

מבוא למערכות לומדות | 67577

הרצאות | ד"ר מתן גביש וד"ר גבריאל סטנובסקי

כתביה | נמרוד רק

תשפ"ב סמסטר ב'

שתי הרצאות לפי ד"ר סטנובסקי, השאר לפי ד"ר גביש. שני התרגולים הראשונים לפי דן דרנזה והשאר לפי גלעד גרין. חלקים מהרצאה הושלמו באדיבות **סיכון של דוד ודן אל.**

תוכן העניינים

I מבוא לאומדים	
6	הרצאה
6	תוכנות של אומדים
6	AMDן נראות
8	התפלגיות וגאוסיאנים מרובי משתנים
11	תרגול
13	
II ריגרסיה לינארית	
16	הרצאה
16	טרמינולוגיה
16	NFL
17	מחלקות היפותזות הלינאריות
18	המקרה חסר המימוש
18	RSS
19	תרגול
21	
III בעיות סיוג	
25	הרצאה
25	איントואציה לטריידז'ו הטיה-שונות בריגרסיה לינארית
25	בעיות סיוג
25	מסוג חצי מרכיב
27	SVM
28	ריגרסיה לוגיסטיבית
29	k-NN
31	עצבי סיוג
33	תרגול
34	
IV מסגרת PAC	
37	הרצאה
37	הגדרות לפני שمبינים אותו
37	גרסה 1.0 של המשחק
38	כנראה, במעט, נכון
39	גרסה 2.0 של המשחק
40	אין ארכוות חיים
41	גרסה 3.0 של המשחק
41	למידת מחלקות היפותזות סופיות
42	
44	VC מימד
45	תרגול

47	V מסגרת PAC אגנסטיית
47	הרצאה
48	בעיות עם PAC קלאסי והכללתן
50	חשיבות ההתקנסות במ"ש
52	פער העלות הגדל ביוטר
52	חסימת פער העלות הגדל ביוטר במקרה הסופי
53	גדילה פולינומיאלית של מחלוקת היפוtheses
54	תרגול
59	VI אלגוריתמי סיוג הסטברותיים
59	תרגול
62	VII שיטות מכלול
62	הרצאה
63	עדות וシיטות ועדת
63	Bootstrap
64	Bagging
64	דה-קורלציה
65	Boosting
65	Adaboost
68	תרגול
71	VIII טריידוף הטיה-שונות ובוסטיניג
71	תרגול
73	IX רגולרייזציה
73	הרצאה
73	רגולרייזציה
74	עצים מקראים
75	ריgresיה ליניארית עם רגולרייזציה
76	ריgresיות Best-Subset
76	ריgresיות Ridge
77	ריgresיות Lasso
79	בחירה ושערוך מודלים
81	בחירה Bootstrap בבחירה מודלים
82	תרגול

85	X	למידה לא מפוקחת
85	הרצאה	
85	שימושים ללמידה לא מפוקחת	
86	PCA	
88	דוגמאות ל-PCA	
90	PCA-קסוד	
90	פסאודו-קסוד ל-PCA	
91	בחירה k ב-PCA	
91	Clustering	
93	k-Means	
95	קלאסטרינג ספקטורי	
	תרגול	
99	XI	שיטות גירעון
99	הרצאה	
99	מודיבציה לKERNEL ודוגמאות	
101	Kernel-Trick	
103	אלגוריתמי למידה מקורנים	
104	KERNELים מפורטים	
106	למידה מטרית	
106	תרגול	
110	XII	איך לפטור בעיה בלמידה
110	הרצאה	
110	לגשת בעיה חדשה	
112	פיתוח מודל	
113	שיקולים חישוביים	
114	דיווח תוצאות	
114	שחזור תוצאות	
114	תרגול	
117	XIII	בעיות קמוריות ו-GD
117	הרצאה	
117	פונקציות קמוריות	
119	תת-גרדיינטאים	
121	אופטימיזציה קמורה	
122	Gradient-Descent	
126	Descent Sub-Gradient	
127	תרגול	

XIV למידה עמוקה

128	הרצאה
128	SGD
129	PAC במסגרת SGD
131	רשתות נוירוניים
132	איך לאמן רשתות
135	תרגול
136	

XV רשתות נוירוניים וסיכום

137	הרצאה
138	תרגול
139	

שבוע II | מבוא לamodelים

הרצאה

בקורס נלמד על אלגוריתמים לומדים (Estimators), איך הם עובדים ועקרונותיהם, בעיות למידה שනפרות באמצעות אלגוריתמים אלו וכיitz לתוכנת את הביעות האלה.

מה זה אומד? נניח שאנו רוצה לדעת מה השעה אבל אין לי שעון. שאל כמה אנשים שלהם שעונים עם סטייה קטנה ואלו יהיו הדגימות שלי.

האומד הכי פשוט הוא ממוצע, $\bar{x} = \frac{1}{m} \sum_i x_i$. אוסף הדגימות נקרא מדגם.

נניח במהלך הקורס שהדגימות שלנו הן מ"מ שווי התפלגות \mathcal{P} .

הגדרות בסיסיות

הגדרה נאמר כי x_m, X_1, \dots, X_m הם $i.d$ אם $x_1, \dots, x_m \sim \mathcal{P}$, כלומר $x_i = x_m$.

הגדרה נאמר כי x_m, X_1, \dots, X_m $\stackrel{i.i.d}{\sim} \mathcal{P}$ אם הם $i.i.d$ ומדובר ב"תונסמן" X_i ו x_m .

הגדרה נניח תמיד ש- \mathcal{P} מוגדר ע"י פרמטר כלשהו $\theta \in \Theta$ כאשר Θ הוא אוסף כל הפרמטרים האפשריים.. אם הוא מוגדר ע"י כמה פרמטרים, θ יהיה וקטור פרמטרים.

דוגמה עבור התפלגות פואסון, ההתפלגות היא עם פרמטר $\lambda \in \mathbb{R}_+$.

הגדרה אומד מקבל דגימות מ- \mathcal{P} כשייננו יודעים את θ . נרצה למצוא את θ^* שהואści מותאים ש- θ יהיה.

הגדרה האمدن מוגדר ע"י פ' החלטה/כלל $\Theta \rightarrow \mathbb{R}^m : \delta$ שבו ניתן דגימה נתן את θ^* הנ"ל.

הגדרה $\{\delta : \mathbb{R}^m \rightarrow \mathbb{R}^m\}$ נקרא מחלקה ההיפותזות ונסמך ב- \mathcal{H} .

מטרה בהינתן Δ , נרצה למצוא $\delta \in \Delta$ שעבורו $\delta(x_1, \dots, x_m) = \delta^*$ מתאר באופן המיטבי את θ .

הדגורה לכל $\delta \in \Delta$, $\delta(X_1, \dots, X_m)$ נקרא אומד נקודתי או אומד.

דוגמה נניח שיש לנו $x_1, \dots, x_m \stackrel{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ כאשר $\mathcal{P} = \mathcal{N}(\mu, \sigma^2)$. כיצד נגלה את σ^2, μ ? נוכל להשתמש באומד ממוצע מדגם ואומד שונות מדגם שייחסבו בהתאם

$$\hat{\mu}_X = \frac{1}{m} \sum x_i, \quad \hat{\sigma}^2 = \frac{1}{m-1} \sum (x_i - \hat{\mu}_X)^2$$

(במהשך נראה מדוע זה $m-1$ ולא m).

תכונות של אומדים

לчисוב השונות אפשר היה לבחור אומדיים אחרים :

$$\hat{\sigma}_1^2 = \frac{1}{m-1} \sum (x_i - \hat{\mu})^2, \quad \hat{\sigma}_2^2 = \frac{1}{m} \sum (x_i - \hat{\mu})^2, \quad \hat{\sigma}_3^2 = \frac{1}{m} \sum |x_i - \hat{\mu}|$$

אבל לכל אחד יש יתרונות וחסרונות בקשר לשערוך שלו של θ האמיטית. נctrיך להגדיר מה הופך אומד להכי טוב.

הערה δ הוא למעשה מ"מ בעצמו כי הוא פ' על וקטורי מקרי.

הגדרה נאמר כי אומד δ הוא חסר הטיה אם בתוחלת הוא מוחזר את הערך הנוכחי, כלומר $E[\delta(X_1, \dots, X_m)] = \theta$

הגדרה הטעות של אומד δ עבור פרמטר θ מוגדרת ע"י

הגדרה ההטיה של אומד δ עם פרמטר θ היא

$$\text{Bias}_{\theta} [\delta(X_1, \dots, X_m) - \theta] = E_{X_1, \dots, X_m | \theta} [d]$$

$$\text{נאמר כי } \delta \text{ חסר הטיה אם } \text{Bias}_{\theta} [\delta(X_1, \dots, X_m)] = 0$$

דוגמה הסיבה שמדובר הוא הרבה פעמים אומד טוב הוא שהוא לא מוטה, שכן

$$\begin{aligned} E_{X_1, \dots, X_m | \mu} [\hat{\mu}(X_1, \dots, X_m)] &= E_{X_1, \dots, X_m | \mu} \left[\frac{1}{m} \sum X_i \right] \\ &= \frac{1}{m} \sum E_{X_i | \mu} [X_i] \\ &= \frac{1}{m} m \cdot \mu = \mu \end{aligned}$$

גם אומד השונות הוא לא מוטה, שכן

$$\begin{aligned}
 E[\hat{\sigma}^2] &= \frac{1}{m-1} \sum_i E[(x_i - \hat{\mu})^2] \\
 &= \frac{1}{m-1} \sum_i E\left[x_i^2 - 2x_i \cdot \frac{1}{m} \sum_j x_j + \frac{1}{m^2} \sum_{j,k} x_j x_k\right] \\
 &= \frac{1}{m-1} \left(\sum_i E[x_i^2] - \frac{2}{m} \sum_{i,j} E[x_i x_j] + \frac{1}{m} \sum_{j,k} E[x_j, x_k] \right) \\
 &= \frac{1}{m-1} \left(\sum_i E[x_i^2] - \frac{1}{m} \sum_{i,j} E[x_i x_j] \right) \\
 &= \frac{1}{m-1} \left(\left(1 - \frac{1}{m}\right) \sum_i E[x_i^2] - \frac{1}{m} \sum_{i \neq j} E[x_i x_j] \right) \\
 &\stackrel{\text{ב''ג}}{=} \frac{1}{m-1} (m-1) E[X^2] - \frac{m(m-1)}{m} E^2[X] \\
 &= E[X^2] - E^2[X] \\
 &= \sigma^2
 \end{aligned}$$

הגדרה יהיו δ אומד לפרמטר θ . השונות של δ מוגדרת ע"י

$$\text{var}(\delta) = E_{X_1, \dots, X_m | \theta} \left[(\delta(X_1, \dots, X_m) - E_{X_1, \dots, X_m | \theta} [\delta(X_1, \dots, X_m)])^2 \right]$$

דוגמה נחשב את השונות של אומד ממוצע המדגם. יהיו $x_1, \dots, x_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}$ עם שונות σ^2 .

$$\begin{aligned}
 \text{var}(\hat{\mu}) &= \text{var}\left(\frac{1}{m} \sum x_i\right) \\
 &= \frac{1}{m^2} \text{var}\left(\sum x_i\right) \\
 &\stackrel{\text{ב''ג}}{=} \frac{1}{m^2} \sum \text{var}(x_i) \\
 &= \frac{1}{m^2} m \sigma^2 \\
 &= \frac{\sigma^2}{m}
 \end{aligned}$$

כלומר ככל שיש יותר דוגמאות, האומד בעל שונות יותר קטנה ולכן באינסויו הוא יהיה אומד מושלם (שונות 0 סביב התוחלת).

אומדן נראות

הערה נוכל לקבוע שהאומד הכי טוב שלנו הוא $\Delta \in \delta$ שהוא לא מותה ובעל שונות מינימלית. לחלופין נוכל להשתמש ביחס נראות.

הגדרה יהיו $f_\theta(x)$ ו- f , הcpfיות של \mathcal{P} . פונקציית הנראות היא $\mathcal{L}(\theta | x)$ לכל דוגמה x -היא.

דוגמה עבור גausian, $(\mu, \sigma^2) = \theta$. פ' הנראות היא

$$\mathcal{L}(\theta | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(x_i - \mu)^2}{2\sigma^2}\right)}$$

כלומר דועך אקספוננציאלית בכל שמותרדים מהתווחת. על דוגמאות x_1, \dots, x_m i.i.d, הנראות היא

$$\begin{aligned} \mathcal{L}(\theta | x_i) &= \prod_{i=1}^m f_\theta(x_i) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} e^{-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2} \end{aligned}$$

הערה פונקציית הנראות מעשה עונה על השאלה "מה הסיכוי שנקבל את x אם אנחנו עם פרמטר θ ?".

למעשה, פ' הנראות על דוגמאות i.i.d נותרת לנו הערכה על הסיכוי שהדוגמאות הגיעו מפרמטר כלשהו.

דוגמה עבור גausian $\mu = 0, \sigma^2 = 1$ ו- $x_1 = -1, x_2 = 0, x_3 = 0, x_4 = 1$ הנראות של x_1, \dots, x_4 $\stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ היא

$$\mathcal{L}(\mu = 0, \sigma^2 = 1 | x_1, \dots, x_4) = \frac{1}{(2\pi)^2} e^{-\frac{1}{2} \sum_{i=1}^4 x_i^2} = \dots \approx 0.00931$$

הנראות של $\mu = 1$ היא בערך 0.00126. אולי נראה שהמספר הזה אומר משהו אך למעשה אין לו יותר מדי משמעות בפני עצמו.

הגדרה יהיו \mathcal{L} פ' נראות של התפלגות \mathcal{P} עם פרמטר $\Theta \in \Theta$. ימי דוגמה ממנה. אומד הנראות המקסימלי (MLE) עבור θ הוא

$$\hat{\theta}^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta | x)$$

דוגמה נחשב את ה-MLE של התוחלת של גausian בהינתן שידועה לנו השונות σ^2 . יהיו x_1, \dots, x_m לכך

$$\begin{aligned} \hat{\mu}^{MLE} &= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \mathcal{L}(\mu | x_1, \dots, x_m, \sigma^2) \\ &= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} e^{-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2} \end{aligned}$$

ונשים לב כי הממוצע ימקסם את הערך הנוכחי גם אם נפעיל עליו פ' מונוטונית, לדוגמה לוגריתמית, ככלומר

$$\begin{aligned}\hat{\mu}^{MLE} &= \operatorname{argmax}_{\mu \in \mathbb{R}} \log e^{-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2} \\ &= \operatorname{argmax}_{\mu \in \mathbb{R}} -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \\ &= \operatorname{argmax}_{\mu \in \mathbb{R}} -\sum_i (x_i - \mu)^2\end{aligned}$$

לכן נגזר נושא לאפס למציאת המקסימום,

$$\begin{aligned}\frac{\partial}{\partial \mu} \left(-\sum_{i=1}^m (x_i - \mu)^2 \right) &= -\sum_{i=1}^m \frac{\partial (x_i - \mu)^2}{\partial \mu} \\ &= -2 \sum_{i=1}^m (x_i - \mu) = 0\end{aligned}$$

ככלומר

$$\hat{\mu}^{MLE} = \frac{1}{m} \sum_{i=1}^m x_i$$

קיבלונו שזה אומד ממוצע המדגם בסוף, ומהיחסוב אנחנו יודעים גם שהוא ממזער את סכום הסטיות הריבועיות של x .

אחרי שהסתכלנו על התוחלת והשונות, נרצה לדעת בכלל גם איך מתפלג האומד שכן גם הוא מ"מ.

דוגמא עבור אומד ממוצע התוחלת לא רק שתוחלתו μ ושונתו $\frac{\sigma^2}{m}$, אלא הוא גם נורמלי עם פרמטרים אלו ובנוסף הוא מרוכז סביב התוחלת (לא מיטה) ותוחלתו פרופורציונית בשנות הנתונים, שדווקע לתינוקת בכל שנקבל עוד נתונים.

כדי לכמת את אייכות האומד, נוכל לשאול מה ההסת' לקבל ערך מסוים? כמה אנחנו בטוחים בניחוש שלנו? מה הסיכוי לסטות מהערך האמיטי כתלות במספר הדגימות?

דוגמא נרצה לאמוד את החטיה של מטבח (לא הוגן), ככלומר מה ההסת' p שנטיל עץ. פורמלית, יש לנו התרפלגות של מ"מ ברנולי

$$\mathcal{D}_p(X) = \begin{cases} p & X = 1 \\ 1-p & X = 0 \end{cases}$$

ובהינתן m הטלות $S = \{x_1, \dots, x_m\}$, נסמן את החתרפלגות של m הטלות ב- \mathcal{D}_p^m ואת ההסת' לקבالت S ב- $\mathcal{D}_p^m(S)$. נציג אלג' לומד \mathcal{A} שיאמוד את p ואז נשאל כמה מדויק הוא.

האלג' מקבל כקלט את S שהוא אוסף דגימות i.i.d. לפי \mathcal{D}^m ופולט אומדן $\hat{p}(S)$. נסמן את הפלט ב- $\hat{p}(S)$.

נשתמש באומד ממוצע המדגם $\hat{p}(S) = \frac{1}{m} \sum_i x_i$ שייתן לנו את היחס האמפירי בין עצים לפליים.

ידעו לנו שהוא אומד לא מוטה, ככלומר שהוא מרוכז סביב התוחלת, נרצה לדעת מה המרחק הגדול ביותר (ϵ) של \hat{p} מ- p ?

לאחר מכון, ב�לל שיכול להיות שקיילנו S מאוד לא יציג, נרצה לדעת מה ההסת' שאנו מדויקים, כלומר מהו (ϵ) $P(|\hat{p} - p| \leq \epsilon)$. לכל רמת זיוק ϵ נוכל לחשב את ההסת' להיות במרקח הזה מ- ϵ , באמצעות מركוב נקבל

$$\mathcal{D}^m(|\hat{p} - p| \geq \epsilon) \leq \frac{E[|\hat{p} - p|]}{\epsilon}$$

נחשב את התוחלת שכן לא נרצה לחשב את הביטוי הלא נוח זהה:

$$\text{var}(|\hat{p} - p|) = E[|\hat{p} - p|^2] - E^2[|\hat{p} - p|] \geq 0$$

ולכן

$$\begin{aligned} E^2[|\hat{p} - p|] &\leq E[|\hat{p} - p|^2] \\ &= E[(\hat{p} - p)^2] \\ &= E[(\hat{p} - E[\hat{p}])^2] \\ &= \text{var}(\hat{p}) \\ &= \frac{1}{m^2} \sum \text{var}(x_i) \\ &= \frac{p(1-p)}{m} \leq \frac{1}{4m} \end{aligned}$$

ושה"כ נקבע $E[|\hat{p} - p|] \leq \frac{1}{\sqrt{4m}}$ ולכן חוזרת להסת' לצאת מטווח ϵ , נקבע

$$\mathcal{D}^m(|\hat{p} - p| \leq \epsilon) \geq 1 - \frac{1}{\sqrt{4m\epsilon^2}}$$

ואם נסמן $\delta = \frac{1}{\sqrt{4m\epsilon^2}}$ נקבע $m = \lceil \frac{1}{4\epsilon^2\delta^2} \rceil$ ולקמן $\delta \in (0, 1)$ אם יש לנו לפחות m נסמן $\delta = \frac{1}{4\epsilon^2\delta^2}$ ונקבל $m = \lceil \frac{1}{4\epsilon^2\delta^2} \rceil$ שההערכה שלנו היא מדויקת מספיק (בטווח ϵ).
לכן, נוכל להגיד ברמת בטיחות $\delta - 1$ שתהערכה שלנו היא מדויקת מספיק (בטווח ϵ).

נוכל באופן יותר חזק לחסום עם א"ש צ'בישב:

$$P(|\hat{p} - E[\hat{p}]| \geq \epsilon) \stackrel{\text{לא מיטח}}{=} P(|\hat{p} - p| \geq \epsilon) \leq \frac{\text{var}(\hat{p})}{\epsilon^2}$$

ואחרי הרבה חישובים מתקבלים שעבור $\delta = \frac{1}{4m\epsilon^2}$ ו- $\epsilon, \delta \in (0, 1)$, בהינתן $m = \frac{1}{4\delta\epsilon^3}$ נסמן $m \geq \lceil \frac{1}{4\delta\epsilon^3} \rceil$ ולקמן $\delta = \frac{1}{4\delta\epsilon^3}$ דגימות נוכל להבטיח שזה הרבי יותר טוב כי כמות הדגימות היא לינארית ב- δ ולא ריבועית כמו במרקוב.
 $\mathcal{D}^m(|\hat{p} - p| \leq \epsilon) \geq 1 - \delta - \delta$

התפלגויות מרובות משתנים

לרוב עסקנו במ"מ עם ערך יחיד, לדוגמה הטלת מטבח, שעאabel במציאות ובמשך הקורס נרצה לנבא וקטוריים מקרים, לדוגמה בן אדם עם גובה ומשקל. נוכל להשתמש בשונות משותפת כדי לתאר את הקשר בין הנתונים (יש קשר כלשהו בין גובה ומשקל לדוגמה).

הגדירה השונות המשותפת של וקטור מקרי ($\Sigma \in M_d(\mathbb{R})$ היא $X = (X_1, \dots, X_d)$)

הערה על האלכסון של Σ יש את השינויות של X_i .

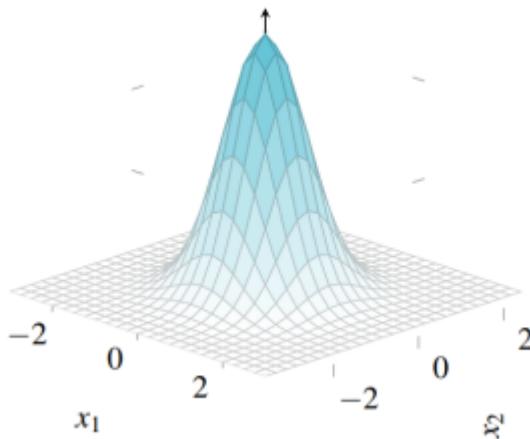
הגדירה נאמר כי $\Sigma \in \mathbb{R}^{d \times d}$ מתפלג נורמלית רב-משתנית עם תוחלת $\mu \in \mathbb{R}^d$ ומטריצת שונות משותפת אם ו惩

הכפיפות שלו מקיימות

$$f(X) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}$$

$X \sim N(\mu, \Sigma)$ ונסמן

דוגמה הגaussian הדו-משתני נראה כבאיור והוא מוגדר ע"י מטריצת שונות (ווקטור תוחלות) והמתוקן הוא $N(0^2, I_2)$.



העובדת שיש 0 בכל מקום שלא על האלכסון מתארת בין היתר את אי התיוות בין הציריים.

הגדירה עבור מ"מ $f(X_A) = \int_{X_B} f(X_A, X_B) dX_B$ ההתפלגות השולית של $A \neq B \in [d]$ היא $X = (X_1, \dots, X_d)$

טענה יhi $(\mu_1, \sigma_1^2, \rho)$ גaussian דו-משתני. אז $X \sim N(\mu, \Sigma)$ והתפלגות השולית היא

$$f(X_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2}$$

דוגמה נוכל לתאר כל דוגמה כוקטור $x \in \mathbb{R}^d$ מתחזק וקטור מקרי $X = \begin{pmatrix} X_{weight} \\ X_{height} \end{pmatrix}$.

