיית variance גבוה עבור
$\gamma=5.0$ לעומת $\gamma=0.0001$, קיבלנו בעיית variance ובמקרה השני
של בעיית Bais.
כעת נראה את טבלה ל. באופן רב המודל, באופן כללי
אני באמת שמחתי את כל המספרים, בחרתי rbf kernel, הוספתי
בטבלה של המודלים שקיבלתי:
- עבור $\gamma=5.0$ נקבל של ערכים מותאמים (למשל מינוס).
מאוד של e^{-213})
- עבור $\gamma=0.0001$ נקבל ... המינימלי הוא 0.8423
כמו עבור המודל של $\gamma=5.0$, קיבלנו מצב שבו אין
שגיאה בכלל! לכן המודל הוא (כמו) ה-Bais זה
הממוצע של y (mean(y)=3.54)
תוצאות כאלה עבור $\gamma=0.0001$ (כמו) כי התוצאות במצב
האימון עם ערכים לכן (כמו), קיבלנו ממוצע של
כל ערכי y במצב האימון (ככל ...)
... בצורת המשוואה ... 3.36)
כמו במודל ... ה-cost ... (כמו)
... לבחון ... קיבלנו ...
... "שגיאה" ...

בע"ה

13.1.14

כלומר אני מאמין (ורצים בשונות את ערכו של ונכון קירוביים

מחיר לממוצע והממוצע קטן $\gamma=5.0$ ג,ם

$$\hat{y}_i \approx 3.59$$