עבור תיאור כלשהו של משקל/גובה של בני אדם (העריכים על האלכסון המשני מתארים את הקשר בין משקל לגובה). כאן אומד התוחלת הוא

זהה לחלוטין רק שאנחנו סוכמים וקטוריים ממשיים ולא רק ממשיים.

כדי לאמוד את סigma נגיד ראשית אומד לא מוטה על השונות המשותפת של שני מ"מ X_i, X_j ע"י

$$\hat{\sigma}(X_i, X_j) = \frac{1}{m-1} \sum_k (x_{ki} - \hat{\mu}_i)(x_{kj} - \hat{\mu}_j)$$

משמעותו של $\hat{\Sigma}_{ij} = \hat{\sigma}(X_i, X_j)$ הוא $\hat{\Sigma} = \frac{1}{m-1} \tilde{X}^T \tilde{X}$ (מטריצה המאנגדת את הדגימות) מחיסום בוטצית מטריצות האומד שלו הוא $X \in \mathbb{R}^{m \times d}$ כאשר $\tilde{X} = \hat{\Sigma}^{-\frac{1}{2}} X$ (מטריצה המאנגדת את הדגימות) מכיסוי $\hat{\mu}$ מכל עמודה.

הערה עבור אוסף $S = \{x_1, \dots, x_m\}$ של דגימות ב- \mathbb{R}^d , נסמן את התוכונה ה- j של הדגימה ה- i ב- x_{ij} ויש עוד כמה סימוני אינדקסים לנוחות שימושינו.

תרגול

הגדרה מטריקה היא $d : X \times X \rightarrow \mathbb{R}$ המקיים את התכונות הבאות:

$$1. x = y \text{ אם } d(x, y) = 0.$$

$$2. d(x, y) = d(y, x).$$

$$3. d(x, y) \leq d(x, w) + d(w, y).$$

הגדרה גורמה היא $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}_+$ המקיים את התכונות הבאות:

$$1. \|v\| \geq 0, v = 0 \text{ אם } \|v\| = 0.$$

$$2. \|av\| = |a| \|v\|.$$

$$3. \|u + v\| \leq \|u\| + \|v\|.$$

הגדרה מעגל היחידה עבור נורמה $\|\cdot\|$ הוא $\{v \in \mathbb{R}^d : \|v\| = 1\}$.

דוגמא הנורמה האוקלידית $\|\cdot\|_2 = \sqrt{\sum_i x_i^2}$. מעגל היחידה הוא מעגל ברדיוס 1 סביב הראשית.

דוגמא נורמת ℓ_1 המוגדרת ע"י $\|v\|_1 = \sum |v_i|$. מעגל היחידה הוא ריבוע המשובב ב- 45° עם צלע $\sqrt{2}$ סביב הראשית.

דוגמא נורמת ℓ_∞ המוגדרת ע"י $\|v\|_\infty = \max_i |v_i|$. מעגל היחידה הוא ריבוע עם צלע 2 סביב הראשית.

דוגמא הכללה של כל הנ"ל היא ℓ_p המוגדרת ע"י $\|v\|_p = \sqrt[p]{\sum_i |v_i|^p}$ ולמעשה נורמת אינסוף היא הגבול כש- p שואף לאינסוף.

הגדרה מכפלה פנימית מעל מ"ז V היא $\langle \cdot | \cdot \rangle : V \times V \rightarrow \mathbb{R}$ המקיים את התכונות הבאות:

$$1. \langle u | v \rangle = \langle v | u \rangle.$$

$$2. \langle \alpha v + u | z \rangle = \alpha \langle v | z \rangle + \langle u | z \rangle.$$

. $\langle v | v \rangle \geq 0$.3. (אי-שליליות)

$$\text{תכונה} \quad \|u - v\|^2 = \langle v - u | v - u \rangle = \langle v | v \rangle - 2 \langle v | u \rangle + \langle u | u \rangle$$

טענה (משפט הקוסינוסים) $\cos \theta = \frac{\langle u | v \rangle}{\|u\| \|v\|}$ כאשר θ היא הזווית בין v, u המוגדרת ע"י $\|v - u\|^2 = \|v\|^2 + \|u\|^2 - 2 \|v\| \|u\| \cos \theta$

הגדירה יהיו V, W מ"ו מעל \mathbb{R} . $T : V \rightarrow W$ תקרא טרנספורמציה לינארית אם היא מקיימת את התכונות הבאות :

$$1. \quad T(u + v) = T(u) + T(v)$$

$$2. \quad T(\alpha v) = \alpha T(v)$$

הגדירה $F : V \rightarrow W$ תקרא טרנספורמציה אפנית אם קיימת טרנס' לינארית T ו- $\vec{b} \in W$ כך ש- לכל $v \in V$ $F(v) = T(v) + \vec{b}$

הגדירה עבור מטריצה A המיצגת את הטרנס' $T : V \rightarrow W$, הגרעין שלה מוגדר ע"י $\ker A = \{x \in V : Ax = 0\}$

$$\text{Row}(A) = \text{Im}A^T \quad \text{Col}A = \text{Im}A = \{w \in W : \exists x \in V, Ax = w\} \quad \text{ע"י}$$

טענה תהי $A \in \mathbb{R}^{m \times d}$

הערה אם $\text{rank } A = \min\{m, d\}$ אז A בעלת דרגה מלאה.

הגדירה $AB = I_d$ -ש $B \in \mathbb{R}^{d \times d}$ תקרא הפיכה אם קיימת

טענה תהי $A \in \mathbb{R}^{d \times d}$. אז התנאים הבאים שקולים :

1. A הפיכה.

2. $\text{rank } A = d$

3. $\text{Im}A = \mathbb{R}^d$

4. $\ker A = \{0\}$

תכונה אם $w = X^{-1}y$ ומ X הפיכה אז גם $y_i = \sum_j X_{ij}w_j$ אז $y \in \mathbb{R}^d, X \in \mathbb{R}^{d \times d}, Xw = y$

הגדירה יהיו $u, v \in V$ הטללה האורתוגונלית של u אל תוך v ע"י $p = \langle u | v \rangle \frac{u}{\|u\|^2} = \frac{\|v\|}{\|u\|} \cos \theta u$

הערה הטללה האורתוגונ' היא הוקטור המקביל ל- u שמשיך עד לאנך מ- v ל- u . היא הוקטור הכי קרוב ל- u המקביל ל- u .

הגדירה המכפלה החיצונית של u, v היא $u \oplus v = v \cdot u^T = \begin{pmatrix} v_1 u_1 & \dots & v_1 u_m \\ \vdots & & \vdots \\ v_n u_1 & \dots & v_n u_m \end{pmatrix}$

הגדירה הטללה האורתוגונ' של על בסיס אורתוגונ' v_1, \dots, v_k של ת"מ מוגדרת ע"י $V \subseteq \mathbb{R}^d$ ועבור $i = 1, \dots, k$ מתקבלים את הנדרש ועבור $j \neq i$ זה מותפס).

$P = \sum_{i=1}^k v_i \oplus v_i = \sum_{i=1}^k v_i v_i^T$ (כופלים כל שורה בעמודה ועבור $j = i$ מקבלים את הנדרש ועבור $j \neq i$ זה מותפס).

טענה תהי P מטריצת הטללה אורתוגונ' אז מתקיים :

$$1. \quad P^2 = P$$

.2 סימטרית.

.3. הע"ע היחדים שלחם הם 1 (עם ריבוי גאומטרי k) ו-0 (עם ריבוי גאומטרי $d - k$).

$$(I_d - P)P = 0 .4$$

$$\forall x \in \mathbb{R}^d, u \in V, \|x - u\| \geq \|x - Px\| .5$$

הגדירה תהי $A \in \mathbb{R}^{d \times d}$ סימטרית. נאמר כי היא PSD (Positive Semi Definite) אם $\forall x \in \mathbb{R}^d$ מתקיים $x^T Ax \geq 0$ ונסמן $0 \succ A$ אם הא-שוויון הוא תמיד נכון.

טענה תהי $A \in \mathbb{R}^{d \times d}$ מטריצה סימטרית. אז התנאים הבאים שקולים:

$$\forall x \in \mathbb{R}^d, x^T Ax \geq 0 .1$$

.2. כל הע"ע של A אי-שליליים.

$$.3. \text{קיימת } B \in \mathbb{R}^{d \times d} \text{ כך ש-} B^T B = A$$

הוכחה: (1) \Leftarrow (2) נניח בשלילה שקיימים ע"ע שלילים λ עבור ו"ע v , لكن $0 \leq v^T Av = \lambda v^T v$ סתירה.

$$(1) \Leftarrow (3)$$

$$x^T Ax = \langle x | Ax \rangle = \langle x | B^T Bx \rangle = \langle Bx | Bx \rangle \geq 0$$

■

הגדירה נאמר כי $A \in \mathbb{R}^{d \times d}$ לכיסינה אם קיימות $P, D \in \mathbb{R}^{d \times d}$ הפיכה, D אלכסונית ומתקיים

משפט (המשפט הספרטורי, EVD) תהי $A \in \mathbb{R}^{d \times d}$ סימטרית אז קיימת U אורתוג'ן כך ש-

הערה בגלל שכל הע"ע של A מטריצה PSD הם אי-שליליים, נוכל להגיד דומה לא- A ויתקיים $A = D_{ii}^{\frac{1}{2}}$ כאשר $D_{ii}^{\frac{1}{2}} = \sqrt{D_{ii}}$ אלכסונית דומה לא- A ויתקיים $B = D^{\frac{1}{2}}U U^T D U = U^T D^{\frac{1}{2}} D^{\frac{1}{2}} U = B^T B$.

דוגמה נרצה לחשב חזקה של מטריצה סימטרית, $A^{200} = UD^{200}U^T$, ואז צריך רק להעלות בחזקה את הע"ע ולא מטריצה שזה יקר.

הגדירה תהי A טרנס' לינארית מ- V -ל- W אז אם מתקיים $Av = \sigma u$ עבור $v \in V, u \in W$, נאמר כי u וקטור סינגולרי משמאלי, v וקטור סינגולרי מימין ו- σ ערך סינגולרי.

הגדירה נאמר כי $B \in \mathbb{R}^{m \times d}$ היא סמי-אלכסונית אם $[B]_{ij} = 0 \forall i \neq j \forall$ מתקיים

משפט (SVD) לכל מטריצה $A \in \mathbb{R}^{m \times d}$ קיימות $U \in \mathbb{R}^{m \times m}$ אורותג'ן ו- $V \in \mathbb{R}^{d \times d}$ אלכסונית כך ש- $A = U\Sigma V^T$ והעמודות של V , U והקטורים סינגולריים משמאלי ומימין של A בהתחמלה והערכיהם על האלכסון של Σ הם ערכים סינגולריים של A .

הערה ככלומר כל מטריצה ניתנת להציג כסיבוב, מתייה ושוב סיבוב. ככלומר כל טרנס' נתנת אליפסoid במרחב אחר.

שבוע III | ריגרסיה לינארית

הרצאה

בלמידת מכונה, יש לנו מנגנון חביי ואנחנו מנסים לעשות לו הנדסה לאחור בתבבש על מאפיינים כלשהם, כדי שנוכל להפעיל אותו גם על מקרים חדשים.

דוגמא המנגנון שקובע האס למשוח יהיה לחץ דם גבוה לא ידוע לנו (mbossed על משתנים כמו סביבה וגנטיקה), אבל נרצה באמצעות מקרים נכפים שהם לחץ דם של מטופלים אחרים, לגלוות האס לאחרים יהיה לחץ דם גבוה בהסתמך על פיצ'רים כלשהם (משקל, גיל וכו').

קטגוריות וטיעמים בלמידת מכונה

• Supervision :

- מפוקחת : התווויות ידועות לנו (התצפיות מתקבלות עם לחץ הדם של המטופל).
- לא מפוקחת : התווויות לא ידועות (לא ידוע לנו לחץ הדם של המטופלים אלא רק המאפיינים שלהם ועלינו לחלקים לקבוצות עם מאפיינים כלשהם).

• אופן תצפיות :

– Batch : מקבלים קבוצה קבועה של תצפיות (רישומות של בי"ח).

– Online : מקבלים תצפיות אחת אחר השניה.

ועוד הרבה אחרות.

בקורס אנחנו עוסקים בלמידה מפוקחת עם תצפיות Batch.

בתכנות קלאסי אנחנו מקבלים נתונים ותוכנה ופולטים פלט, בלמידת מכונה אנחנו מקבלים נתונים ואת הפלט ומחזירים את התוכנה (שתוכן לדמות פלט בדומה למידע-פלט שקיבלו עד כה).

”קובעים“ בלמידת מכונה

במהלך למידת המנגנונים של למידת מכונה, ננתח את האלגוריתמים שלנו מכמה קובעים שונים.

• ידוע לנו המנגנון החביי במדוייק, זה לא מצב ריאליסטי אבל הוא עוזר למציאת חסמים תאורטיים וניטוחים אחרים. Oracle Mode :

• מצב מבחן, הערכה : בהינתן מודל, נוכל לבדוק כמה טוב הוא בתיאור המציאות (גם בלי לדעת מה המנגנון החביי במדוייק), לדוגמה כמה התקפי לב מענו באמצעות חיזוי לחץ דם יקבע האס מודל שנבנה הוא מוצלח או לא.

• מצב התאמת/אימון : בנייתו וכיוול של מודל למידה.

הערה במהלך הקורס הרבה פעמים נבדק התאמה למצב מבחן כדי לבדוק את עצמו אפרי שוננה את המודל.

סיכוםונים

- סקלר: אוטיות יווניות קטנות (a).
- וקטורים ופ': אוטיות לטיניות קטנות (v).
- מטריצות: אוטיות לטיניות גדולות (בהרצאות בbold A, אני אכתוב רגיל A).
- אוסףים ומ"מ: אוטיות לטיניות גדולות (S).
- ערכים משוערכים: מסומנים עם כובעים (\hat{w}).

הערה על אף חשיבותם של נוסחים, פעמים רבות טוב להזכיר מה המשתנים אומרים במצבות.

הגדרה מידע אימון הוא אוסף $S = (x_i, y_i)_{i=1}^m$ כאשר $x_i \in \mathbb{R}^d$ הוא הקלטים (פיצ'רים), ו- y_i ברגירסיה לינארית הוא ב- \mathbb{R} ומשמש כמידע שאנו מנסים לנבא, כלומר אם f הוא המנגנון החבוי אז $y_i = f(x_i)$. נגידו את מטריצת הקלט $X = \begin{pmatrix} \dots & x_1 & \dots \\ \vdots & \vdots & \vdots \\ \dots & x_m & \vdots \end{pmatrix} \in \mathbb{R}^{m \times d}$.

דוגמא בלחץ דם, x_i הוא משקל, גובה, גיל וכו' ו- y_i הוא לחץ הדם.

הגדרה מודל ברגירסיה לינארית הוא פ' $\hat{f}_S : \mathbb{R}^d \rightarrow \mathbb{R}$ שמכoon (אינטואיטיבית)קיים

הגדרה למידה היא בניית מודל על בסיס מידע אימון.

הגדרה מידע מבחן הוא מידע שיש לנו אחרי הלמידה ואנו לא יכולים לגעת/לשנות אותו.

הגדרה לדוגמה מידע המבחן שלנו הוא לחץ הדם האמייתי של המטופל, ובאמצעותו נקבע האם המודל היה מוצלח (לדוגמה, אם הוא ניבא קרוב מספיק ניתן לו ניקוד ואחרת נגער ממנו ניקוד).

משפט (No Free Lunch) כדי שנוכל לבנות מודל כלשהו, אנחנו חייבים להגביל את המודל שלנו לחלוקת היפווזה כלשהי, כלומר קיימת $\hat{f}_S \in \mathcal{H}_\theta$ עבורה תמיד מתקיים

הערה ככלומר, אנחנו חייבים לקבוע אוסף מוגבל של מודלים שיכולים להיות לדענו, כי אחרת נוכל לבנות אוסף שרירותי וAKERAI עם ערכים שמתנהגים בכל דרך שנרצה וכך לכואורה זה אף פעם לא ייגמר כי אולי זה המודל הזה ואולי זה מודל אחר. לכן נצטרך לשולח חלק מההיפותזות כדי לקבל תוצאה בעלת ערך. אנחנו מחפשים תבנית ולכן נגביל את ההיפותזות שלנו רק לתבניות, ולא כל הבא לדי.

הערה קביעת מחלוקת ההיפותזות היא שלב קריטי שכן אם המציאות לא מצאת בחלוקת ההיפותזות לעולם לא נוכל לנבא נאמנה את המנגנון החבוי.

הערה למעשה לעולם לא נוכל לדעת שהמנגנון האמייתי נמצא בחלוקת ההיפותזות אלא אם אנחנו במצב אורקל ואין דרך קבועה לבחור את \mathcal{H}_θ ואוסף הפיצ'רים הנדרשים לנו כדי לנבא משלו.

עתה נלמד על מחלוקת היפותזות לינאריות שעבורה התווית תלויות באופן לינארי בפיצ'רים. כלומר אנחנו מניחים שבחרו עבורנו בעיה שidue
שהיא מתנהגת באופן לינארי.

הגדרה מחלקת ההיפותזות הלינאריות היא w_1, \dots, w_d ו- $\mathcal{H}_{lin} = \left\{ (x_1, \dots, x_d) \mapsto w_0 + \sum_{i=1}^d w_i x_i : w_i \in \mathbb{R} \right\}$ נקראות המשקולות ו- w_0 הesisט.

הערה אם $w_j = 0$ אז נוכל לדעת שהפיצ'ר x_j לא רלוונטי ולכן לא צריך לדגום אותו.
הערה כדי לעשות שיתהיפה, נגידר מעשה $x_i = (1, x_1, \dots, x_d)$ וזו למעשה מודל לינארי הוא מהצורה $x \cdot w$ ולמזה של $\hat{f}_S \in \mathcal{H}_{lin}$ היא ממצא $\hat{w} \in \mathbb{R}^{d+1}$ בחתבסט על S .

כדי לנבא \hat{y} עם מודל \hat{w} מקרה חדש x' , נחשב $\hat{y}' = \hat{w} \cdot x'$.

הערה כדי להבין את המודל (איך הוא עובד) נוכל להסתכל על יחס המשקולות w_i (ערך חיובי גדול אומר שימושי מאוד בכיוון אחד, ערך שלילי גדול בכיוון אחר, ערך קטן משפייע פחות וכיווץ).

נסתכל על שני מקרים של מיידת מנוקדת מבט אל אורקל:

- המקרה בר-המיימוש (Realizable) בו $f \in \mathcal{H}_{lin}$ (המנגנון החבוי האמתי) בלי שום רעש.
- המקרה חסר-המיימוש (Non-Relizable) בו $f \notin \mathcal{H}_{lin}$ או שיש לנו רעש על הישר.

הערה בשני המקרים האלה, פירוק SVD ייתן לנו את המודל האידיאלי.

המקרה בר-המיימוש

נניח כי לחץ D הוא אכן תלוי לינארית בנתונים. אם קיימת $f \in \mathcal{H}_{lin}$ (כפי שאנו מניחים) או קיימת $w \in \mathbb{R}^{d+1}$ כך ש- $y_i = \langle x_i | w \rangle = y_i$. אם קיימת $f \in \mathcal{H}_{lin}$ (כפי שאנו מניחים) או קיימת $w \in \mathbb{R}^{d+1}$ כך ש- $Xw = y$. כלומר $y = Xw$.

למעשה $y = Xw$ אם ו- $y = Xw$ יש לפחות פתרון אחד, זו מערכת עם m משוואות ו- $d+1$ נעלמים.

הערה זה לא מאד ריאלי ולא מעניין לניתוח.

המקרה חסר-המיימוש

ל- $y = Xw$ אין בהכרח פתרון יוכל להיות פתרון שעבוד עבור הדגימות, אבל לא בהכרח במקרה הכללי לכל הערכים. במקרה, נסה למצואו משהו מספיק קרוב.

המקרה הזה קורה או אם f באמת לא לינארית אבל אנחנו מנסים לקרב אותה לכזו עדין, או אם היא כן לינארית אבל יש רעש כלשהו בדגימות (אחוז סטייה כלשהו, חסר איחדות בחישובים).

הגדירה דוגמאות עם רעש חן $y = f(x) + z$ הוא וקטור.

הערה הכוודינטת ה- i - של z קובעת כמה רעש יש לפיצ'ר ה- i ואם z לא טריויאלי אז התכפיות שלנו יוצאת מ- $\text{Im } X$.

עלות (Loss)

בhinintן מודל, נרצה לאמוד כמה הוא טוב. פ' ה

עלות
 של השגיאה של $\hat{f}_S(x_i)$ על דוגמיה x_i מהערך האמתי y_i ונרצה למזערה ככל שניתן.

הגדירה נאמר כי $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ היא פ' עלות אם $\hat{f}_S(x_i) = y_i$ אם $L(y_i, \hat{f}_S(x_i)) = 0$

דוגמא פ' עלות ערך מוחלט $L(y_i, \hat{f}_S(x_i)) = |y_i - \hat{f}_S(x_i)|$

דוגמא פ' סטייה ריבועית $L(y_i, \hat{f}_S(x_i)) = (y_i - \hat{f}_S(x_i))^2$

לפי שיטת ERM נרצה בהinintן פ' עלות, למצוא את המודל שמצמצם את סך ה

עלות
,

$$\underset{\hat{f}_S}{\operatorname{argmin}} \sum_{i=1}^m L(y_i, \hat{f}_S(x_i))$$

הערה המינימום על סכום ה

עלות
 הוא שונה לפי

עלויות
 שונות (סטייה ריבועית, סטייה בערך מוחלט).

RSS

תחת מחלוקת היפותזות לינארית, סך ה

עלות
 נקרא Residual Sum of Squares והוא

$$\sum_{i=1}^m (y_i - x_i^T w)^2 = \|y - Xw\|^2$$

כלומר ברגression לינארית נחפש את

$$\hat{w} = \underset{w}{\operatorname{argmin}} RSS(w) = \underset{w}{\operatorname{argmin}} \|y - Xw\|^2$$

הערה אם X בדרגה מלאה אז $RSS(w)$ היא רציפה עם מינימום יחיד. הדרגה לא מלאה אם יש תלות לינארית בין פיצ'רים (לדוגמא שאנו מה הגיל בשנים ובשבועות).

כדי למצוא מינימום נגורר ונמצא נקודה בה זה מתאפס:

$$\frac{\partial}{\partial w_j} RSS(w) = -2 \sum_{i=1}^m x_{ij} (y_i - x_i^T w) = 0$$

$$\text{או לחלופין } 0 \cdot \nabla RSS(w) = -2X^T(y - Xw) = 0$$

מטרה אנחנו מחפשים \hat{w} עבורו $0 = X^T y = X^T X w$. כלומר נקבעות משוואות נורמליות.

נכל לרשום $X = X$ כאשר f_i הוא הוקטור שמייצג בכל הדוגמאות את הפיצ'ר i -ה. את המשוואות הנורמליות נפרק פר עמודה ונקבל שhn למשה

$$\langle f_j | y - Xw \rangle = 0, \quad \forall j \in \{0, \dots, d\}$$

מתקיים $\hat{z} = y - \hat{y} \in (\text{Im } X)^\perp$ ו- $\hat{y} = X\hat{w} \in \text{Im } X$ נסמן $\{f_0, \dots, f_d\} = \text{Im } X$ ו- $y - X\hat{w} \perp \text{Im } X$ נדרש \hat{y} כרך לקיימים $\|y\|^2 = \|\hat{y}\|^2 + \|z\|^2$, שכן z הוא למעשה מעשה אותו הרעש מקודם.

אם עמודות X בת"ל אז החטלה האורתוגונית היא ייחודה (הוכחנו בלינארית 2) ואם הם ת"ל אז יש אינסוף הצלות (יש אינסוף ק"ל שיצרו אותה).

- אם הפיצ'רים בת"ל: $\hat{w} = (X^T X)^{-1} X^T y$ וגם $\dim \ker X^T X = \dim \ker X = 0$ ישירות מהמשוואות הנורמליות ואחרי מצום מקבלים $\hat{w} = X^{-1} y$

- אם הפיצ'רים ת"ל: נבצע פירוק SVD ל- X ונרשום אותו כ- $X = U\Sigma V^T$ כאשר $\Sigma \in \mathbb{R}^{m \times (d+1)}$ פסאודו-אלכסונית עם ערכי סינגולריים על האלכסון.

הסיכוי ש- X תהיה הופכית מאוד נמוך כי זה דורש שמספר הפיצ'רים ומספר הדוגמאות יהיה שווה, לכן נחשב את פסאודו-הופכית של X המוגדרת ע"י $X^\dagger = V\Sigma^\dagger U^T$ כאשר $\Sigma^\dagger \in \mathbb{R}^{(d+1) \times m}$ מוגדרת ע"י $\Sigma^\dagger_{ii} = \frac{1}{\sigma_i}$ אם $\sigma_i > 0$ ו-0 אחרת. כמובן ש- $X^{-1} = X^\dagger$ - קלומר הכללו גם את המקרה הקודם!

$$\text{הערה} \quad \text{עבור } \hat{w} \text{ כנ"ל מתקיים } \|\hat{w}\| = \min \{\|w\| : Xw = y\}.$$

לסיכום, המודל האידיאלי ביחס לעולות סטטיסטיה ריבועית מוצאים באמצעות RSS ומשתמש ב- SVD למציאת פסאודו-הופכיה.

אם $X^T X$ "כמעט ולא הפיך" אבל כן הפיך, לדוגמה יש תלות לינארית אבל קוורדיינטה אחת לא עונה לתלות, אז לעיתים הערכים הסינגולריים הם מאוד קטנים ואז $\frac{1}{\sigma_i}$ יהיה מאוד גדול ולא מדויק ונקבל מודל לא יציב נומריית ולכן נוכל להגיד באמצעות דיקון מכונה $\sum_{ii}^{\dagger, \epsilon} = \begin{cases} \frac{1}{\sigma_i} & \sigma_i > \epsilon \\ 0 & \sigma_i \leq \epsilon \end{cases}$.

כזכור, הטיה היא כמו קרובה מחלוקת ההיפותזות ל- f האמיתית ושותות היא כמו רגישה \hat{f}_S לדוגמאות שונות של f וכעיקרון, ככל ש- \mathcal{H} "מתוחשבת" יותר בדוגמאות כך השונות ועלה וההטייה תרד.

דוגמא עבור מחלוקת ההיפותזות המתאימות פולינומיAli המוגדרת ע"י $\mathcal{H}^0 \subseteq \mathcal{H}^1 \subseteq \dots \subseteq \mathcal{H}^d = \left\{ a \mapsto \sum_{k=0}^d w_k a^k \right\}$ וככל שהמעלה עולה כך הטעיה יורדת אבל השונות ועלה.

נדגום f - בלי רוש ונתאים מודלים עם פולינומיים ממעלה שונות, וכן אם נסתכל על הגרפים יש שיפור משמעותי ככל שהמעלה עולה (החותטיה יורדת).

עתה נדגום עם רוש אקראי z לכל דוגמה x ונראה שגם עבור מודל מאוד אקספרסייבי (פולינום מדרגה 27) הוא עוקב הרבה יותר אחריו הרוש מאשר אחריו הפ' האמיתית - ככלומר השונות מאוד גבואה.

ככל, אנחנו מוחפשיםazon טוב בין יציבות ביחס לרוש והתאמה לנוטוי האימון.

תרגול

לעתים קרובות בעיה בלמידה תתואר כצמצום עלות, ככלומר נctrיך לגוזר ולמצוא מינימום ולכן איןפי רב מימדי חשוב לנו.

הגדרה תהי $f : \mathbb{R} \rightarrow \mathbb{R}$. הנגזרת בנקודה x היא $\frac{df(x)}{dx} = \lim_{a \rightarrow 0} \frac{f(x+a)-f(a)}{a}$.

דוגמה עבור $\text{Relu}(x) = \max\{0, x\}$ פופולרית בלמידת מכונה.

הגדרה עבור $f : \mathbb{R}^d \rightarrow \mathbb{R}$, נגדיר את הנגזרת החלקית בנקודה $x \in \mathbb{R}^d$ ע"י $\frac{\partial f}{\partial(x_i)} x_i = \lim_{a \rightarrow 0} \frac{f(x+ae_i)-f(x)}{a}$, שזה בדיק לכיוון הגידלה של x_i .

הערה ככלומר אנחנו גוזרים לפי ערך אחד וכל השאר נשארים קבועים.

הגדרה תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$. הגרדיאנט בנקודה $x \in \mathbb{R}^d$ מוגדרת ע"י $\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix}$.

דוגמה עבור הנקודת $(1, 1)$ הגרדיאנט הוא $(3, 3)$ שזה בדיק לכיוון הגידלה של x_1 וכך גם עבור $(-3, -3)$ שזה בדיק לכיוון הגידלה של x_2 .

משפט (כלל לייבניץ) $\frac{\partial h(x)g(x)}{\partial x} = \frac{\partial h(x)^T}{\partial x} g(x) + h(x) \frac{\partial g(x)^T}{\partial x}$.

דוגמה $f(x) = w^T x$ לכן

$$\frac{\partial f(x)}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{i=1}^d w_i x_i \stackrel{\text{השאר מתאפסים}}{=} w_j \frac{\partial x_j}{\partial x_j} = w_j$$

ככלומר $\nabla f(x) = w$ ולכן בדיקת הגידלה תמיד יהיה כלפיו w .

דוגמה $\nabla f(x) = 2x$ ככלומר $\frac{\partial f(x)}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{i=1}^d x_i^2 = \sum_{i=1}^d \frac{\partial}{\partial x_j} x_i^2 = 2x_j$, $f(x) = \|x\|^2$

הגדרה תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$. היעקביאן של f בנקודה x מוגדר ע"י $J_x(f) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \dots & \frac{\partial f_m(x)}{\partial x_d} \end{pmatrix}$.

הערה למעשה, השורה ה- i של היעקביאן היא הגרדיאנט של f_i בנקודה x (טרנספוז, כי זו שורה ולא עמודה).

דוגמה $f(x) = Ax$ כאשר $A \in \mathbb{R}^{m \times d}$. את זה כבר חישבנו ולכן שזה נחמדה כי זה מقلיל את הרענון של פ' לינארית אם A היה סקלר (הנגזרת הייתה פשוטה וככלול $A = \begin{pmatrix} \dots & a_1 & \dots \\ \vdots & \ddots & \vdots \\ \dots & a_m & \dots \end{pmatrix} = A$ הסקלר).

הגדירה תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ פ' גזירה פעמיים. נגידר את החסיאן של f ב- x “וי”

$$H(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{pmatrix}$$

$$\text{כלומר } [H(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

הערה החסיאן היא סימטרית.

דוגמה עבור f (החסיאן שלו היה $H(f) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$) ו $f(x, y) = x^2 + xy + y^2$ וצורך $\frac{\partial^2 f}{\partial x^2}$.

משפט (כלל השרשרת) תהיינה $f : \mathbb{R}^d \rightarrow \mathbb{R}^m, g : \mathbb{R}^m \rightarrow \mathbb{R}^k$

$$J_x(f \circ g) = J_{g(x)}(f) J_x(g) \in \mathbb{R}^{m \times k}$$

או במפורש

$$[J_x(f \circ g)]_{ij} = \sum_{l=1}^d \frac{\partial f_i(g(x))}{\partial g_l(x)} \frac{\partial g_l(x)}{\partial x_j}$$

דוגמה עבור $J_x(g) = A, \nabla f = 2x$ וצורך $f \circ g = \|Ax\|^2$ ונחשב את היעקביאן של ההרכבה של g . $f(x) = Ax$ ו $J_x(f) = \|x\|^2$ ולכן $J_{Ax}(f) = 2(Ax)^T$ ומעבר לכך $J_x(f) = 2x^T$

$$J_x(f \circ g) = 2x^T A^T A$$

ובגלל שיש רק מימד אחד אז הגרדיינט הוא פשוט הטרנספוז של היעקביאן, כלומר $\nabla(f \circ g) = 2A^T Ax$

הגדירה תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ פ'. קירוב מסדר ראשון של f הוא $f(x) \approx f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle$

הגדירה תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$. קירוב מסדר שני של f הוא

$$f(x) \approx f(x_0) + \langle \nabla f(x_0) | x - x_0 \rangle + \frac{1}{2} \langle x - x_0 | H(f(x_0))(x - x_0) \rangle$$

הגדירה תהי $S \subseteq \mathbb{R}^d$. נאמר כי קמורה אם $C \subseteq S$ ואם $\forall \alpha \in [0, 1] \forall u, v \in C$

הערה למעשה הכוונה היא שהקטוע בין u ל- v כולל ב- C .

דוגמה כדור היחידה הוא $\{v \in V : \|v\| = 1\}$.

$$\begin{aligned} \|(1-\alpha)u + \alpha v\| &\stackrel{\Delta}{\leq} \|(1-\alpha)u\| + \|\alpha v\| \\ &= (1-\alpha)\|u\| + \alpha\|v\| \\ &\leq (1-\alpha) \cdot 1 + \alpha \cdot 1 = 1 \end{aligned}$$

תכונות

1. תהיינה קבוצות קמורות, אזי $\bigcap_i C_i$ קמורה.
2. עבור C_1, C_2 קמורות, $C_1 + C_2 = \{c_1 + c_2 : c_1 \in C_1, c_2 \in C_2\}$ קמורה גם כן.
3. עבור C קמורה, $\lambda C = \{\lambda c : c \in C\}$ קמורה גם כן.

הגדלה תהי $f : C \rightarrow \mathbb{R}^d$ קמורה אם $\alpha \in [0, 1]$ -ו $u, v \in C$ מעל ערכי הפ' על הקטע (כמו פרבולה קלאסית). אם יש אפילו

$$f((1-\alpha)u + \alpha v) \leq (1-\alpha)f(u) + \alpha f(v)$$

הערכה אינטואיטיבית, הכוונה היא שהקטע שמחבר בין כל שתי נקודות הוא מעל ערכי הפ' על הקטע (כמו פרבולה קלאסית). אם יש אפילו קטע אחד שיש לו נקודת מתחת לערך של הפ' שם, הפ' כבר לא קמורה.

טענה תהי $\|\cdot\| : \mathbb{R}^d \rightarrow \mathbb{R}$ נורמה, אזי היא קמורה.

הוכחה: יהי $u, v \in \mathbb{R}^d, \alpha \in [0, 1]$

$$\|(1-\alpha)u + \alpha v\| \stackrel{\Delta}{\leq} \|(1-\alpha)u\| + \|\alpha v\| = (1-\alpha)\|u\| + \alpha\|v\|$$

דוגמה עבור $g(x) = \sum a_i g_i(x)$ קמורות היא קמורה (חייב להיות חיובי כי אחרת עבור $g_1 = x^2, a_1 = 1$ קיבל פרבולה הפוכה שווה לא קמורה).

דוגמה $\sup_i f_i(x)$ כאשר f_i קמורות היא קמורה.

טענה הרכבה של קמורות היא קמורה.

הגדירה ה-אפיגרף של $f(x)$ מוגדר ע"י $\text{epi } f = \{(x, \beta) : x \in \text{dom } f, f(x) \leq \beta\}$.

משפט (אפיון קמירות באמצעות קירוב מסדר ראשון) תהי $f : C \rightarrow \mathbb{R}^d$ קמורה ומתקיים $\forall u, w \in C \subseteq \mathbb{R}^d$ אז $f(w) \geq f(u) + \langle \nabla f(u) | w - u \rangle$

$$f(u) \geq f(w) + \langle \nabla f(w) | u - w \rangle$$

משפט (אפיון קמירות באמצעות קירוב מסדר שני) תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ גזירה. אז f קמורה אם $\exists H(f) \subseteq 0$ (כלומר $x^T Ax \geq 0$ מטרכיה סימטרית, נגיד $Ax = f(x)$) וקיים $J_x(g) = A$ ו $J_x(h) = I_d$. נקבל $J_x(g) = A$ ו $J_x(h) = I_d$.

דוגמה עבור $A \in \mathbb{R}^{d \times d}$ מטריצה סימטרית, נגיד $Ax = f(x)$. נוכיח כי f קמורה. מכיל השרשרת עבור $J_x(g) = A$ ו $J_x(h) = I_d$. נקבל $J_x(g) = A$ ו $J_x(h) = I_d$.

$$\nabla f = J_x^T g(x) + A^T x = Ax + A^T x = 2Ax$$

$$J_x^T g(x) = 2A$$

הערה קבוצות קמירות הן כל כך חשובות לנו משום שככל מינימום מקומי הוא גם מינימום גלובלי (לא יכולה להיות גבעה, אלא רק עמק ושאי אפשר לעלות ולרדת ממנו למינימום מקומי אחר).

הגדירה בעיתת אופטימיזציה קמורה היא בהינתן $f_0, \dots, f_n : C \rightarrow \mathbb{R}^n$ קמירות, מציאת

$$\min_{x \in C, f_i(x) \leq b_i} f_0(x)$$

הגדירה בעיתת תכנון לינארי (LP) היא בהינתן $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$, מציאת

$$\min_{x \in \mathbb{R}^n, Ax \leq b} c^T x$$

הגדירה בעיתת תכנון ריבועי (QR) היא בהינתן $A \in \mathbb{R}^{m \times n}, Q \in \mathbb{R}^{n \times n}, a \in \mathbb{R}^m, b \in \mathbb{R}^n$ כאשר Q סימטרית, מציאת

$$\min_{w \in \mathbb{R}^n, Aw \leq d} \frac{1}{2} w^T Q w + a^T w$$

הערה אם Q היה PSD אז בעיתת התכנון הריבועי היה בעיתת אופטימיזציה קמורה כי $\frac{1}{2} w^T Q w + a^T w$ קמירות וגם סכומן, אחרת זו לא בעיתת אופטימיזציה קמורה.

שבוע III | בעיות סיוג

הרצאה

איןטואיציה ליחס בין דגימות לפיצ'רים

נזכר כי מה שעשינו בהרצאה הקודמת בנושא ריגרסיה לינארית הוא להסתכל על u ביחס לת"מ $\mathbb{R}^m \subseteq \text{Im}(X)$, ולמצוא מה הקירוב הכי טוב שלנו של u בתחום $(X) \text{Im}$, זהה למעשה הטלת האורטוג', והצלחנו באופן יציב נומרית (באמצעות SVD, שעובד גם אם עמודות X בת"ל וגם אם לאו) למצוא מודל לא רע.

ההנחה כאן הייתה שיש לנו יותר דגימות (m) מפיצ'רים (d), אך במצבות כל מאוד לאסוף מידע ואולי קשה יותר לאסוף דגימות, נבחן את הפער ביניהם. אם יש לנו הרבה דגימות, $(X) \text{Im}$ הוא יותר קטן ביחס למ"ז כלו וזו הטלת האורטוג' מחלוקת להתחמಡ יותר עם הרעש (היא מזיהה את u הרבה כי אין לה מספיק מדדים). לכן חשוב שיחיו לנו הרבה דגימות. כאמור, אם $m > d$ **בערך** באותו הגודל זו הטלת לא מחלוקת לנוקות את הרעש במקביל- $(X) \text{Im}$ (בתוכו) ואז מקבלים מודל לא איקוני כי בלבו רושם לתקן אותו, אבל אם m מאוד גדול לעומת d , יש לנו ת"מ יחסית קטן וזו הטלת יותר שימושית. ובפעם השלישייה:

כל שהת"מ יותר גדול ($m \approx d$), **ההטיה של המודל קטנה** (אנחנו יכולים לtarget יותר תופעות) אך **השינויים גדלה** כי דגימה חדשה (שהה יש רושם) תנסה את המודל הרבה (אנחנו בולעים את הרעש).

כל שהת"מ קטן ($m < d$), **ההטיה גדלה** (כל נתון מוטל מאוד אחרית מערכו האמתי) אך **השינויים קטנה** (אין יותר מדי אפשרויות בתחום הת"מ ולכן דגימות מטאורופות יצומצמו לטוחה די מוגבל שלא משפייע כל הרבה).

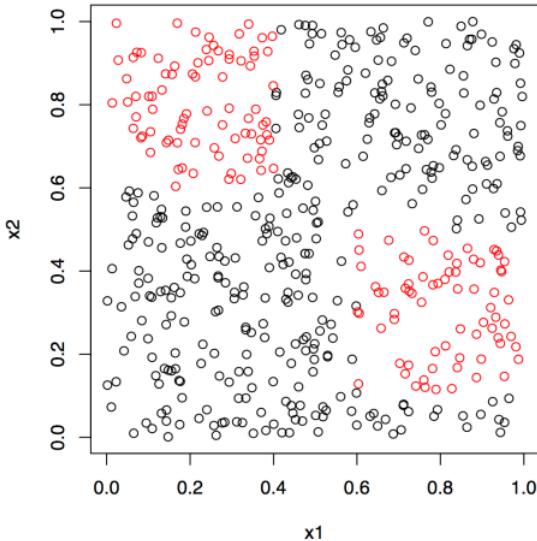
נעסק עתה ב-**Prediction**, **Classification**, כלומר קביעה ביינרית על דגימה, במקרה מה שעשינו עד עכשיו ונעשה בהמשך, שהוא קביעה פרמטר רציף עבור דגימה.

בעיות סיוג

בעיות קלאסיפיקציה, התמונה היא בדידה, כלומר $\{1, \dots, k\}$ וailo במקרה הבינארי מרחב הפיצ'רים הוא $\{-1, 1\}$.

דוגמא בעיות קלאסיפיקציה מאוד שימושיות, לדוגמה קביעה האם מטופל הולך למות, האם משתמש ירצה לקנות מוצר, האם אימיל הוא ספאם, האם תקשורת חשודה היא מתcaptת סייבר וכיוצא"ב.

עבור ניבוי מחלות לב, בדקנו לחץ דם, קיומם חומרים רעלים בגוף וכו', אך באירועים שלנו נוכל רק להשתמש בשני פיצ'רים מפותח חוסר יכולת לציר גרפים בממדים גבוהים (ראו איור לדוגמה).



לבחירה מודל תלמיד צריך פ' עלות, בבעיות קלסיפיקציה נשתמש הפ' הכי אינטואיטיבית היא

$$L_S(h) = \sum_{i=1}^m \mathbb{1}_{y_i \neq h(x_i)} = |\{i : y_i \neq h_i\}|$$

אבל היא עדין לא תספק לנו את המשמעות הנדרשת לשגיאה, שכן במצבות יש הבדל בין טעויות שונות. לדוגמה, אישור תרופה מסוכנת חמורה הרבה יותר מפסילת תרופה בטוחה.

נסמן $-1 = y$ ה"שלילי" ו $1 = y$ ה"חיובי" ואז נחלק לשני סוגים של שגיאות:

- שגיאה סוג 1 (false-positive) $y = -1$ אבל הכרזו $y = 1$.

- שגיאה סוג 2 (false-negative) $y = 1$ אבל הכרזו $y = -1$.

נבחר כקונבנצייה שחיובי יהיה ההנחה המסוכנת, לדוגמה "האם התרופה הבאה בטוחה?" - אם תקרה לנו טעות מסווג ראשון, קרי false-positive, אז הוצאנו תרופה מסוכנת לשוק.

הערה: ככל בבחירה "חיובי" כקיצוני יותר נקבל ששגיאה מסווג ראשון היא תמיד חמורה יותר שגיאה מסווג שני, וכך נקבע תלמיד.

טרמינולוגיה בשמות דומים מדי אחד לשני

נסמן P מספר החיוביים, N מספר השיליליים, TP,FP,TN,FN הם ראשית התוצאות של מה שזה נשמע (וכיו"ב).

\bullet Error Rate : $\frac{FP+FN}{P+N}$ (חלוקת הניחושים השגוים).

\bullet Accuracy : $\frac{TP+TN}{P+N}$ (חלוקת הניחושים הנכונים).

\bullet Precision : $\frac{TP}{TP+FP}$ (חלוקת הניחושים החיוביים הנכונים מתוך הניחושים חיוביים).

• Recall : $\frac{TP}{P}$ (חלוקת הניחושים החיוביים הנכונים מtower החיוביים באמת).

• Specificity : $\frac{TN}{N}$ (חלוקת הניחושים השליליים הנכונים מtower שליליים באמת).

• False Positive Rate : $\frac{FP}{N}$ (חלוקת השגיאות מסוג ראשון מtower שליליים באמת).

בקלסיפירים, מעוניין אותנו מה גבול ההחלטה, כלומר עבור איזה גבול (על מישור במקורה הרבה ממד) מגדיר מה זה 1 – ומה זה 0. לא תמיד יש גבול ייחיד כזה כי אפשר להשתמש בפ' שמתנהגות מזרע עם אי רציפות וכו' אבל זה עדין מעוניין אותנו.

שאלות חשובות בבחירה קלסיפיר

1. מהי מחלוקת ההיפותזות? איך נראה גבול ההחלטה?

2. מה העיקרונו שליליו מתבססת הבחירה של $\mathcal{H} \in h_S$ בהינתן S ?

3. איך ממשיכים את עיקרונו הבחירה בקורס?

4. איך נשמר את המודל \mathcal{H} ?

5. בהינתן $\mathcal{H}, h_S \in \mathcal{H}$, איך נקבע מהו (x) עבור x דוגמה חדשה?

6. האם אפשר להבין למה המודל קבע את הגבול כפי שעשה?

7. האם הלומד נותן הסט' משוערכות לקביעות (כמו זה נראה כמו 1, כמו זה נראה כמו -1).

8. האם זה מודל ייחיד או שהוא דרש כיוול (מtower משפחה)?

9. באיזו קונטקטט השתמש בו?

קלסיפיר חצי-מרחב

הגדירה קלסיפיר חצי מרחב הוא פ' מהצורה $\{h_w : w \in \mathbb{R}^d\}$, מחלוקת ההיפותזות היא

הערה לעתה נניח כי $b = 0$ לשם פשוטות. למעשה הקלסיפיר נותן 1 בחצי מרחב אחד ו-1 – בחצי מרחב האחר (המכ'פ' על כיוונים שונים ביחס לעל-מישור משנה סימן).

עקרון למידה מה עיקרונו הלמידה? השתמש בפ' הערות הקלסית שסופרת שגיאות, שערכה | $\{i : y_i \langle w | x \rangle \leq 0\}$ | (אם הסימן של y_i לא תואם לניחוש שלנו זו שגיאה), ונשתמש בה לשם עיקרונו הלמידה ERM (העיקרונו שմזעර את הערות,

נניח כרגע $-S$ ניתנים להפרדה לינארית (לא ריאלייטי), כלומר אפשר לשים על מישור בין הדוגמאות השליליות לחובייתו, או במפורש, קיימים $w \in \mathbb{R}^d$ כך $\sum_{i \in S} y_i \langle w | x_i \rangle = 0$.

אם $w \cdot \sum_{i \in S} y_i \langle x_i | w \rangle > 0$ אז $L_S(w) \geq 1$ לכל i , כאשר w_0 הוא w כפול סקלר. נוכל לפטור זאת באמצעות simplex, שכן זו בעיית תכנון לינארית!

$$\text{נרצה למצוא } 0 \quad \underset{x \in \mathbb{R}^{d+1} \text{ s.t. } y_i \langle x_i | w \rangle \geq 1}{\operatorname{argmin}} \quad \text{(ולא simplex Perceptron)}$$

הערה אם הדוגמאות לא ניתנות להפרדה לינארית זו בעיה NP קשה כי המינימום הוא לא 0 אלא מספר כלשהו אחר.

ת"ז לקלסיפייר חצאי-מרחבי

- מחלקת היפותזות: חצאי-מרחבים.
- עיקרונו למידה: ERM על הנקודות שבסיד הלא נכון של העל-מישור.
- שימוש בקוד: תכונון לינארי, Perceptron
- איך לשמר את המודל: באמצעות וקטור המשקלות w .
- מתי משתמש: אף פעמיים כנראה, זה סתם יהיה מובא למסוגים מורכבים יותר.

Support Vector Machines

הلومד משתמש באוטה מחלקת היפותזות של חצאי-מרחבי רק שהפעם משתמש בעיקרונו למידה אחר שגם יהיה שימושי.

הערה נמשיך להניח $-0 = b$ אבל זה כמובן לא מציאותי.

הפעם במקומות להשתמש בפ' הערות הקלאסית, ננסה לבחור ישיר שכמה שפחות מתקרב לנקודות, ככלומר, שהוא מפריד ה' “חזק” בין האיברים.

הגדרה בהינתן על-מישור $\{v : \langle w | v \rangle = 0\}$

טענה אם $1 = |\langle w | x \rangle|$ אז $\|w\| = 1$.

מסקנה חישוב המרחק של הדוגמאות מהעל מישור היא חישוב מכ"פ ייחודי (אין סיבה לא להגיד את המשקלות בוקטור מנורמל).

הגדרה השול (margin) של על מישור ביחס לדוגמאות x_i הוא $\min_i |\langle w | x_i \rangle|$, וקטור שהמירחק שלו מהעל-מישור הוא השול נקרא support vector.

הערה עיקרונו הלמידה שלנו יהיה למקסם את השול.

עקרון למידה

- נניח שהדוגמאות ניתנות להפרדה לינארית. האלג' הلومד שלנו, Hard-SVM, מ Chapman

$$\underset{w: \|w\|=1, y_i \langle w | x_i \rangle}{\operatorname{argmax}} \min_{i \in [m]} |\langle w | x_i \rangle| \stackrel{\text{בתרגיל}}{\iff} \underset{w: y_i \langle w | x_i \rangle \geq 1}{\operatorname{argmin}} \|w\|^2$$

בעיה כזו היא בעיית אופטימיזציה קמורהRiboult עם אילוצים לינאריים (QP). יש אלג' יעילים לפתרון QP ככלומר הוא יעיל.

- במקרה הכללי, האלג' הלומד שלנו, Soft-SVM, מփש

$$\operatorname{argmin}_{w, \xi: y_i \langle w | x_i \rangle \geq 1 - \xi_i \wedge \xi \geq 0} \left(\lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right)$$

כלומר מוצא את המינימום לשקלול של המרחקים כי שראינו במקרה ה-Hard-SVM, יחד עם ממוצע ה- ξ_i , כאשר ξ משמעו

כמה נכנסו אל תוך המישור (ככל שהוא יותר גדול זה יותר רע) - $\xi_i = \text{הכוונה שאחננו בצד הנכון וככל שהערך יותר גדול כך ניתן}$

$$. y_i \langle w | x_i \rangle \geq 1 - \xi_i$$

ולשווים לו כמוה חשיבות ההפרות של המישור - אם לא מאוד קטן אז האלג' יצטרך לעבוד קשה כדי כמו שפחות להפר את השול,

ואילו אם לא מאוד גדול משמשו שחריגות מהשול לא קריטיות כי הרבה יותר חשובה שהעל מישור יהיה כמו שייתר חזק בין

הדוגמאות (רחוק מהן).

ולקרה משתנה רגולרייזציה, והוא למעשה עוזר להשפייע על הטרידיזוסוף הטיה-שונות (ולקטן יהיה עם שונות גבוהה כי הוא ינסה

לזוז כמו שאפשר כדי לא להפר את הגבול ואילו לא גדול יזם רעשנים קטנים ויראה את התמונה הגדולה, וכן יקטין שונות).

הערה אמנים נדמה ש- ξ , w נתונים לנו מודל אחד, למעשה הפכנו אותם למשפחה כשהוספנו λ שיכול להשתנות.

SVM ל-“z”

- מחלוקת היפטוזות: חצאי מרחבים.

- עיקנון הלמידה: שול מקסימלי.

- שימוש בקוד: אלג' לפתרון QP.

- קביעה על דוגמה חדשה: חישוב mc''_w עם w .

- איך לשמר את המודל: רק את w .

- משפחת מודלים: קיימת, עם פרמטר יחיד $\lambda \in [0, \infty)$.

- מתי להשתמש? מודל בסיסי וטוב לפני אחרים מורכבים יותר.

רגression לוגיסטי

הערה לעתת $\{0, 1\} = \mathcal{U}$ לנוחות.

ראיינו ברגression לינארית ש- $y_i = \langle x_i | w \rangle + z_i \sim N(\langle x_i | w \rangle, \sigma^2)$, נרצה לדמות משחו דומה כאן רק באופן דיסקרטי, לכן נמדל w_0 $w \in \mathbb{R}^{d+1}$, $p_i = \phi(\langle x_i | w \rangle)$ כאשר p_i לינארי (בערך) ב- x_i . כמובן, נניח שיש לנו פ' ϕ : הפיכה כך ש- $\phi(p) \rightarrow (0, 1)$: נסיף קוודינטה ל- x שהיא 1 לועפי כמו בהרצאה הקודמת.

סה"כ אנחנו מניחים שהדוגמאות ב"ת מקיימות $y_i \sim Ber(\phi(\langle x_i | w \rangle))$.

נשותמש בפ' הלוגיסטיית, $\pi(x) = \frac{e^x}{1+e^x}$ שהיא הפיכה ומונוטונית ולכון יפה.

מחלקת ההיפותזות שלנו היא $\{\pi(\langle x_i | w \rangle)\}$. נשותמש עקרון נראות מקסימלית (בחירה פרמטר שהחשת' לקבל את המדגם בהינתן פרמטר גובהה ביוטר). מתקיים מהיות הדגימות ב"ת

$$\mathcal{L}(Y=y | w) = \prod_{i=1}^m p_i^{y_i}(w) (1 - p_i(w))^{1-y_i}$$

כאשר $p_i = \pi(\langle x_i | w \rangle)$. עברו ללוג-לייקליהוד האהוב, נסמן $\ell = \log \mathcal{L}$ ונקבל

$$\ell(w | y) = \dots = \sum_{i=1}^m \left(y_i \langle x_i | w \rangle - \log \left(1 + e^{\langle x_i | w \rangle} \right) \right)$$

זו פ' קמורה, כלומר נוכל לחשב אותה (באופן יותר עיל מסתם בעיות אופטימיזציה קמורה). אנחנו רוצים למצוא

$$w = \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{i=1}^m \left(y_i \langle x_i | w \rangle - \log \left(1 + e^{\langle x_i | w \rangle} \right) \right)$$

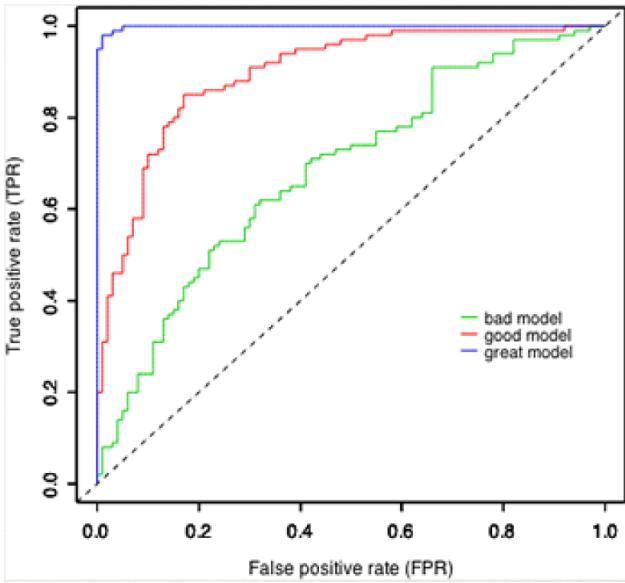
ירירסיה לוגיסטיבית היא interpretable, כלומר אפשר לבירר אילו פיצ'רים רלוונטיים וายלו לא, במקרה זהה באמצעות בדיקות גודל המשקלות המותאיימה להם - ככל שהיא גדולה יותר כך יש לה יותר השפעה. בנוסף, הניבואה מתבצע ע"י $\pi(\langle x_i | w \rangle)$ ואפשר לבירר "כמה" x_i הוא 1 או 0 וכך לדבג יותר בקלות המודל.

$\hat{y} = \begin{cases} 1 & \pi(\langle x | w \rangle) > \alpha \\ 0 & \text{אחרת} \end{cases}$ מסוווג אמנים לאנוון הסט' שזה 0 או 1 אלא תשובה ביןארית, לכן נציב פ' גף $\alpha \in (0, 1)$ והקביעה שלנו תהיה

- אם $1 \sim \alpha$ אז כמעט הכל יהיה חיובי ולכון יהיו לנו מעט מאוד false-positive אך גם מעט true-positive.
- אם $0 \sim \alpha$ אז כמעט הכל יהיה שלילי ויהיו לנו הרבה false-positive שזה לא טוב.

יש טריידוף בין רמת ה-true-positive וה-false-positive בבחירה ה- α .

כדי למדוד את הטריידוף הזה משתמשים ב-ROC, שהוא גраф עם צירים שמייצגים את יחס ה-true-positive-false-positive ולכל α מתאים זוג ערכים על הציריים האלה (ראו איור).



הישר באמצע מייצג את המקירה שבו אנחנו מטילים מטבב באקראי בקביעה שלנו ומצלחים בהסת' חצי לנחש נכון. העקומה שלנו היא מתחת לישר זהה כנראה שהמודל שלנו לא שווה כלום כי ניחוש שרירותי יותר מוצלח ממנו. ככל שהעקומה עולה יותר מהר למעלה, ככל המודל יותר טוב. נוכל בהינתן דרישות על ה-TPR, לגנות כמה נctrיך "לשלים" מבחינת FPR. אפשר לכמה את ביצועי המודל באמצעות השיטה מתחת עקומה, AUC, שיתן הערכה לאיוכות המודל. עם זאת, זה כימות פשטי של המודל וראוי לחשב יותר על איוכות המודל מעבר ל-AUC.

ת"ז לריגרסיה לוגיסטיבית

- **מחלקת היפותזות:** $H_{logi}^d = \{x \mapsto \pi(\langle x | w \rangle)\}$.
- **עקרון למידה:** נראהות מקסימלית.
- **מימוש בקוד:** פותר בעיות אופטימיזציה קמורה (ספקטיבי).
- **איך נקבע על דוגמה חדשה:** באמצעות פ' רג', $\hat{y} = \mathbb{1}_{h(x) > \alpha}$ כאשר $0 < \alpha < 1$.
- **ניתן לפירוש:** כן, באמצעות משקלות המודל.
- **משפחת מודלים:** רק אם מוסיפים או מורידים פיצ'רים.
- **איך שומרים את המודל:** רק צריך את w המשקלות.
- **מתי להשתמש:** תמיד כדאי.

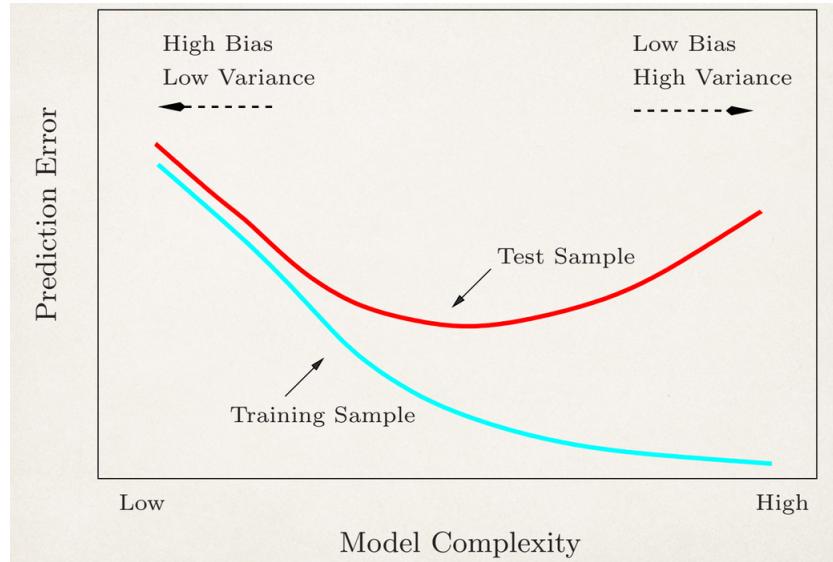
k-Nearst Neighbors

זהו מודל שונה לגמרי ממנו שלמדנו עד כה, כי אין לו מחלקת היפותזות או שלב אימון. הפרמטר היחיד שאפשר לשנות הוא $m \leq k \leq 1$ (לרוב אי זוגי).

בحينطن דגימה $x \in \mathbb{R}^d$, נמצא את k השכנים הקרובים ביותר ל- x ונחזר את הקביעה של רוב השכנים. נוכל להשתמש במרקח לפי נורמה אוקלידית, נורמה משוקללת או כל דבר אחר.

איך נבחר את k ? עבור $k=1$ לדוגמה, השונות מאוד גבוהה כי אנחנו מושפעים רק מדגימה אחת קרובה שיחסית שרירותית ביחס לאוסף הדגימות כולו. הגרף של 1-NN דומה לחלוקת לשטחי A,B,C,B,C בגדה - לא מודל שימושי מאוד.

אפשר לחשב את אחוז שהגיאיה וסטיית התקן שלו כפ' של מספר השכנים ואז למצוא את האחד שנutan את הטריידוף הכי טוב לטעמו. בכלל, זה מה שנרצה לעשות בפועל במידה - למצוא את מינימום הטריאידוף (ראו אייר).



איך נחשב את השכנים הקרובים ביותר?

1. ברוט-פורס : חישוב מרחקים על כל הדגימות.
2. בניית נתונים מייעל : בשל המקרים הנשмар את כל הדגימות במבנה נתונים שעזר לנו לחשב מרחקים (לדוגמה k-d tree) וכך נמצא שכנים בייעילות, ונזורק את דגימות האימון אחרי.
3. שיטות הסט' : בהסת' גובהה נוכל בייעילות לחשב את השכנים הקרובים ביותר.

ת"ז k-NN

- אין מחלוקת היפותזות.
- אין עיקרונו למידה.
- שימוש בקוד : בניית נתונים/הסט'/ברוט-פורס.
- שערוך הסט' קביעה : יש רק אם משתמשים בהרבה שכנים.

- משפחת מודלים : כו, עם פרמטר k .

- איך שומרים את המודל : או עם מבנ"ת, או את האוסף שלו.

- מתי להשתמש : תמיד לנסות.

עצי סיווג

יעיר מקרי הוא מודל מאוד פופולרי של לומדים, ואפשר להשתמש בו גם לריגרסיה וגם לסיווג. מעבוד על בעיות סיווג ונחלה את הבנייה שלו בשלושה שלבים : גידול (ההרצאה הזו) ; ניזום ; והפרדה (בהרצאות הבאות).

חלוקת עצים היא חלוקה של המרחב האוקלידי (איחוד קבוצות זרות) כאשר כל חלוקה מוגדרת ע"י הפרדה של המרחב הנתון אל באמצעות ישר שמקביל לאחד מהציריים e_1, \dots, e_d , וניתן לייצג כל חלוקה כזו ע"י עץ ביןארי. אם נקבע לכל קבוצה בחלוקת ערך $(1, 0)$, אז הגדרנו כלל החלטה לפי הקבוצה בה הדגימה הנדרשת נמצאת.

$$h(x) = \sum_{j=1}^N c_j \mathbb{1}_{B_j}(x), \quad \text{הינתן עץ החלטה } h \in \mathcal{H}_{CT}, \quad \text{נקבע לכל קבוצה } B_j \text{ תווית } c_j \in \{0, 1\} \text{ ולכן המודל שלנו הוא}$$

מחלקת ההיפותזות שלנו היא \mathcal{H}_{CT} המכילה כל פ' מהצורה הנ"ל . נגביל את מחלקת ההיפותזות שלנו לחלוקות עם עומק לכל היותר k בעץ הביארי המיצג אותן. כל פ' $h \in \mathcal{H}_{CT}$ מיצגת עץ החלטה באופן אינטואיטיבי, שגם תואם למשמעותו אונשי על עץ החלטה בידי מקבלי החלטות שצרכיהם לקבוע מדיניות (האם לחתם חמצן למטופל לדוגמה).

מהו עיקנון ההחלטה שלנו? ERM יכול להיות בעיתי כי אם הוא קובע קופסה קטנה סביב כל דגימה ומזער לגמרי את הרиск האמפירי לא עשינו בזה כלום, מה גם שחייב כל העצים הללו הוא NP-קשה. لكن נפנה להיוריסטיקה, שקובעת באופן בלתי מוכח מהו $h \in \mathcal{H}_{CT}^k$ "טוב" (המקבילה של רפואה משלימה לתרופות אמיתיות).

יש כל מיני היוריסטיות אבל אנחנו משתמשים ב-CART (Classification And Regression Trees). היוריסטיקה זו מסבירה לנו כיצד לגדל עץ זה ווגם איך לגוזם אותו אחריו, נעסק ברגע בගידולו. CART הוא אלג' חמדן - בכל חלוקה באמצעות ישר המקביל לציר כלשהו נבדוק את כל הערכים האפשריים עבורו (בין דגימות) ובחר את האחד שמקטין כמה שיותר את הרиск האמפירי (ERM) ונצדיד לו תווית לפי הצבעת רוב, כך עד שנגיע לעומק המקסימלי. אף על פי שהסבירו כראות גדולה אפשר למסח את זה מאד עיל.

ת"ז לעצי סיווג

- מחלוקת היפותזות : פ' קביעות למקוטען על קביעות ב- \mathbb{R}^d המיצגות ע"י עץ בחירה בעומק לכל היותר k .

- עקרון מידת : (דרך ERM).

- שימוש בקוד : מימוש קלסי ופשוט של CART.

- קביעה בהינתן דגימה חדשה : ריצה בעץ הסיווג.

- ניתן לפירוש : מאוד (כל מסלול לעלה מסביר מה הקביעה).

- שערוך הסת' קביעות : אין.

- משפחת מודלים : כו, עם פרמטר k - עומק עץ הסיווג.
- מתי להשתמש : כבסיס או כפרשון הוא קריטי - עצי סיווג הם בסיס ליער מקרי, שנראה בהמשך.

תרגול

סכמה לפתרון בעית למידה

1. תרחיש : הגדרת הבעיה - בבהינתן הדגימות (p.i.) שנותנות לנו, נרצה למצוא את האומד המתאים ביותר למ"מ.

$$\text{לדוגמה, } x_1, \dots, x_m \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2).$$

2. מצום מרחב הפ' : הגדרת \mathcal{H} , מחלוקת ההיפותזות שלנו.

$$\mathcal{H} = \{\hat{\mu} : \mathbb{R}^m \rightarrow \mathbb{R}\}.$$

3. עקרון הלמידה : איך נבחר את $h \in \mathcal{H}$ - מה מוגדר כמתאים ביותר?

בדוגמא הנ"ל, העקרון היה נראהות מקסימלית, כלומר $\hat{\mu} = \operatorname{argmax}_{\mu \in \mathcal{H}} \hat{L}(\mu | X)$, בירגרסיה ליניארית העקרון הוא אחר.

4. אלגוריתם : מציאת $h \in \mathcal{H}$ לפי עיקנון הלמידה (החלק התכנוני).

בדוגמא הנ"ל, אחרי ניתוח אורך תאורטי הענו לכך שצריך למש ממוצע על מערך בטח'כ.

דוגמה נפעיל את הסכמה על ריגרסיה ליניארית :

1. תרחיש : נתון $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ התגובה (מה הנביוחת), נרצה מיפוי טוב $y \xrightarrow{f} \mathcal{X}$.

2. מצום : בירגרסיה ליניארית מחלוקת ההיפותזות היא

$$\mathcal{H} = \{f(x) = x^T w + b : w \in \mathbb{R}^d, b \in \mathbb{R}\} = \{x \mapsto x^T w : w \in \mathbb{R}^{d+1}\}$$

כאשר בהגדירה השנייה מגדרים $x = (1, x)$ ו- $w = (b, w)$.

3. עיקנון הלמידה : נרצה למצוא w כך ש- $Xw = y$ (כאשר $X \in \mathbb{R}^{m \times (d+1)}$, $y \in \mathbb{R}^m$ הם הנתונים).

• רייאלייזבילי : קיימים $w \in \mathbb{R}^{d+1}$ שמקיים זאת.

• לא Reiailizyibeli : לא קיימים w שמקיים את מערכת המשוואות בין אם כי מה שאנחנו מנסים לקרב הוא לא לינארי או שיש בו רעש, אבל עדין ניתן מודל לינארי כדי לקרב את הבעיה.

בגלל שלא נוכל תמיד למצוא $w = X$, נגיד מה זה "הכי קרוב" לקיים המשווהה הזה. בעקרון שלנו נשים זאת באמצעות מצום פ' עלות, במקרה שלנו פ' העלות היא הסטיות הריבועיות,

$$l(w, (x, y)) = (w^T x - y)^2$$

והפ' שאותה נמצאים היא פ' העלות האמפירית,

$$L_S(w) = \sum_i l(w, (x_i, y_i))$$

במקרה הרלייזאbialי, נוכל למצוא w עבורו $L_S(w) = 0$

העקרון שלנו הפעם נקרא Empirical Role Minimization, כלומר,

$$\hat{w} = \underset{w \in \mathcal{H}}{\operatorname{argmin}} L_S(w) = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \sum_i (x_i^T w - y_i)^2 = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \|Xw - y\|_2^2$$

4. אלגוריתם : כדי למצוא את המינימום מטרך למצוא קודם מתי הגריאנט מתאפס :

$$\begin{aligned} \nabla RSS(w) &= J_w (RSS)^T \\ &= J_w \left(\|Xw - y\|_2^2 \right)^T \\ &\stackrel{(*)}{=} (J_{g(w)}(f) J_w(g))^T \\ &= (2(Xw - y) X)^T \\ &= 2X^T(Xw - y) = 0 \end{aligned}$$

$$. f(z) = \|z\|_2^2, g(w) = Xw - y \text{ כאשר } RSS = f \circ g (*)$$

כלומר אנחנו מחפשים w כך ש- y - X . כלומר שאם $\varphi_0, \dots, \varphi_d$ הן העמודות של X אז מתקיים

$$.\hat{y} - y = Xw - y \in (\operatorname{Im} X)^\perp$$

$$\bullet \text{ אם } X \text{ הפיכה, ולכן } X^T X \text{ הפיכה ולכן}$$

$$H(RSS) = \nabla(\nabla RSS) = \nabla_w (X^T X w - X^T y) = \nabla_w (X^T X w) = X^T X \succeq 0$$

ולכן כל RSS קמורה ולכן נקודת קיצון היא מינימום גלובלי ככלומר \hat{w} אכן המינימום שאנו מחפשים.

\bullet אם X לא הפיכה ($X \in \mathbb{R}^{(d+1) \times m}$, $m \geq d+1$), תחת ההנחה ש- SVD את $X = U\Sigma V^T$, $r(X) = r < d+1$,

ולכן

$$X^T X w = X^T y$$

$$(U\Sigma V^T)^T (U\Sigma V^T) w = (U\Sigma V^T)^T y$$

$$U^T \Sigma^T V^T V \Sigma U^T w = U^T \Sigma^T U^T y$$

$$\Sigma^T \Sigma V^T w = \Sigma^T U^T y$$

$$\begin{aligned}
X &= U\Sigma V^T \\
&= \begin{pmatrix} \vdots & & \vdots \\ u_1 & \cdots & u_m \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_r \\ 0 & & & \ddots & 0 \end{pmatrix} \begin{pmatrix} \cdots & v_1 & \cdots \\ & \vdots & \\ \cdots & v_{d+1} & \cdots \end{pmatrix} \\
&\stackrel{(*)}{=} \left(\begin{array}{c|c} \underline{U_R} & \underline{U_N} \\ \hline (d+1) \times r & (d+1) \times (m-r) \end{array} \right) \left(\begin{array}{c|c} \frac{\underline{S}}{r \times r} & 0 \\ \hline 0 & 0 \end{array} \right) \left(\begin{array}{c|c} \underline{V_R^T} & \underline{V_N^T} \\ \hline r \times m & r \times (m-r) \end{array} \right) \\
&= U_R S V_R^T
\end{aligned}$$

הן מטריצות שעמודותיהן הן בסיסים U_R, V_R ($*$) בהתאמה, U_N, V_N מטריצות שעמודותיהן בסיסים של $\ker U, \ker V$ בהתאמה. אנחנו יכולים לעשות את המעבר הזה פשוט ע"י מצום מימי הטרנס' במאוץ (המעבר ל- U) כי כל הוקטורים שעוברים דרך U_N, V_N ובלוק האפסים ב- Σ מותאפסים ולא מעניינים.

הגדרה הפירוק SVD קומפקטי.

הגדרה תהי Moore-Penrose ההפסודו-הופכית של X לפי $X^\dagger \in \mathbb{R}^{(d+1) \times m}$.

$$X^\dagger = V_R \Sigma^{-1} U_R^T \in \mathbb{R}^{m \times (d+1)}$$

כאשר מוגדים מפירוק SVD הקומפקטי של X .

מתקיים

$$\Sigma^T \Sigma = S^2 = S^T S$$

נשתמש בפירוק הקומפקטי $(\Sigma = S, U = U_R, V = V_R)$ ויחד עם המשוואות לעיל (בחצבת $X = U_R S V_R^T$) ובהזיהות S בדרגה מלאה, היא הפיכה ולכן ניתן לכפול בהופכית שלה וב- $V_R^{-1} = V_R^T$ וובקרה הלא ריאלייזאבל, נבחר את האומד שלנו להיות

$$\hat{w} = V_R S^{-1} U_R^T y = V \Sigma^\dagger U^T = X^\dagger y$$

: $\text{Im } X$ נציג מנוקדת מבט אחרית את מינימליות $RSS(X^\dagger y)$. עברו המקרה הריאלייזבילי, קל להראות ש- \hat{y} הטלה אורתוג' על

$$\hat{y} = X \hat{w} = \frac{X (X^T X)^{-1} X^T y}{P_{\text{Im } X}(y)}$$

$$RSS(\hat{w}) = \|\hat{y} - y\| = \|X\hat{w} - y\| = \|P_{\text{Im } X}(y) - y\|$$

ובגלל ש- (y) הוא הוקטור חci קרוב ב- X ל- y ולכן מינימלי הריש(\hat{w}) מינימלי לפי האלג' שלנו.

שבוע VII | מסגרת PAC

הרצאה

בURITY קלסיפיקציה היא בעיה בה נתון לנו דוגמא $\mathcal{A} = \mathbb{R}^d$ אבל לא בהכרח, יכול להיות טקסט או תמונה) ו- \mathcal{Y} אוסף לייבלים (דוגמיה $\{0, 1\}$ הצלחה/כשלון) ואנחנו מחפשים כלל החלטה $\mathcal{Y} \xrightarrow{h} \mathcal{A}$ מוצלח.

הגדרה לומד הוא אלג' העתקה $(\mathcal{Y} \rightarrow (\mathcal{A} \times \mathcal{Y})^m \rightarrow \mathcal{A})$.

הערה לפעמים הכוונה ב"אלגוריתם" היא ל- \mathcal{A} ולפעמים ל- h .

מה המטרה של הלומד? אינטואטיבית, ש- h יצדוק לדגימות בעtid. נctrיך נוסחה מתמטית שקובעת איך נוצרות דוגימות חדשות כדי לקבוע האם h מוצלח עבורה.

למד תורה חישובית תאורטית למורי ונענה על שאלות בסיסיות: מה אפשר ללמוד ומה לא? כשאפשר ללמוד, כמה דוגימות אימון צריך?

הערה במודל הlienרי ומקרים אחרים לא הנחנו שום דבר על התפלגותן של הדוגימות אבל בעצם זה נדרש.

נניח שיש הפלגות \mathcal{D} מעל \mathcal{A} כך x_i - x_m נוגם מ- \mathcal{D} באופן ב"ת באחרים, כולם $\mathcal{D} \stackrel{\text{i.i.d.}}{\sim} x_1, \dots, x_m$. בנוסף, נניח שיש כלל אמייתי שעבורו $y = f(x)$ נתון לנו אוסף אימון $s = \{x_i, f(x_i)\}_{i=1}^m$.

הגדרה העלות המוכללת על כלל החלטה כלשהו בהינתן כל החלטה נכונה והפלגות היא ההסת' שהכל לא יסכים עם הכלל האמייתי בהינתן ההסת', כולם,

$$L_{\mathcal{D},f}(h) = P_{x \sim \mathcal{D}}(h(x) \neq f(x)) = P(\mathcal{D} \in \{x \in \mathcal{A} : h(x) \neq f(x)\})$$

עלות זאת נקראת גם הסיכון או misclassification error

הערה f לא ידועות לנו.

הערה צריך לזכור שהסיכון לא מתייחס לשגיאות מסווג ראשוני ושני אבל כבר למדנו שהוא כן חשוב.

סבירום ביןיגים יש לנו $S = \{(x_i, y_i)\}_{i=1}^m$ שוגרים מ- \mathcal{D} שמקיימים $y_i = f(x_i)$ ואנחנו מחפשים \mathcal{A} שהינתן S מחזיר $L_{\mathcal{D},f}(h)$ שמשמעותו $h : \mathcal{A} \rightarrow \mathcal{Y}$

נכזה להבין לעומק בהרצאה את המושגים PAC-למידתיות, מרכיבות דגימות ומימד-VC של מחלוקת היפותזות $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$.

הגדרה נאמר כי \mathcal{H} היא PAC-למידה אם קיימת $f : \mathbb{N} \rightarrow (0, 1)^{\mathcal{H}}$ וalgo' למידה \mathcal{A} שמקיימים את התוכנה הבא:

לכל $0 < \epsilon, \delta < 1$, הטענות \mathcal{D} על \mathcal{X} וכלל החלטה $\{ \pm 1 \}$ $h^* \in \mathcal{H}$ המקיימים $L_{\mathcal{D}, f}(h^*) = 0$, כמפורטים

את \mathcal{A} על יותר מ- $\tilde{m}_{\mathcal{H}}(\epsilon, \delta)$ דגימות i.i.d. מתוך \mathcal{D} עם ליבלים מ- f , האlg' מחזיר היפותזה $h_S = \mathcal{A}(S)$ (כלל החלטה) כך שהסת'

$$L_{\mathcal{D}, f}(h_S) \leq \epsilon.$$

הערה לפני המשכתי לראות את ההרצאה, אינטואיטיבית מה שקרה כאן זה הדבר הבא: לכל סיטואציה שבה יציבו אותנו (\mathcal{D}, f) יש מספר דגימות מינימלי כלשהו שם יש לנו מספר דגימות כזה, ככל ביחס' גבואה בכל שרגצה ($\delta - 1$) לקבל עלות נמוכה ככל שרגצת (ϵ).

הערה מחלוקת היפותזות (אוסף כללי החלטה לאלו ששכחו) היא או PAC-למידה או לא PAC-למידה, זו הגדרה בינהית.

הערה מחלוקת היפותזות יותר מדי מרכיבות לא יהיו PAC-למידות בעוד עם מחלוקת פשוטות ניתן יהיה ללמוד.

הגדרה תהי \mathcal{H} מחלוקת היפותזות PAC-למידה ו- $\epsilon, \delta \in (0, 1)$. נגידר את סיבוכיות המודגם של \mathcal{H} עבור δ, ϵ להיות המספר המינימלי של דגימות $m_{\mathcal{H}} : (0, 1)^{\mathcal{H}} \rightarrow \mathbb{N}$ שבעורן מתקיימת ההגדרה הנ"ל עבור δ, ϵ . פ' סיבוכיות המודגם של \mathcal{H} מסומן ע"י

הגדרה תהי $\mathcal{H} \subseteq \mathcal{X}$ (חלוקת היפותזות בעיית סיוג). עבור $\mathcal{X} \subseteq \mathcal{H}$, נגידר $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\} \subseteq \mathcal{H}$ כאשר לכל $\mathcal{Y} \subseteq \{\pm 1\}^{\mathcal{X}}$ $C \in \mathcal{H}$ הינה החיצונים של h_C היא C -ל- \mathcal{H} . מימד ה-VC של \mathcal{H} מוגדר להיות

$$\text{VCdim}(\mathcal{H}) = \max \left\{ |C| : C \subseteq \mathcal{X} \wedge |\mathcal{H}_C| = 2^{|C|} \right\} \leq \infty$$

הערה כל ההגדרות האלה מגיעות בסופו של דבר מובילות אותנו למשפט היסודי של למידה סטטיסטי שעונה על השאלה מתי אפשר ללמוד מה מספר הדגימות המינימלי ואייך ללמוד.

משפט \mathcal{H} היא PAC-למידה אם $\text{VCdim}(\mathcal{H}) < \infty$. בנוסף, סיבוכיות המודגם של מחלוקת היפותזות עם מימד-VC סופי היא בקשר

$$m_{\mathcal{H}}(\epsilon, \delta) \sim \frac{\text{VCdim}(\mathcal{H}) + \log \frac{1}{\delta}}{\epsilon}$$

הערה עיקרון הלמידה ERM מגיע למינימום של סיבוכיות המודגם הנ"ל. ככלומר כשלמידה היא אפשרית, ERM לומד עם מספר קרוב למינימום של דגימות.

הערה אנשים שמכורים למשפט יסודיים חוזים relapse כשהם מבינים את המשפט היסודי של הלמידה הסטטיסטי.

ນחשוב על המסקנת בתורו משחק: אנחנו בוחרים את \mathcal{A} והטבע בוחר את f .

גרסת המשחק 1.0

נקבע מרחב דגימות m . נבחר אסטרטגיה \mathcal{A} . הטבע יודע מה האסטרטגיה שלנו ואחרינו, בוחר \mathcal{D} ו- f . בורר (שופט) מגעה ודוגמ m דגימות מתוך \mathcal{D} עם ליבלים מתוך S . מועבר ל- \mathcal{A} שמחזיר כלל החלטה $h_S = \mathcal{A}(S)$. התשלום שלנו הוא $L_{\mathcal{D}, f}(h_S)$, למעשה תוחלת החלוקת של שגיאות סיוג- S עשויה על דגימות i.i.d. התשלום אקראי כי S אקראית ואז גם h_S אקראית.

נניח שהטבע רשע ופועל נגדנו ולכון מוצא f , כמה שיותר גרוועים ביחס לומד שלנו \mathcal{A} , ולכון נחפש לומדים שיש להן עלות המקסימלית מובטחת (h) לכלבחירה של הטבע $L_{\mathcal{D},f}$.

כלומר, מניחים שבהינתן \mathcal{A} , הטבע יבחר f , \mathcal{D} שיקשו עליינו להכליל מ- S ובכל מקרה יש סיכוי לקבל S שלא מייצג. נרצה לדעת האם יש לנו סיכוי לנצח, כלומר שתמיד יהיה תשלום "טוב" (נמוֹך) בלי קשר לאיך הטבע משחק.

Probably, Approximately, Correct

שאלה האם יש אסטרטגיה שתבטיח חסם עליון $1 < \epsilon \leq 0$ על העלות בהסת' 1 (על S מקרי) עבור כל f, \mathcal{D} ? לא. יהיו $1 < \epsilon < 0$. הלומד לא יכול לייצר בלי תלות במה הטבע עושה בהסת' 1 כל החלטה עם $\epsilon \leq L_{\mathcal{D},f}(h)$. זאת מושם שיש סיכוי קטן מאוד שנתקבל מוגם. אימונן פתולוגית שלא מייצג את \mathcal{D} בכלל ואז h_S טוענה עבור רוב \mathcal{A} . אם S ממש גרווע, העלות של S h_S יכולה לעלות עד 1 (לא כולל כMOV).

דוגמה יהיו $\{x_1, x_2\} = \mathcal{A}$ ו- $0 > \gamma$. הלומד שלנו צריך לקבוע מה הוא אומר על דגימה שהוא עוד לא ראה +1 (אחרת נהפוך את הסימנים). הטבע משחק עם \mathcal{D} שמחזיר x_1 בהסת' $\gamma - 1 - x_2$ בהסת' γ וכלל אמייתי $h_S(x_1) = +1$ $h_S = \mathcal{A}(S)$. במקרה כזה, במקרה x_2 הוא יטעה בהסת' γ . במקרה, עבור γ מספיק קטן, $\epsilon > \gamma - 1$. ככלומר, $L_{\mathcal{D},f}(h) = 1 - \gamma$

כלומר, גם על דוגמה מאוד בסיסית, הטבע יכול לנצח כל אסטרטגיה בהסת' לא 0.

הגדרה נאמר כי \mathcal{A} הוא לומד PC (Probably Correct), כנראה צודק אם ל- \mathcal{A} יש לו פ' עלות גדולה באופן שריוןתי בהסת' לכל היותר $\delta > 0$ ובמקרה זה נאמר כי יש לו ביטחון δ (confidence).

הערה ככלומר, אנחנו מאפשרים ל- \mathcal{A} להיכשל לחלווטין בהסת' נמוכה בלבד (delta).

שאלה במקרה לא פתולוגי (בהסת' $\delta - 1$), האם יש אסטרטגיה שתוכל להקנות לנו עלות 0 בהסת' $\delta - 1$ בכל מקרה שהוא? לא. יהיו $0 > \delta$. הלומד לא יכול לייצר עלות 0 בהסת' $\delta - 1$ כי זה אומר ש-0 $L_{\mathcal{D},f}(h) = 0$ ככלומר $P(\mathcal{D}) \in \{h_S(x) = f(x)\} = 1$ ואילו אם \mathcal{D} נבחר כך ש- $\mathcal{A} \in x$ הוא בעל הסת' זניחה, במקרה S לא יכול את x ואז h_S לא יודע מה לעשות לגביו.

דוגמה במקרה הדוגמה לפני, אם נקבע $0 > \delta$, ההסת' לא כולל את x_2 ב- S היא $\gamma - 1$ (ולכן לא יוכל לקבוע נכון על x_2 בהסת' יותר גדולה מ- δ עבור γ מספיק קטן).

הגדרה נאמר כי \mathcal{A} הוא לומד AC (Approximately Correct), בערך צודק אם קיים חסם עליון $0 > \epsilon$ על העלות ובמקרה זה נאמר כי יש לו דיוק ϵ .

סיכום ביניים ראיינו עד כה שצריך להתחשב גם במקרה של כישלון קולוסאלי וגם במקרה שפיספסנו דגימות זניחות. נרצה לאפשר מקרים כאלה.

הגדרה נאמר כי \mathcal{A} הוא לומד PAC אם יש עלות של S ($h_S = \mathcal{A}(S)$) חסם עליון $0 > \epsilon$ בהסת' לפחות $\delta - 1$ עבור $0 > \delta$, ככליהם ובעל עלות גבוהה באופן שריוןתי (יותר מ- ϵ) בהסת' לכל היותר δ ובמקרה זה נאמר כי יש לו δ ודיוק ϵ .

כלומר, $\mathcal{A} : S \rightarrow h_S$, $f \in \mathcal{D}$, ϵ ו- δ אם לכל

$$P_{\mathcal{D}^m} (\{S \in (\mathcal{X} \times \mathcal{Y})^m : L_{\mathcal{D},f}(h_S) \leq \epsilon\}) > 1 - \delta$$

הערה האקרαιות מכה פעמיים: פעם ראשונה כ- S -מוגרבאקראי ואו δ היא ההסת' שניכשל בהינתן S מזר (לא יציג) ופעם שנייה כشدיגיות חדשות מוגרלות באקראי מתוך \mathcal{D} ואז החסם על השגיאה הוא ϵ . הסטודנטית המשקיעה תקרה את המשפט הזה לפחות $\frac{1}{\epsilon}$ פעמים עד שתבין לעומק את הבדל ביניהם.

גרסת המשחק 2.0

מעתה m לא קבוע מראש, אלא δ, ϵ , כלומר ידוע לנו מה הדרישות שלנו, אנחנו נוכל לבחור באסטרטגיה שלנו את \mathcal{A} (כמו לפני) אבל עכשו גם את m כדי לנסות לעמוד בדרישות הדיקוק והביחוח. למעשה יש לומד A_m לכל m שנבחר אבל ממשיך להתייחס לזה בתור לומד אחד \mathcal{A} . לכל $0 < \delta, \epsilon$ נחקק משחק נגד הטבע עם תשלום אקראי. נבחר $\mathcal{X} \times \mathcal{Y}^m \rightarrow \mathcal{A}$: גודל מוגם m שתלוויים ב- (ϵ, δ) . הטבע ידוע מה האסטרטגיה שלנו ובוחר \mathcal{D} ו- f . השופט דוגם m דגימות מתוך \mathcal{D} , אנחנו נותנים לו את h כל החלטה והתשלום שלנו הוא (h_S) שהוא אקראי כי S ולכנו h_S אקראי).

נניח שהטבע רשע ופועל נגדנו ולכנו מփש לומדים \mathcal{A} עם עלות מקסימלית מובטחת (h) לכל אסטרטגיה f שהטבע בחר. נחקק את המשחק הזה הרבה פעמים ונספר כל פעם את את ההסת' $\{S \sim \mathcal{D}^m : L_{\mathcal{D},f}(h_S) \leq \epsilon\}$ על פני מוגם אימון S . אם מוצאים שהסת' הזה גבוהה מ- $\delta - 1$ (כלומר ש- \mathcal{A} הוא לומד PAC עם דיקוק ϵ ו- δ), נאמר שניצחנו, בלי תלות בערכיים הספציפיים של δ .

נרצה לבחור את m להיות קטן ככל שניתן כל עוד \mathcal{A} הוא עדין לומד PAC. ככל $\delta - \epsilon$, קטנים, m גדל (ככל שהדרישות נוקשות, נדרש יותר מידע).

אי אפשר לנצח את גרסת המשחק השנייה במקרה כללי (עבור δ, ϵ , כלשהם) כי אם לא ידוע לנו \mathcal{D} או f יש יותר מדי אפשרויות עבור f ואז לא משנה כמה גדול m , לא נוכל להיות מספיק בטוחים שנמצא כל החלטה מדויק מספיק h_S .

דוגמה נניח כי $\mathcal{X} = \mathcal{A}$. נניח שבחרנו m כללי. נבחר ב- (x) את ההחלטה שלנו על דגימה חדשה כלשהי. הטבע בוחר $\mathcal{X} \subseteq C$ כך $|C| > 2m$ (אפשרי כי הקבוצה לא סופית) ומוציאר $\mathcal{D} = \text{Unif}(C)$ (כלומר קיבל אחד איברים מ- C ובהסת' 0 כל איבר אחר). ככל החלטה האמיתית f מוגדר ע"י, $f(x) = -g(x)$ לכל $x \in \mathcal{X} \setminus S$ (תמיד נטעה עבור דגימות שלא ראיינו). ברור שאנו טועים על כל $C \setminus S$.

הסטודנטית המשקיעה תשים לב שיש לנו שגיאה מתמטית איפשהו (אין לי מושג מה היא) ואם היא רוצה להבין לעומק את החומר היא תצטרך לחשב טוב טוב.

עבור S מוגם אימון, נסמן $\text{supp}(S) \subseteq C$ את כל הנקודות שקיבלו ב- S ומהיות $m \leq |\text{supp}(S)|$ (יכולות להיות כפליות) ו- \mathcal{D} אחיד על C , מתקיים $P_{\mathcal{D}}(\{x \in \mathcal{X} \setminus \text{supp}(S)\}) \geq \frac{1}{2}$. כאמור, ההסת' לדגימה שלא ראיינו גדולה ממחצית. בעצם, \mathcal{A} מוציאר לנו כל החלטה $h_S = \mathcal{A}(S)$. הצלות היא גדולה ממחצית כי בהסת' יותר מחצית אנחנו מקבלים דגימה מ- (S) כלומר דגימות שלא ראיינו ועבור כולם אנחנו טועים כאמור, لكن $L_{\mathcal{D},f}(h_S) \geq \frac{1}{2}$

זה קורה לכל מודם S ולכל אם הינו מתחשים לומד PAC \mathcal{A} עבור δ, ϵ , כלשהם בלתי תלות באסטרטגיה של הטבע f , לא הינו מוצאים אחד כזה אף פעם. בנוסף m גדול יותר לא יהיה עוזר כי פשוט הינו בוחרים C גדול יותר ($-D$ אחדה עליו). הבעה כאן הייתה ש- f יכול להיות כל מה שהטבע רוצה, וזה מקשא עליו מאוד.

אין ארכות חיים

משפט (אין ארכות חיים) יהיו \mathcal{X} מרחב מודם אינסופי ו- $\mathcal{Y} \subseteq \mathcal{H}$. אזי קיים $0 < \delta < \epsilon$ כך שלכל לומד \mathcal{A} עם מודם אימון m , קיימת התפלגות D על \mathcal{X} וכל החלטה אמיתית $\mathcal{U} \rightarrow \mathcal{X}$: כך שביחסו לפחות δ על S ה- $L_{D,f}(h_S) \geq \epsilon$ כאשר $L_{D,f}(S) \leq \epsilon$.

הערה פשוט הכלנו את מה שראינו זה עתה - אם \mathcal{X} אינסופי ולא ידוע לנו כלום על f , נוכל תמיד להרים את המסיבה.

כדי להיות יכולים ללמידה, הלומד חייב לקבל מידע מקדים על מחלוקת היפותזות $\mathcal{Y} \subseteq \mathcal{H}$. נניח את הנחת הרילאייזביליות, כלומר, נניח ש- H ידוע בתחילת המשחק והוא הטבע חייב לבחור $\mathcal{H} \in f$. למעשה, מספיק שהוא יבחר $\mathcal{Y} \subseteq f$ שווה כמעט תמיד $\mathcal{H} \in h^*$ עבור התפלגות D , כלומר ש- f בוחר פ' כך שקיים $\mathcal{H} \in h^*$ שקיימים $0 = L_{D,f}(h^*)$.

הلومד יודע את \mathcal{H} בתחילת המשחק ויזיר רק $\mathcal{H} \in h_S$, כלומר עבשו \mathcal{A} הוא העתקה $\mathcal{H} \rightarrow (\mathcal{X} \times \mathcal{Y})^m$. ואם $\infty = |\mathcal{X}|$ או $\mathcal{H} = \mathcal{Y}^\mathcal{X}$ היא גודלה מדי שמלמד, כלומר לא קיים m שעבورو נוכן ללמידה.

עבשו עלות לנו שאלות חדשות: מי הון \mathcal{H} שהן מספיק קטנות שעבורי קיימים לומדי PAC? אילו גודלות מדי שמלמד?

נניח שיש לנו \mathcal{H} "קטנה מספיק". לכן לכל δ, ϵ קיימת לפחות אסטרטגיה אחת \mathcal{A}, m כך ש- \mathcal{A} הוי לומד PAC עם דיק ϵ וביתחון δ . לכן קיים m מינימלי עבור הפרמטרים הנ"ל, האם נוכל לאפיין את $(\epsilon, \delta)_H$ זהה? האם יש קשר בין הגודל של \mathcal{H} ל- m_H ? האם נוכל לפרט מי הוי \mathcal{A} הספציפי שהיה לומד PAC על \mathcal{H} ? כמה דוגמאות אימון יctrck \mathcal{A} כזה כדי ללמידה PAC? האם אפשר למצוא את הלומד הכי יעיל מבחינת מספר דוגמאות האימון הנדרשות, כלומר לומד שדורש $(\epsilon, \delta)_H$?

גרסת הלמידה 3.0

נתונים $0 < \delta, \epsilon < 1$ מחלוקת היפותזות. נבחר גודל מודם m ולומד $\mathcal{H} \rightarrow (\mathcal{X} \times \mathcal{Y})^m$. שיכולים להיות תלויים ב- (ϵ, δ) . הטבע בוחר התפלגות D על \mathcal{X} וכל החלטה אמיתית $\mathcal{Y} \subseteq f$ שווה כמעט תמיד לפי D לפ' \mathcal{H} . השופט בוחר m דוגמאות מתוך D עם לייבלים f, S . הוא מקבל $L_{D,f}(h_S) = \mathcal{A}(S)$ והתשלום שלו הוא \mathcal{H} . אותן הנחות כמו הגרסה הקודמת חלות.

האם נוכל ללמידה כאשר $\infty = |\mathcal{H}|$? כן, גם אם $\infty = |\mathcal{X}|$ עדין אפשר ללמידה, כלומר עוצמה היא לא המדי גנדול/קטן מספיק.

דוגמה נבחר $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0, 1\} = \{0, 1\}$ ומחלוקת היפותזות של פ' סף, $\{\pm\infty\}$ כאשר $\mathcal{H}_{th} = \{x \mapsto h_\theta : \theta \in \mathbb{R}\} \cup \{\pm\infty\}$ כאשר $L_{D,f}(h_S) = \mathcal{A}(S)$. למעשה אנחנו שוללים כאן כל החלטה עם 0-ים ואו 1-ים ואו 0-ים שוב.

נבחר אלג' למידה \mathcal{A} שיחזיר x_i כאשר $h_{\theta_{alg}}(x) = \max_{y_i=0} x_i$, כלומר קו ההפרדה שאנו יכוו בהינתן דוגמאות כלשהן. אם כל הדוגמאות הן 1 אז $\theta_{alg} = \infty$ ואם כל הדוגמאות הן 0 עם לייבל 0 אז $\theta_{alg} = -\infty$.

טענה יהי $0 < \epsilon, \delta < 1 - \delta$. אם $m \geq \frac{\log \frac{1}{\delta}}{\epsilon}$ אז לכל התפלגות \mathcal{D} על \mathbb{R} ולכל $f_\theta \in \mathcal{H}_{th}$ עם הסת' לפחות δ על S מקרי, העלות $L_{\mathcal{D}, f_\theta}(h_{\theta_{alg}}) \leq \epsilon$.

הוכחה: תהי \mathcal{D} התפלגות על \mathbb{R} ו- $f_\theta \in \mathcal{H}_{th}$ אופן הבחירה שלו. האלג' יטעה רק עבור ערכאים $\theta_{alg} < \theta < x$ בין.

$$\bullet \text{ אם } \epsilon < P_{\mathcal{D}}((-\infty, \theta]) < \infty \text{ סימנו.}$$

אחרת $\epsilon \geq P_{\mathcal{D}}((-\infty, \theta])$. נגיד θ' להיות מספר שמיים $\epsilon = P_{\mathcal{D}}((\theta', \theta))$, כלומר אנחנו מוצאים את הטווח שעבורו ההסת' $\theta' \leq x \leq \theta$ אז העלות היא לכל יותר ϵ (במקרה הגורע הרף נקבע על ידי x והוא θ' ואז העלות היא בדיק ϵ) וההסת' לא לקבלת דוגמה כזו היא $e^{-(1-\epsilon)^m} < \epsilon$. מאינפי 1 מתקיים $e^{-m} < \epsilon$ ולכן ההסת' לשגיאה גדולה מ- ϵ הוא לכל יותר $e^{-\epsilon m}$ וזה קטן מ- δ עבור $m \geq \frac{\log \frac{1}{\delta}}{\epsilon}$.

■

מסקנה \mathcal{H}_{th} היא למידה-PAC.

עכשו אנחנו יכולים להבין את ההגדלה של מחלוקת היפותזות למידה-PAC (חזרו וhabenoid איך הכל מתכנס יחד).

למידה מחלוקת היפותזות סופיות

דוגמה מחלוקת היפותזות שהיא כל הפ' שאפשר כתוב עם לכל יותר 10,000 תווים בפייתון היא סופית - لكن המשפט היסודי במקרה הסופי הוא כן מעניין.

הערה מסתבר שיש לומד פשוט שתמיד מצליח למדוד מחלוקת PAC סופיות עם אותו עיקרונו למידה. העיקרונו הוא ERM. הסיכון האמפירי של S הוא $L_S(h) = \frac{1}{m} |\{i : h(x_i) \neq y_i\}|$.

$$\mathcal{A}_{ERM} : S \mapsto \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$$

אמנם המינימום לא יחיד אבל נזכיר בכל מקרה את אחד המזערים. בגל שהאלג' הזה כל כך מיוחד נסמן $ERM_{\mathcal{H}}$.

משפט (המשפט היסודי של למידה סטטיסטית עבור מחלוקת היפותזות סופית) יהי \mathcal{X} מרחב דוגמאות, $\mathcal{Y} = \{0, 1\}$ עם $|\mathcal{H}| < \infty$ ו- $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. אם $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$ אז לכל \mathcal{D} , f מתקיים $L_{\mathcal{D}, f}(ERM_{\mathcal{H}}(S)) \leq \epsilon$ בהסת' לפחות δ .

הוכחה: בהינתן S , יכולים להיות כמה כלליים שמזערים סיכון אמפירי. נסמן את כולם ב- $ERM_{\mathcal{H}}(S)$. נבחר m שנקבע בעתיד באמצעות ϵ, δ ונבחר את הלומד $\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D}, f}(h) > \epsilon\}$. יהי $S \mapsto h_S \in ERM_{\mathcal{H}}(S)$ (קבוצת היפותזות עם סיכון גורע מ- ϵ) ויהי S מבחן אימון מקרי.

נרצה להוכיח כי מתקיים $\delta \leq P_{\mathcal{D}^m}(\{S : (\exists h \in \text{ERM}_{\mathcal{H}}(S) : h \in \mathcal{H}_B)\})$ כלומר $P_{\mathcal{D}^m}(\{S : (\exists h \in \text{ERM}_{\mathcal{H}}(S) : h \in \mathcal{H}_B)\}) \leq \delta$. כלומר ϵ הוא לכל היותר δ .

מהנחתה הריליאנטיביליות, כל כלל h שנבחר באמצעות ERM הוא בעל סיכון אמפירי 0 (מההגדירה). נסמן את אוסף הדוגמאות המטעה, כלומר $\{S : \text{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B\} \subseteq M = \{S : (\exists h \in \mathcal{H}_B : L_S(h) = 0)\}$. מתקיים $\{S : \text{ERM}_{\mathcal{H}}(S) \in \mathcal{H}_B\} \subseteq M = \{S : L_S(h) = 0\}$. כלומר $M = \{S : L_S(h) = 0\}$ כל החלטה מטעה (ההכרה שאוסף דוגמי האימון שגורמים ללמידה לחזור כל החלטה מטעה מוכל באוסף דוגמי האימון שעבורם קיימים כל החלטה מטעה) אינה שוויה כי יכול להיות שהלומד מחוזיר כל החלטה לא מטעה אף שקיים לו מקבילים כן מטעים).

לכן $\{S : L_{\mathcal{D},f}(\text{ERM}_{\mathcal{H}}(S)) > \epsilon\} \subseteq M$ (זה נכון ולכנן אם נכון $\delta < P_{\mathcal{D}^m}(M) < \epsilon$ נסימן).

$$\text{נכטו} M = \bigcup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\} \text{ לכן}$$

$$P_{\mathcal{D}^m}(M) \stackrel{\text{א.ש. בול}}{\leq} \sum_{h \in \mathcal{H}_B} P_{\mathcal{D}^m}(\{S : L_S(h) = 0\})$$

נשים לב כי $P_{\mathcal{D}^m}(\{S : L_S(h) = 0\})$ היא החסת' לדגם S כך ש- h בדיק נכונה על כל x_i נבחרים p.i.i. ולכנן נוכל להפריד את הדוגמאות. החסת' של h לטעות על x אקראי היא בדיק $L_{\mathcal{D},f}(h)$ (מההגדירה). לכן החסת' ש- h -מטעה צודקת על כל S היא $(1 - L_{\mathcal{D},f}(h))^m$ ולכנן

$$P_{\mathcal{D}^m}(\{S : L_S(h) = 0\}) = (1 - L_{\mathcal{D},f}(h))^m$$

לכל $h \in \mathcal{H}$ ובפרט

$$P_{\mathcal{D}^m}(\{S : L_S(h) = 0\}) < (1 - \epsilon)^m$$

לכל $h \in \mathcal{H}_B$ סה"כ

$$\begin{aligned} P_{\mathcal{D}^m}(\{S : L_{\mathcal{D},f}(\text{ERM}_{\mathcal{H}}(S)) > \epsilon\}) &\leq P_{\mathcal{D}^m}(M) \\ &\leq \sum_{h \in \mathcal{H}_B} (1 - \epsilon)^m \\ &< |\mathcal{H}_B| (1 - \epsilon)^m \\ &\leq |\mathcal{H}| (1 - \epsilon)^m \\ &< |\mathcal{H}| e^{-\epsilon m} \end{aligned}$$

ולכן נקבל שהחסת' בשורה הראשונה היא לכל היותר δ עבור $m \geq \frac{\log \frac{|\mathcal{H}|}{\delta}}{\epsilon}$

הערה הביטוחו וכי טוב שנוכל לקבל הוא $\frac{|\mathcal{H}|}{e^{m\epsilon}}$ והדיק hei גובה הוא $\frac{|\mathcal{H}|}{e^{m\epsilon}}$ (כשקובעים את שני הפרמטרים האחרים).

האם החסם שנתנו הוא אדווק? מתי אפשר לעשות את מה שאמרנו במקרה יותר כללי? אנחנו יודעים שם \mathcal{H} סופית אפשר ללמידה ואמ

$\mathcal{H} = \mathcal{X}$ אי אפשר ללמידה, אז איפה עבר הגבול באמצע?

דוגמה נגידר $.h_{a,b,c}(x) = \begin{cases} 1 & x \in [a, b], x \geq c \\ 0 & \text{אחרת} \end{cases}$ כאשר $\mathcal{H} = \{h_{a,b,c} : a < b < c \in \mathbb{R}\}, \mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0, 1\}$ קלומר פ' ס' כפול. המחלקה הזו יותר עשירה מ- \mathcal{H}_{th} אבל עדין אינטואיטיבית היא למידה-PAC.

מימד VC

עבור $\mathcal{X} \subseteq \mathcal{C}$ בגודל m , נניח כי כל כללי החלטה על C האפשריים נמצאים ב- \mathcal{H} . אז אי אפשר למצוא לומד PAC עם פחות מ- $\frac{m}{2}$ דגימות (ראינו זאת בדוגמה ל-No Free Lunch) כי הטבע יכול להטעות אותנו על כל הדגימות האחרות ב- C וכו' וכו'.

אם לכל m , נוכל למצוא $\mathcal{X} \subseteq C$ כך שכל הצטומים h_C הם ב- \mathcal{H} וגם $|C| > 2m$ או אין לומד PAC ל- \mathcal{H} כל פעם חדש נתן את הטיעון הנ"ל.

למעשה הדבר הזה הוא אס"ם, קלומר, קיימת סדרת קבוצות C_i בגודל שווה לאינסוף שכל הצטומים של C_i הם ב- \mathcal{H} אס"ם \mathcal{H} היא לא-PAC-למידה.

זה אומר שהגודל המקסימלי של C כזה ב- \mathcal{H} הוא כמות קריטית: הוא נותן חסם תחתון על $m_{\mathcal{H}}$ ואם הוא ∞ (קלומר קיימים C כאלה עם גדים באופן שכיח) אז \mathcal{H} לא למידה PAC.

הגדרה תהי $\mathcal{X} \subseteq C = \{x_1, \dots, x_{|C|}\}$ ותהי \mathcal{H}_C הצטומים של \mathcal{H} ל- C , כאשר \mathcal{H} מנתצת את C אם $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\}$. נאמר כי \mathcal{H} מנתצת את C אם קיימים $C \rightarrow \mathcal{H}_C$ קיימים $C \rightarrow \mathcal{H}$ ששווה לה.

הגדרה תהי \mathcal{H} מחלוקת היפוטזות. **מימד ה-VC** של \mathcal{H} מוגדר ע"י \mathcal{H} מנתצת את C .

הערה למעשה אם \mathcal{H} מנתצת את C זה אומר שהוא אף דבר על ליבלים מצומצמים ל- C (קלומר היא כללית על אוסף דגימות מצומצם).

דוגמה מהו $\text{VCdim}(\mathcal{H})_{th}$

האם $C = \{x\}$ מנתצת ע"י \mathcal{H}_{th} ? עבור h_{x-1}, h_{x+1} נקבל שהצטומים ממהה את כל $\mathcal{Y} \rightarrow C$ (כי זה ± 1 וכל אחת מהפ' נותנת ערך אחר בהתאמה), קלומר כן.

נראה כי לכל פ' בגודל 2 אינה מנתצת. יהיו $a < b$. יש 4 פ' אפשרויות מ-{ a, b } (++, --, +-,-+). ההשמה $+$ לא תתקבל (פ' הס' תמיד מתחילה מ-- ומתיישחו עוברת ל-+).

הערה קלומר הוכחנו כאן שקיימות קבוצה בגודל 1 שהיא מנתצת ובנוסף שאין אף קבוצה בגודל יותר גדול שהיא מנתצת, וכך הגענו ל- $\text{VCdim}(\mathcal{H})_{th} = 1$.

אם $\infty = \text{VCdim}(\mathcal{H})$ אז אי אפשר ללמידה כמו שאמרנו לעיל, אחרת, המשפט היסודי נותן לנו שכן.

הערה אם אין קבוצה בגודל d מנתצת אז גם אין קבוצה בגודל 1 $\geq d + 1$ שמננתצת.

דוגמה עבור פ' דו-סימetric, \mathcal{H} , נחשב את $\text{VCdim}(\mathcal{H})$. בروم שיש קבוצה בגודל 1 שמנוטצת. בנוסף הפעם קיימת גם קבוצה בגודל 2 שמנוטצת, לדוגמה $\{1, 2\}$ (פשוט מזיזים את הספים קדימה ואחורה, זה לא מעניין במילוי).

עבור קבוצה בגודל שלוש הקיימים הבאים אפשריים:

1	2	3
+	+	+
+	+	-
+	-	+
+	-	-
-	+	+
-	+	-
-	-	+
-	-	-

הסטודנטית המשקיעה תראה שאנו קיימות פ' סך לכל הקיימים האלה (לא מעניין במילוי).

עבור C בגודל 4 לעומת זאת כבר לא נוכל לספק את כל הדרישות (הסטודנטית המשקיעת תראה זאת).

לכן הוכחנו שאנו מנתצים 3, 2, 1 ולא מצליחים לנתח קבוצות בגודל 4. לכן $\text{VCdim}(\mathcal{H}) = 3$.

לכורה המשפט היסודי עושה טריוויאלייזציה לבעה - פשוט נפעיל את כלל ERM תמיד ונגיע כמעט למינימום הדגימות הנדרשות. הבעה עם זה היא שчисוב ERM יכול להיות קשה מאוד ואז מודלים אחרים עוזרים יותר.

תרגול

נזכיר ביריגסיה לינארית:

1. תרחיש: נתון לנו $\mathcal{Y} \xrightarrow{f} \mathcal{X}$ והנחה ש- f לינארית.

2. מצום מחלוקת היפותזות: בגלל שהכל לינארי, ההיפותזות הן $w^T x \mapsto x^T w$.

3. עקרון למידה: ERM שנמצא אותו עם $\underset{w}{\operatorname{argmin}} RSS(w)$

4. אלגוריתם: $\hat{w} = X^\dagger y$

עד עכשיו לא התעסקנו ברעש, ועכשו נתיחס אליו באמצעות מקסימום נראות ונגיע למסקנה ש- $\hat{w}^{MLE} = X^\dagger y$ אבל עכשו הרבה יותר מושכל כי נדע שהוא מתחשב ברעש.

הערה אם RSS היה פ' עלות אחרת, נגד מצום על נורמה ℓ_1 ולא נורמה ℓ_2 , לא נקבל את אותו האומד עבור \hat{w}^{MLE} .

הערה יש סיבות טובות לעבד עם ℓ_1 ולא ℓ_2 , למשל אם יש נקודות מאוד חרגות לעומת איזשהו קו יחסית לינארי עם רעש, הם הרבה יותר יכולים ל- ℓ_2 מאשר ל- ℓ_1 .

נניח כי $y = Xw + \epsilon_i$ כאשר $\epsilon_i \stackrel{\text{i.i.d}}{\sim} N(0, \sigma^2)$ (olumn ממורכזים, עם אותה שוננות וב"ת אחד בשני) וביצוג מטריציוני. לכן $y \sim N(Xw, \sigma^2 I_m)$. דרוש הוכחה אבל אינטואיטיבית הגיונן).

הערה רעש יכול להויכר משגיאה ב- y או מוגינה לא מדויקת של פיצ'ר.

נסמן $\text{נסמן}(y | Xw, \sigma^2 I_m) = N(y | Xw, \sigma^2 I_m)$ PDF של y כדי לא לכתוב את הצפיפות של גאוסיאן רב ממדים כל פעם. נמצא את הנראות של w ,

$$\begin{aligned}\mathcal{L}(w | y) &= f_w(y) \\ &= N(y | Xw, \sigma^2 I_m) \\ &\stackrel{\text{i.i.d}}{=} \prod_{i=1}^m N(y_i | x_i^T w, \sigma^2) \\ &= \dots \\ &= (2\pi\sigma^2)^{-\frac{m}{2}} \prod_{i=1}^m e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2}\end{aligned}$$

$$\begin{aligned}\hat{w}^{MLE} &= \underset{w}{\operatorname{argmax}} \mathcal{L}(w | y) \\ &= \underset{w}{\operatorname{argmax}} \log \mathcal{L}(w | y) \\ &= \underset{w}{\operatorname{argmax}} \log \left((2\pi\sigma^2)^{-\frac{m}{2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - x_i^T w)^2 \\ &= \underset{w}{\operatorname{argmax}} -RSS(w) \\ &= \underset{w}{\operatorname{argmin}} RSS(w) \\ &= \text{ראינו בתרגול הקודם } X^\dagger y\end{aligned}$$

כלומר הגענו לאותו הדבר אפיו כשהתחשבנו ברעש!

כיצד נוכל ללמידה פולינום כמו $y = x^3 + \frac{1}{2}x^2 - 7x - 3$ באמצעות המודל שיש לנו כרגע? נוכל להגיד שאליה מתייחסת x . למעשה רצוי שעל הפיצ'רים אנחנו עושים ריגרסיה לינארית (אפשר להפעיל עליהם סינוסים ולוגנים ולפניהם מגיעים לריגרסיה וזה לא ישנה).

נניח שיש לנו $x \in \mathbb{R}^d$ וקטור פיצ'רים. תחילה $h : \mathbb{R}^d \rightarrow \mathbb{R}^k$ ו- $h_1, \dots, h_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$. כלומר $h_i(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_k(x) \end{pmatrix}$ והוא איזשהו פרי-פרוססינג של x . נחשפ $w \in \mathbb{R}^k$ כך ש-

במקרה של התאמת פולינומיים, מדובר במקרה פרטי של ההרחבה הנ"ל של ריגרסיה ליניארית. עכשו, יש לנו פיצ'ר מקורי יחיד $x \in \mathbb{R}$ ושלכל $\{h_0, \dots, h_m\}$ מתקיים $\mathcal{H}_{poly}^k = \left\{ x \mapsto \sum_{j=0}^k h_j(x) w_j \right\}$. מחלוקת היפותזות היא $h_j(x) = x^j$ ונשים לב שהעדרון ריגרסיה ליניארית כי מה שחשוב זה שהפ' תהיה ליניארית ב- w . המטריצה X עכשו היא אן-דר-מנדיה.

אם אין רוש ואנחנו מנסים להשתמש בחתامة פולינומיאלית לפולינום, כמשמעותו נקלע בול ונפסיק שם (כל שאר המקדים הגבוהים יותר יתאפסו). אם יש רוש, עד שלא נגיע לדרגה m (שאז אפשר לעבור דרך כל נקודה ונקודה), ה- RSS ישתפר כל פעם עוד ועוד.

הערכה מתקיים $\dots \subseteq \mathcal{H}_{poly}^0 \subseteq \mathcal{H}_{poly}^1 \subseteq \dots$

נסתכל עכשו על הטריידוף של ההטיה-שונות. $\text{Bias}(\hat{\theta}) = E_{\text{Data}}[\hat{\theta}] - \theta$

$$\begin{aligned} E_y[\hat{w}] &= E_y[(X^T X)^{-1} X^T y] \\ &= E[(X^T X)^{-1} X (X^T w + \epsilon)] \\ &= \frac{(X^T X)^{-1} X^T X w}{=1} + (X^T X)^{-1} X^T \frac{E[\epsilon]}{=0} \\ &= w \end{aligned}$$

כלומר זה אומד לא מוטה.

Mean Squared) MSE $E[(\hat{y} - y)^2] = \dots = \text{var}_y(\hat{y}) + \text{Bias}_y^2(\hat{y})$ אבל הביאס הוא 0 כמו שעכשו הראננו ולכן (Error) מושפע רק מהשונות של \hat{y} על y במקרה הזה.

עם זאת, ככל מתקיים

$$\text{MSE} = \text{Variance} + \text{Bias}^2$$

ואנחנו רוצחים לצמצם את MSE ולבסוף כמה שיותר לצמצם גם את ההטיה והשונות, אבל ככל שנתאים יותר מדויק, ההטיה תרד אבל השונות תעליה. לדוגמה אם נתאים פולינום ממעלה גבוהה מדי לתופעה שהיא פולינום מדרגה נמוכה יחסית עם איזשהו רוש, הפולינום יעקוב אחרי הרוש הרבה יותר מדי וישכח את "התמונה הכללית".

שבוע VII | מסגרת PAC אגנוסטית

הרצאה

נזכר שיש לנו לומד שהוא אלג' \mathcal{A} . שמקבל $\{(x_i, y_i)\}$ כאשר $x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$ והוא הכלל האמתי שלא ידוע. \mathcal{A} פולט כל החלטה $\mathcal{Y} \rightarrow \mathcal{X}$ שמצויר על $h(x) \neq f(x)$ ($L_{\mathcal{D},f}(h) = \Pr_{x \sim \mathcal{D}}(h(x) \neq f(x))$) והוא אקראי.

לומד PAC הוא לומד שמצויר כל החלטה שמצויר בהסת' $\delta - 1$ להזיר עלות חסומה ע"י ϵ (ביחס ל- S). אמרנו ש- \mathcal{H} היא מחלוקת היפותזות

למייה PAC אם היא לא מורכבת מדי, ושיש לה סיבוכיות מדגם כלשהי. דיברנו על מימד VC והראנו שם יש גודל מקסימלי של קבוצה מנותצת ע"י \mathcal{H} אז \mathcal{H} למייה PAC ואחרת לא. למדנו איך לחשב מימד VC.

משפט (המשפט היסודי של הלמידה הסטטיסטי, שוב) תהיו \mathcal{H} מחלקת היפותזות של מסויימים ביןאריים עם ∞ . אז $d = \text{VCdim}(\mathcal{H}) \leq \infty$. במקורה זה, קיימים קבועים c_1, c_2 כך ש-

$$c_1 \frac{d + \log \frac{1}{\delta}}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq c_2 \frac{d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}}{\epsilon}$$

ועלISON-H-ERM מושג את החסם העליון של הסיבוכיות זו.

נrichib את תורה PAC-L-PAC אגנוסטי, כולם נקל את הדרישות.

בעיות עם PAC קלאסי

- אנחנו לא מתייחסים לרעש בלייבלים.
- הנחת הריאלייזביליות היא לא ריאלית.
- העולות שלנו מוגבלות לעלות ℓ_0 ולא עלות גנרייה.

רעש במסגרת האגנוסטי

כדי לאפשר רעש מקרי, נניח שההתפלגות שלנו היא לא על \mathcal{X} , אלא התפלגות על $\mathcal{U} \times \mathcal{X}$ (התפלגות משותפת). אם ההתפלגות השיווריות ב"ת לא השנו כלום, אבל אם הן לא, יוכל ללמידה דברים.

ונכל לכתוב

$$P(X = x, Y = y) = P(X = x) P(Y = y | X = x)$$

כלומר קודם כשוברים y, x , אנחנו בוחרים x כלשהו ואז דוגמים ברנולי את y על בסיסו.

למעשה יוכל לרשום $P(x) = p$ $P(Y = +1 | X = x) = p$ כאשר $\mathcal{X} \rightarrow [0, 1]$ כאמור כמה רעש יש לנו. אם $p = 0, 1$ אז אנחנו באלאג' דטרמיניסטי כרגע. אם לא, (x) מיצג את הרעש בתווית של הדגימה - אם קרוב לחצי אז המון רוש, אם קרוב ל-0 או 1 אז יש יחסית מעט רוש.

לחלוופין, יוכל לכתוב

$$P(X = x, Y = y) = P(Y = y) P(X = x | Y = y)$$

כלומר קודם מטילים מטבע לראות מה התווית, ורק אז בודקים מה הסיכוי שהתוויות נלקחה מדגימה כזו או אחרת.

הכללת העלות

בגלל שאנו עכשו עם התפלגות \mathcal{D} על $\mathcal{Y} \times \mathcal{X}$, אין יותר f שהיא נcona באמת, כי אנחנו לא דטרמיניסטיים. הדבר כי קרוב לעלות 1 – 0 היא $P(Y = y | X = x)$

הגדרה פ' **עלות כללית** היא $\int_{\mathcal{H}} \ell(h(x), y) d\mathcal{D}(x, y)$ כמה ההערכה שלנו טוענה (כמה שיוור גדור כמו שיוור גורע).

$$L_{\mathcal{D}}(h) = E_{\mathcal{D}}(\ell_0(h, (x, y))) = \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$$

הגדרה עברו $L_{\mathcal{D}}(h) = E_{x \sim \mathcal{D}}(\ell(h, z))$, נגדיר את העלות המוכללת ע"י $\ell : \mathcal{H} \times Z \rightarrow [0, \infty]$

המקרה הלא ריאלייזабילי וההגדרה האגנוסטית

בעבר העלות של כלל החלטה המינימלי ב- \mathcal{H} היה 0 כי הנחנו ריאלייזבילים, אבל עכשו לא נוכל להניח את זה, שכן השתמש במינימום גנרי.

הגדרה ϵ ומחלוקת היפותזות \mathcal{H} . נאמר כי $h \in \mathcal{H}$ הוא Correct אם מתקיים

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

הגדרה נגדיר את הUDENT**הבייזני האופטימלי** על התפלגות \mathcal{D} של $\mathcal{Y} \times \mathcal{X}$ על ידי. הבעה כאן היא שאנו לא יודעים מה \mathcal{D} ולכן לא נוכל לחשב אותו. $f_{\mathcal{D}}$ הוא הלומד עם העלות המוכללת הקטנה ביותר. הסטודנטית המשקיעה תוכיה זאת.

הסטודנטית המשקיעה תגידר מהו \mathcal{A} , לומד PAC-אגנוסטי עם ביטחון δ ודיוק ϵ . היא תוכיה גם שלומד PAC אגנוסטי עם עלות 1 – 0 הוא בפרט לומד PAC קלאסי.

הגדרה תהי \mathcal{H} מחלוקת היפותזות. נאמר כי היא למידה-PAC-אגנוסטי ביחס לעלות $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \rightarrow [0, \infty)$ אם קיימת פ' $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$ שמקיימים את התכונה הבא:

$$\text{לכל } 1, \text{ התפלגות } \mathcal{D} \text{ על } \mathcal{Y} \times \mathcal{X} \text{ מתקיים } m \geq \tilde{m}_{\mathcal{H}}(\epsilon, \delta) \text{ ו- } \mathcal{A}(\mathcal{S}) \text{ הוא אוסף דגימות i.i.d מ-} \mathcal{D} \text{ עם עלות } 1 - \delta.$$

$$P_{\mathcal{D}^m} \left(\left\{ S_m : L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon \right\} \right) \geq 1 - \delta$$

כאשר $h_S = \mathcal{A}(S)$ והוא אוסף דגימות i.i.d מ- \mathcal{D} .

הסטודנטית המשקיעה תוכיה שאם \mathcal{H} מחלוקת היפותזות למידה PAC-אגנוסטי עם עלות 1 – 0 אז היא למידה PAC.

הערה למעשה PAC-למידה שקול ל-PAC-אגנוטי למידה עם עלות ℓ_0 אבל לא נוכחת את זה.

הגדירה ה**סיכום האמפירי** ביחס ל- $\mathcal{Y} \rightarrow \mathcal{X}$: $\mathbb{P}, h : \mathcal{F}$, עלות ℓ ודוגמאות $S = \{(x_i, y_i)\}_{i=1}^m$ מוגדר ע"י, "אבל שמקיים $L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, (x_i, y_i))$

הגדירה לומד **ERM** במסגרת PAC האgnוטי הוא אלג' לומד שמקיים $\mathcal{A}_{\text{ERM}} : S \mapsto \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$

משפט (חוק החלש של המספרים הגדולים) עבור X_i דגימות i.i.d או $\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m X_i = E[X_i]$ כשהתכנסות היא בהסת' , כלומר, לכל $\delta > 0$ מתקיים

$$\lim_{m \rightarrow \infty} P \left(\left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| > \delta \right) = 0$$

. $P \left(\left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| > \delta \right) < \epsilon$ קיים $m_0 \in \mathbb{N}$ כך שכל $m > m_0$ מתקיים $\left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| < \delta$.

הערה החוק החלש נותן לנו קשר בין ממוצע אמפירי לעלות המוכלلت (עבור מספיק דגימות).

משפט (המשפט הייסודי של הלמידה הסטטיסטיית האgnוטי) תהיו \mathcal{H} מחלקת היפותזות של מסווגים ביןאריים עם מימד $\text{VCdim } \mathcal{H} \leq d$. אז \mathcal{H} היא למידה PAC-אגנוטי אם $\exists \epsilon < \delta$ ובקשה כזו,

$$c_1 \frac{d + \log \frac{1}{\delta}}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq c_2 \frac{d + \log \frac{1}{\delta}}{\epsilon^2}$$

וכן ERM מושג את החסם העליון של סיבוכיות המדגם.

את הצד הראשון כבר הוכיחו (אם בשילילה $\infty = d$ יש סדרת קבועות מנותצות لكن לא ניתן ללמידה).

עתה נניח $\text{VCdim } (\mathcal{H}) < \infty$ ונוכיח ש- \mathcal{H} למידה PAC-אגנוטי בכך כל יתר ההרצתה. אנחנו לא הולכים להוכיח הכל פורמלית אבל ניגע בGESOT בתוכן ההוכחה (אנלוגיה בעיינית).

חשיבות התוכנות במ"ש

מתקיים $(h) \in \mathcal{L}_S$ מהחוק החלש. לכן, $\forall \epsilon, \delta > 0$, קיים $m_0 \in \mathbb{N}$ כך שכל $m > m_0$ מתקיים $\mathbb{P}(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) < \delta$

$$\mathbb{P}(|L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) < 1 - \delta$$

זה לא מסיים את ההוכחה כי זה מוכיח עבור h ספציפי, ולא עבור h שימושה בכל שיש לנו יותר נתונים (שמתබול מהתבנית \mathcal{A} על S).

למעשה הPUR שנוצר כאן הוא ההבדל בין התוכנות של סדרת פ' לבין התוכנות במ"ש של סדרת פ' (h כלשי בטוח מתוכנת, אבל מינימום על h -ים שמתוכנים עובד אחרת).

סיכום ביניים נרצה להראות ש- $(S) \in \text{ERM}_{\mathcal{H}}(S)$ עובד טוב במקרה הכללי כי הוא מזעך סיכון אמפירי, כלומר, נוכיח כי

$$P_{\mathcal{D}^m}(\{S \in (\mathcal{X} \times \mathcal{Y})^m : |L_{\mathcal{D}}(h_S) - L_S(h_S)| < \epsilon\}) > 1 - \delta$$

ומשם אם $L_S(h_S)$ קרוב למינימום ואז $L_{\mathcal{D}}(h_S)$ קרוב למינימום (בהתאם' הכלל כמובן) ואז נקיים את תנאי ה-PAC האגונסטי ונסיים.

הגדירה תהיה סדרת פ' על \mathbb{R} . נאמר כי f_n מתקנת במ"ש $-f$ אם $\epsilon > 0$ קיים $N \in \mathbb{N}$ כך שלכל $x, m_0 \geq n$ מתקיים $|f_n(x) - f(x)| < \epsilon$.

הערה כאמור זה אומר להוכיח התכונות במ"ש של $L_S(h)$ כאשר כרגע יש לנו לכל h, D, m_0 כלשהו שאחריו אנחנו לומדים, ואילו נרצה m_0 שאחריו לכל h, D אנחנו לומדים (סדר הנקודות הוא ההבדל בין התכונות נקודתית במ"ש), כלומר שאינו תלוי $-h$.

הגדירה נאמר כי S הוא ϵ -מייצג עבור ℓ , D, H, ℓ אם לכל $h \in \mathcal{H}$ לכל $h \in \mathcal{D}$ $|L_S(h) - L_{\mathcal{D}}(h)| < \epsilon$, כלומר $L_S(h) - L_{\mathcal{D}}(h) < 2\epsilon$ ליותר (ERM), יהיה קרוב ב- 2ϵ ליותר (ERM).

טענה יהיו S מוגן אימון $\frac{\epsilon}{2}$ -מייצג עבור ℓ, D, H, ℓ . יהיו $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$ כלומר החלטה לפי עקרון (Uniform Convergence) או $h \in \mathcal{H}$ כל החלטה על כל \mathcal{D} .

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

כלומר, אם S ϵ -מייצג אז ERM עבר הכללה בהצלחה (הוא לומד היטב על כל \mathcal{D}).

סיכום ביניים מעתה נועל כדי לקבל בהסתמך גבורה S מייצג מספיק.

הגדירה נאמר כי \mathcal{H} יש את תכונת ההתכנסות במ"ש (Uniform Convergence) אם קיים N כך שלכל $1 < \delta, \epsilon$ $m_{\mathcal{H}}^{UC}((0, 1)^2) < \epsilon$ וכל ההפולגות \mathcal{D} על $\mathcal{X} \times \mathcal{Y}$ מתקיים

$$P_{\mathcal{D}^m}(\{S \in (\mathcal{X} \times \mathcal{Y})^m : S \text{ מימייצג}\}) > 1 - \delta$$

הערה למעשה נפטרנו מהתלות ב- $-h$, וגם m בא לפני \mathcal{D} עכשו (כמו שרצינו בשביל הבמ"ש).

טענה אם \mathcal{H} היא בעלת התכונות במ"ש אז היא למידה PAC-אגונסטי עם סיבוכיות מוגן $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\epsilon}{2}, \delta)$.

הוכחה: אם הסיכוי לקבל מוגן $\frac{\epsilon}{2}$ -מייצג הוא גדול, לעומת עיקרונו ERM תמיד מוצא כל החלטה עם סיכון אמפירי לכל היותר ϵ יותר גדול מהעלות המוכללת כלומר הוא מקיים את ההגדרה של PAC-למידה. הטענה מושקיעה תרכיב את כל ההגדרות יחד פורמלית כדי למשם את הוכחה זו. ■

סיכום ביניים מספיק שנראה כי אם $\text{VCdim}(\mathcal{H}) < \infty$ אז \mathcal{H} היא בעלת תכונת UC. כלומר, עבור m מספיק גדול שלא קשור ל- \mathcal{D} , מוגן ϵ -מייצג עם הסת' $\delta = 1 - \text{כל } \mathcal{D}$.

פער העלות הגדול ביותר

הגדולה תהי \mathcal{D} התפלגות על $\mathcal{Y} \times \mathcal{X}$, \mathcal{H} מחלקת היפותזות ו- $m \in \mathbb{N}$. נגיד $F_m^{\mathcal{D}} : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathbb{R}$ מחלקת היפותזות ו- $m \in \mathbb{N}$.

$$F_m^{\mathcal{D}}(S) = \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$$

כלומר כלל החלטה עם פער הכי גדול בין סיכוי אמפירי עלות מוכלلت.

הערה $\{F_m^{\mathcal{D}}\}_{m \in \mathbb{N}}$ היא סדרת מ"מ על $(\mathcal{X} \times \mathcal{Y})^m$.

סיכום בינוניים מספיק שנווכיה שלכל $1 < \epsilon, \delta < 0$ קיימת $N \in \mathbb{N}$ כך שלכל התפלגות \mathcal{D} ו- $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ נקייב S לא (ϵ -)מייצג הוא לכל היותר δ , שזו ההגדרה של התכנסות במ"ש.

חסימת פער העלות הגדול ביותר במקרה הסופי

טענה תהי \mathcal{H} מחלקת היפותזות סופית ויהי $0 < \epsilon, \delta < 1$. אזי קיים $m_0 \geq m_0(\epsilon, \delta)$ ולכל \mathcal{D} , $m \geq m_0$ כך שלכל

הערה לא הוכחנו פורמלית אבל קל להיווכח שאם \mathcal{H} סופית אז $\text{VCdim}(\mathcal{H})$ סופי.

הוכחה: מתקיימים

$$\begin{aligned} P_{\mathcal{D}^m}(\{S : F_m^{\mathcal{D}}(S) > \epsilon\}) &= P_{\mathcal{D}^m}(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\ &\leq \sum_{h \in \mathcal{H}} P_{\mathcal{D}^m}(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\ &\leq |\mathcal{H}| \max_{h \in \mathcal{H}} P_{\mathcal{D}^m}(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \end{aligned}$$

כלומר כל שנותר הוא לחסום את $\{P_{\mathcal{D}^m}(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\})\}_{h \in \mathcal{H}}$

משפט (א"ש הוודיניג) תהי $\{\theta_m\}_{m \in \mathbb{N}}$ סדרת מ"מ תמיד איז לכל $0 < \epsilon < \mu$ מתקיימים

$$P\left(\left|\frac{1}{m} \sum_{i=1}^m \theta_i - \mu\right| > \epsilon\right) \leq 2e^{-2\frac{m\epsilon^2}{(b-a)^2}}$$

כיצד השתמש בו כאן? נגיד $L_S(h) = \frac{1}{m} \sum_{i=1}^m \theta_i$ ו- $L_{\mathcal{D}}(h) = E_{\mathcal{D}}[\theta_i]$ מההגדרה. לכן, מא"ש

הוודיניג מתקיימים

$$P_{\mathcal{D}^m}(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2e^{-2m\epsilon^2}$$

לכן בזרה לא"ש מההתלהה, ההסת' חסומה ע"י, ועבור $m \geq \frac{\log \frac{2|\mathcal{H}|}{\delta}}{2\epsilon^2} 2e^{-2m\epsilon^2}$ מתקיים δ

הערה הוכחה זו לא עובדת כי אי אפשר להשתמש בחסם האיחוד במקרה האינסופי.

גדרה פולינומיאלית של מחלוקת היפותזות

נזכיר בסימן $\text{VCdim}(\mathcal{H})$ לצטום כל \mathcal{H} - C . אם (\mathcal{H}) מונצת את C אז $|\mathcal{H}_C| = 2^{|C|}$ יכול להיות ש- \mathcal{H} מונצת. אם $|\mathcal{H}| \leq \text{VCdim}(\mathcal{H})$ לא יכול להיות ש- \mathcal{H} מונצת.

מסתבר שאם אנחנו יודעים כמה מהר $|\mathcal{H}_C|$ גדלה (לחסום את הגדרה איכשהו), נוכל להוכיח את המקרה הכללי.

הגדרה תהי \mathcal{H} מחלוקת היפותזות, נגדיר $\tau_{\mathcal{H}}(m) = \max \{|\mathcal{H}_C| : C \subseteq \mathcal{X}, |C| = m\}$.

הערה קל לראות ש- $\tau_{\mathcal{H}}$ גדל עם m , אבל כמה גדול? אם ∞ או קיימת סדרת מספרים כך ש-

הגדירה תהי מחלוקת היפותזות $\mathcal{U} \subseteq \mathcal{H}$. נאמר כי \mathcal{H}_C גדלה פולינומיאלית-ב- $|C|$ אם קיימים $N, \beta, b > 0, m_0 \in \mathbb{N}$ כך שלכל $m > m_0$ נובעת $\tau_{\mathcal{H}}(m) \leq b \cdot m^\beta$.

סיכום ביניים כדי להוכיח את המקרה הכללי, נחלק לשני חלקים.

1. אם \mathcal{H}_C גדלה פולינומיאלית-ב- $|C|$ אז ל- \mathcal{H} יש UC ולכנן למידה PAC אגוסטי עם עיקרונו ERM.

2. אם ∞ או \mathcal{H}_C גדלה פולינומיאלית-ב- $|C|$.

טענה (חלק הראשון) אם $|\mathcal{H}_C|$ גדלה פולינומיאלית-ב- $|C|$ אז \mathcal{H} היא עם UC.

הוכחה: נזכיר כי כדי להוכיח UC אנחנו מוכחים כי לכל δ, ϵ, α מתקיים $P_{\mathcal{D}}^m(F_m^{\mathcal{D}}(S) > \epsilon) < \delta$ מספיק גדול. משתמש בא"ש מרכיב,

$$\text{כלומר שלכל } m \text{ אי שלילי (כמעט תמיד) מתקיים } P(X > \alpha) \leq \frac{E[X]}{\alpha}$$

הו הוא אי שלילי מבון. נמצא סדרת מספרים α_m ייה תלוי ב- \mathcal{H} אבל לא ב- \mathcal{D} , כלומר החסם הוא במ"ש על \mathcal{D} . ממש, מתקיים

$$P_{\mathcal{D}^m}\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| > \epsilon\right) = P_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S) > \epsilon) \leq \frac{E_{\mathcal{D}^m}(F_m(S))}{\epsilon} \stackrel{\text{מרכיב}}{\leq} \frac{\alpha_m}{\epsilon}$$

כלומר, ההסת' לקבלת מרחב מדגם באורך m שהוא ϵ -מייצג היה לפחות $\frac{\alpha_m}{\epsilon} - 1$. כלומר, השגנו התוכנות במ"ש גם על $\mathcal{H} \in \mathcal{H}$ (אנו לוקחים סופריום) וגם על \mathcal{D} (אנו חסומים תוחלת).

זה לא מספיק כי יכול להיות ש- α_m ממשיך גדול, וזה החסם על ההסת' לא שווה כלום. אם נוכל למצוא סדרה כזו שמתאפסת (שואפת ל-0), אז לכל δ, ϵ , נקבע m_0 כך שלכל $m > m_0$ $\tau_{\mathcal{H}}(m) < \delta, \epsilon$, כלומר ל- \mathcal{H} יש UC.

מצא סדרה כזו.

лемה מרובית (שלא נוכחה) $E_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S)) = \mathcal{O}\left(\sqrt{\frac{\log(\tau_{\mathcal{H}}(2m))}{2m}}\right) + o(m)$

מסקנה גודל פולינומיאלית ב- $|C|$ ולכן קיימים m_0 כך שלכל $m > m_0$ ולבן $\tau_m(\mathcal{H}) \leq b \cdot m^\beta$, כלומר $\tau_m(\mathcal{H})$ מוגבל ב- $b \cdot m^\beta$.

$$E_{\mathcal{D}^m}(F_m^{\mathcal{D}}(S)) \leq \mathcal{O}\left(\sqrt{\frac{\beta \log 2m}{2m}}\right) + o(m) \xrightarrow[m \rightarrow \infty]{} 0$$

כלומר זו סדרת המספרים שלנו.

טענה (הוכחה בתרגיל) אם $m > \text{VCdim}(\mathcal{H}) = d$ אז פולינומיאלית ממש.

מסקנה אם $\text{VCdim}(\mathcal{H}) < \infty$ אז $\tau_{\mathcal{H}}(m) \geq \text{VCdim}(\mathcal{H})$.

■

סה"כ קיבלנו שאם $\text{VCdim}(\mathcal{H}) < \infty$ אז פולינומיאלי ב- m ולכן \mathcal{H} הוא עם UC ולבן \mathcal{H} היא למידה-PAC-AGONSTI.

תרגום

אנחנו עוסקים בבעיות סיווג לינארית $\mathcal{Y} = \{\pm 1\}$. כלומר מחלוקת הhipotheses שלנו היא $\mathcal{X} = \mathbb{R}^d$. מיפוי f מ \mathcal{X} ל \mathcal{Y} מוגדר על ידי $y = \text{sign}(\langle x | w \rangle + b)$.

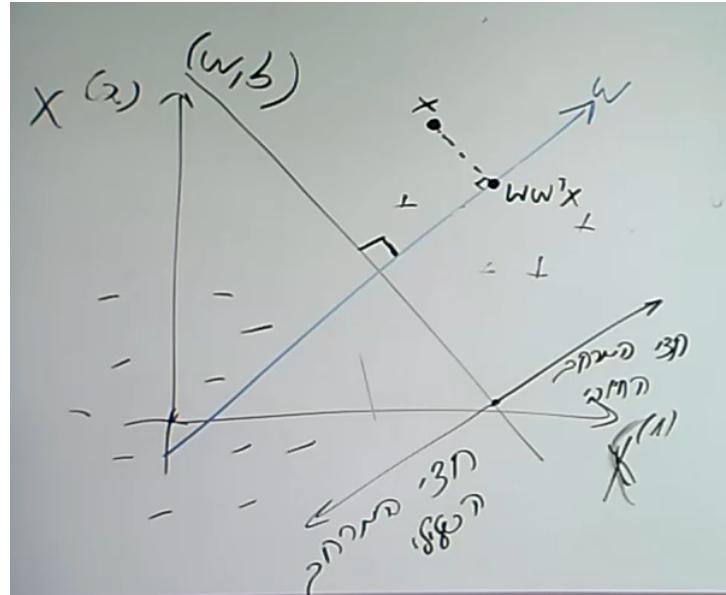
$$\mathcal{H}_{lin} = \{x \mapsto \text{sign}(\langle x | w \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

הגדרה יהיו $w, b \in \mathbb{R}^d$. העל מישור (w, b) הוא $\{x \in \mathbb{R}^d : \langle w | x \rangle + b = 0\}$.

הערה הנקודות שיתנו ערך חיובי עברו החישוב הנ"ל יקרו חצי המרחב החיובי ובהתאם אלו עם ערך שלילי יקרו חצי המרחב השילי.

הערה w הוא וקטור הניצב לקו של העל-מישור. x הוא הנטלה האורתוגונלית של x על w ואפשר להסתכל על זה בכמה דרכים: w היא מטריצת הנטלה על w מדרגה 1; x היא הקוורדינטה לפי הבסיס $\{w\}$ בתוך המרחב הנפרש על w .

לכן אם נטיל נקודה במרחב על w נוכל לראות באיזה צד הוא באמצעות הפרדה לפי ערך על ישר אחד (ערך $x^T w$ על הישר שהוא בכיוון w).



b. מאפשר לנו לחסיט את ה-0 של הישר הנ"ל.

נעסוק ברגע במקרה הרלייזאכטיל ובהמשך נכליל.

נעבוד עם עיקרונו הלמידה ERM, עבור דוגמה אחת נגדיר

$$\ell_{0-1}(h, (x_i, y)) = \mathbb{1}[y \neq \text{sign}(x^T w + b)] = \mathbb{1}[y(\langle x_i | w \rangle + b) < 0]$$

$$. L_S(h) = \sum \ell_{0-1}(h, (x_i, y_i))$$

$$\begin{aligned} \text{כולם ליותר עלות על על-מישורים שפרידים באופן מושלם את הדוגמאות.} \\ \text{בעיית האופטימיזציה שלנו היא} \\ \underset{h_{w,b} \in \mathcal{H}_{lin}}{\operatorname{argmin}} L_S(h) \\ \text{s.t., } y_i (\langle x_i | w \rangle + b) \geq 0 \end{aligned}$$

ב מקרה שיש כזה מיותר, הערות היא 0 מהגדה. כלומר אנחנו מזעררים 0 argmin פשוט. אלג' אחד שפותר בעיה כזו הוא פרספטורן.

Algorithm 1 Batch-Perceptron

```
procedure PERCEPTRON( $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ )
     $\mathbf{w}^{(1)} \leftarrow 0$ 
    for  $t = 1, 2, \dots$  do
        if  $\exists i$  s.t.  $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$  then
             $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$ 
        else
            return  $\mathbf{w}^{(t)}$ 
        end if
    end for
end procedure
```

האלג' מआחל פתרון כלשהו (לא בהכרח חוקי) ובאופן איטרטיבי מזיז את הפתרון כך שיהיה יותר טוב. אם המקרה הוא לא ריאלייזабילי האלג' יכול לזרץ עד אין סוף.

למה האלג' משפר את עצמו כל איטרציה? נניח ש- w^t טועה עבור y_i لكن

$$y_i \langle w^{t+1} | x_i \rangle = y_i \langle w^t + y_i x_i | x_i \rangle = y_i \langle w^t | x_i \rangle + \|x_i\|^2$$

כלומר שאנו "יוטר צודקים" על x עכשו. לא בהכרח שהאיטרציה זו תתקן את הטעות אלא עוד כמה איטרציות זה יתקן את זה.

בעיות בפרשפטרון

1. הפתרון שלו הוא לא יחיד - בהינתן סדר שונה של דוגמאות הוא היה מחזיר אלג' (חוקי) אחר.
2. פרשפטרון לא מתרחק כמו שצריך מנתונים אלא רק זו קצת כל פעם כך שהוא יכול להיכ说得 לקבוצה כלשהי ולתת לנו טעות על ערך מאד קרוב מצד השני העל-מיישור.
3. הוא מניח ריאלייזביליות.

מסקנה בטור לומד, פרשפטרון הוא אומד עם שונות מאוד גבוהה.

כדי לפטור את הבעיה האלה, משתמשים ב-SVM.

Hard-SVM

הגדרה יהיו (w, b) על-מיישור וקטור $x \in \mathbb{R}^d$, המרחק של x מ- (w, b) הוא $\|x - v\|$.

הגדירה ייחי (w, b) על מישור ואוסף דגימות S , השול של (w, b) (m -S) הוא

נרצה על מישור שמקסם שול, כלומר מתרחק מהנקודות בהפרדה כמה שיותר. בעצם בעיית האופטימיזציה שלנו היא

$$\begin{aligned} & \underset{h_{w,b} \in \mathcal{H}_{lin}}{\operatorname{argmax}} \text{margin}(h, S) \\ & s.t. y_i (\langle w | x_i \rangle + b) \geq 0, \forall i \end{aligned}$$

טענה יהיה על מישור (w, b) עם $\|w\| = 1$ $\text{margin}(h, S) = \langle w | w \rangle + b$.

הוכחה: נסמן $P = ww^T$. המטריצה $I - P$ היא מטריצה הטלה על w^\perp . נניח בה'כ $b = 0$. נוכיח כי היא מטריצה הטלה אורתוגונלית ווגם $\text{Im}(I - P) = w^\perp$

מתקיים $(I - P)^2 = I - 2P + P^2 = 1 - P + P = 1 - P$ ומכיוון $w \in \mathbb{R}^d$ ו- w^\perp מתקיים $(I - P)u \perp w$.

נוכיח כי $w \in \text{Im}(I - P)$ או נקבע $u \in \text{Im}(I - P)$ כך ש- u מימד התמונה של w הוא $d - 1$. נוכיח כי $(I - P)u \perp w$. נוכיח כי $(I - P)u \perp w$ ומכיוון $(I - P)u \perp w$ ו- $w \in \text{Im}(I - P)$ נקבע $u \in \text{Im}(I - P)$ כך ש- u מימד התמונה של w הוא $d - 1$.

$$\begin{aligned} \langle (I - P)u | w \rangle &= \langle u - Pu | w \rangle \\ &= \langle u | w \rangle - \langle u | Pw \rangle \\ &= \langle u | w \rangle - \left\langle u | w \frac{ww^Tw}{\|w\|^2} \right\rangle \\ &= 0 \end{aligned}$$

$$\begin{aligned} \|x - (I - P)x\| &= \|x - x + Px\| \\ &= \|ww^Tx\| \\ &= \|w\| |\langle w | x \rangle| \\ &= |\langle w | x \rangle| \end{aligned}$$

■

לכן נוכל לנתח מחדש את בעיית האופטימיזציה בתוור

$$\begin{aligned} & \underset{(w,b): \|w\|=1}{\operatorname{argmax}} \min_i |\langle x_i | w \rangle + b| \\ & s.t. y_i (\langle w | x_i \rangle + b) \geq 0, \forall i \end{aligned}$$

אבל אם $0 \geq \langle w | x_i \rangle + b$ אז אפשר לכתוב פשוט

$$\begin{aligned} \operatorname{argmax}_{(w,b):\|w\|=1} \min_i y_i (\langle w | x_i \rangle + b) \\ s.t. y_i (\langle w | x_i \rangle + b) \geq 0, \forall i \end{aligned}$$

ולמקרה הכל פתרון לא חוקי יתנו ערכים שליליים ולכון הוא מוגלם בתחום המיקסום ולכן אפשר להיפטר מהתנאי למיטה, כלומר אנחנו נשארים

$$\begin{aligned} \operatorname{argmax}_{(w,b):\|w\|=1} \min_i y_i (\langle w | x_i \rangle + b) \\ \text{עם} \\ \text{עכשו ננסח בעיה חדשה,} \end{aligned}$$

$$\operatorname{argmin}_{(w,b):y_i(\langle x_i | w \rangle + b) \geq 1, \forall i} \|w\|^2$$

טענה עבור $(\hat{w}, \hat{b}) = \frac{b^*}{\|w^*\|}$ פתרון אופטימלי לעוביה הנ"ל מתקיים כי הפתרון (\hat{w}, \hat{b}) הוא פתרון אופטימלי לעוביה שלפניה עברו.

$$\hat{w} = \frac{w^*}{\|w^*\|}$$

הערה למשמעות הביעות האלה הם בעיות דו-אלגבריות.

קשה לנו לפתרון בעיות בסגנון השול הקליני, ולכן פתרה של הבעיה הנ"ל הרבה יותר נוחה כי לא רק שזו בעית אופטימיזציה קמורה שאנו יכולים לפתור אלא היא גם בעיה ריבועית.

Soft-SVM

במקרה הלא ריליאזובי, נצורך לאפשר קצת לעובור את הצד השני. כלומר, הבעיה שלנו עתה היא

$$\begin{aligned} \operatorname{argmin}_{w,b,\{\xi_i\}} \|w\|^2 \\ s.t. \quad \begin{cases} y_i (\langle x_i | w \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \wedge \frac{1}{m} \sum_{i=1}^m \xi_i \leq C \end{cases} \end{aligned}$$

כאשר אנחנו קובעים את C , שמחלית כמה אפשר לסתות לכיוון הלא נכון של העל מישור.

C זהה לא נוח ולכן ננסח מחדש בתו

$$\begin{aligned} \operatorname{argmin}_{w,b,\{\xi_i\}} \lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \\ s.t. y_i (\langle x_i | w \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned}$$

עבור ג'רמן רגולרייזציה כלשהו שאנו קובעים. ג'רמן את היחס בין החישבות של $\|w\|^2$ לבין השטייה, אם הוא מאוד גדול זה

אומר שלא מספיק לנו לסתות וחשוב לנו מאוד השול והוא מאוד קטן אז זה אומר שהוא חשוב לנו מאוד לא לסתות ולא כל כך השול.

ונכל שוב לנשח את הבעה מחדש ולקבל ש-Soft-SVM שcool ל-

$$\underset{w,b}{\operatorname{argmin}} \left(\lambda \|w\|^2 + \frac{1}{m} \sum \ell^{hinge} (y_i \langle x_i | w \rangle) \right)$$

$$\text{כאשר } \ell^{hinge} (a) = \max \{0, 1 - a\}$$

למעשה אם $\lambda = 0$ נקבל את בעיית ה-H-SVM שוב.

שבוע VII | אלגוריתמי סיווג הסתברותיים

תרגול

בריגרסיה ליניארית היה לנו $(\phi_w(x_i), \sigma^2)$ כאשר $w \sim N(\phi_w(x_i), \sigma^2)$ (כרגע אין משמעות ל"בහינתן x_i " אבל בהמשך יהיה).

בבעיות סיווג יש לנו $y_i \sim Ber(\phi_w(x_i))$ כאשר $w \sim N(\phi_w(x_i), \sigma^2)$. לאחר מכן, נגדיר איזשהו סף שאם אנחנו מעליו נסוווג 1 ואם מתחתנו נסוווג 0.

דוגמא אם נגדיר רף 0.7, אם $\phi_w(x_i) = 0.45$ נקבע שזה יהיה 0.

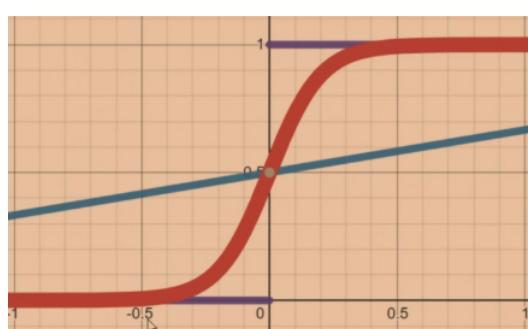
אנו מנהים קשר ליניארי בין x להתפלגות p_i של y_i .

הפ' הכיטוריואלית היא $\phi_w(x_i) = \mathbb{1}[x_i^T w > 0]$. הבעה כאן היא שאנו לא מתייחסים לרעש והמעבר בין הסיווגים. לכן העדיפה כי היא מתייחסת גם למצב ביןניים כלשהו וגם לרעש, וכך לסוג נגדיר עליה סף כלשהו כאמור. $w \phi_w(x_i) = x_i^T w$

נרצה פ' לא ליניארית שמדגימה את זה שיחסית ברור באיזה צד אנחנו החל מנקודת מסוימת, לדוגמה הפ' הלוגיסטיבית

$$\phi_w(x) = \sigma(w^T x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

ראו השוואה בין כל הפ' שדנו בהן עד כה (הסתודנטית המשקיעה תבין מי זו מי).



כל ש- $\|w\|$ יותר גדול, כך הפ' הלוגיסטיבית מפרידה באופן גס יותר (מושת יותר קרוב ל-1 או ל-0 ערכים במרכז) ואילו אם $\|w\|$ קטן מאוד הפ' הלוגיסטיבית תשפיע פחות על x ביחס למיקומו המקורי.

בגלל שאנו עוסקים באלג' סיווג הסט', עיקרונו הלמידה היחיד שRELONETI (מבין MLE ו-ERM) הוא (Maximum Margin MLE, ERM) כי הוא הסט' .

$$p_i = \sigma(x_i^T w), \text{ נניח } S = \{(x_i, y_i)\}, \text{ עבור } y_i \sim \text{Ber}(p_i)$$

$$\begin{aligned} \mathcal{L}(w | y) &= P(y_1, \dots, y_m | w) \\ &\stackrel{\text{i.i.d}}{=} \prod_{i=1}^m P(y_i | p_i) \\ &= \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1-p_i) \\ &= \text{זהות אלגברית} \prod_i p_i^{y_i} (1-p_i)^{1-y_i} \end{aligned}$$

$$\begin{aligned} \ell(w) &= \sum_i y_i \log p_i + (1-y_i) \log (1-p_i) \\ &= \sum_i y_i \log \frac{e^{w^T x_i}}{1+e^{w^T x_i}} + (1-y_i) \log \frac{1}{1+e^{w^T x_i}} \\ &= \dots \\ &= \sum_i y_i x_i^T w - \log \left(1 + e^{x_i^T w} \right) \end{aligned}$$

ונוכיח שהפ' הזו קמורה ומשם למצוא אומד MLE זה ממש פשוט.

בעבר הוכיחו שיש לנו איזושהי התפלגות על \mathcal{U} של לייבלים שימושפעת מהדגימות (לדוגמא ברגression לוגיסטי - הסט' עם הפ' הלוגיסטי) ושההשפעה הזו היא דטרמיניסטית. במקומות זאת, נניח שיש איזושהי הסט' משותפת \mathcal{F} על $\mathcal{U} \times \mathcal{X}$.

עתה, ניתן להסתכל על הלמידה בכמה דרכים,

$$\frac{f_{Y|X=x}(y) f_X(x)}{\text{הגישה הדיסקרטטיבית}} = f_{X,Y}(x,y) = \frac{f_{X|Y=y}(x) f_Y(y)}{\text{הגישה הganrtivit}}$$

עד כה עסקנו בגישה הדיסקרטטיבית, ככלומר בהינתן x דגימה כלשהי, בדקנו את ההסת' לקבל כל תווית ובחרנו את הטובה ביותר.

דוגמה נניח שיש לנו אוסף תמונות עם אובייקט עליהם ואנחנו צריכים לקבוע האם מדובר כלב או חתול. בגישה הדיסקרטטיבית, נזהה איזשהו פיצ'ר שונה בין כלבים לחתולים (כלבים לרוב יותר גודלים) ולפיו נבדיל ביניהם (יחד עם פיצ'רים אחרים). העיקרונו הוא שלא מספיק לנו איך נראה האובייקט, אלא רק שהוא כלב או חתול.

בגישה הganrtivit, בהינתן שיש לנו דגימה (תמונה של חתול), נרצה ללמד איך הוא נראה. זה מותbeta בנוסחה לכך שאנו בודקים בהינתן דגימה כלשהי, מה ההסת' לקבל מאפיין מסוים אצלו (גובה גובה).

$$\frac{f_{Y|X=x}(y)}{\text{posterior}} = \frac{\frac{\text{likelihood}}{f_{X|Y=y}(x) \cdot f_Y(y)}}{\frac{\text{prior}}{f_X(x)}} \cdot \frac{\text{evidence}}{f_X(x)}$$

כאשר הפוסטוריון הוא התפלג' אחרי שראינו את הדוגמה, הפרIOR הוא ההסת' לפני שראינו את הדוגמה (בכללי), הליקליזוד וחסירות קבל דוגמה עם ליבר צזה, והראייה לא צזו מעניינת.

כשאנו מבאים תווית, נרצה לבחור את הליבר שנותן לנו פостוריון הכى גבוה שאפשר, ככלומר

$$\hat{y}^{\text{MAP}} = \underset{y}{\operatorname{argmax}} f_{Y|X=x}(y) = \underset{y}{\operatorname{argmax}} f_{X|Y=y}(x) f_Y(y)$$

כאשר הראייה נעלמת כי היא קבועה. מסוג שעבוד על פי עקרון זה נקרא Bayes Optimal Classifier

טענה תהיו התפלגות \mathcal{D} על $\mathcal{Y} \times \mathcal{X}$ כאשר $\mathcal{Y} = [k]$, כלומר $\mathbb{1}[h(x_i) \neq y_i]$ אוסף אופטימלי ביחס ל- ℓ_{0-1} (המודרת ע"י).

$$L_{\mathcal{D}}(h^{Bayes}) \leq L_{\mathcal{D}}(h)$$

הוכחה:

$$\begin{aligned} L_{\mathcal{D}}(h) &= E_{x,y} [h(x) \neq y] \\ &= \int_{x,y} f_{X,Y}(x,y) \mathbb{1}[h(x) \neq y] dx dy \\ &= \int_x f_X(x) \sum_y f_{Y|X=x}(y) \mathbb{1}[h(x) \neq y] dx \\ (*) &= \int_x f_X(x) (1 - f_{Y|X=x}(h(x))) dx \\ &\geq \int_x f_X(x) (1 - f_{Y|X=x}(h^{Bayes})) dx \quad \text{לומד בייזני מצמצם פוסטוריוני} \\ &= E_{x,y} [h^{Bayes}(x) \neq y] \\ &= L_{\mathcal{D}}(h^{Bayes}) \end{aligned}$$

■ (*) במקומות לבדוק מה ההסת' לטעות (זה מה שהסכים עושים), זה המשלים להסת' שאנו צודקים.

הערה החולשה של לומד בייזני היא שהוא דורש פרIOR מדויק - ככלומר שנדע את ההסת' בעולם האמתי לתוויות.

ונכל עם זאת להניח כל מי הנחות על התפלגות המשותפת (לדוגמה שהוא גאוסיאני), ועל בסיס זה לבנות מסוג בייזני אופטימלי.

LDA

נניח שלכל $[m] i, y_i \in \{1, \dots, k\}$ נדגם מותoxic ($\pi_i = \text{multinom}(\pi)$ כאשר $\sum_i \pi_i = 1$, $\pi \in [0, 1]^k$) הtcpלגות מולטינומית אומרת שההסת' לקבלת את k היא π_k . אחרי זה, נניח שבבינהן k נדגם מותoxic (μ_k, Σ) Naive Bayes- Σ -היא אלכסונית.

כדי להשתמש במסוג שכזה, נצטרך לנחות מה הם הפרמטרים μ, Σ, π . בהנחה שהם לא נתונים לנו (וain סיבה שהם יהיו נתונים לנו), נוכל לחשב אותם באמצעות אומדי MLE, כאשר מקבלים בסופו של דבר

$$\begin{aligned}\hat{\pi}_k^{MLE} &= \frac{m_k}{m} = \frac{1}{m} \sum_i \mathbb{1}[y_i = k] \\ \hat{\mu}_k^{MLE} &= \frac{1}{m_k} \sum_{i:y_i=k} x_i \\ \hat{\Sigma}^{MLE} &= \frac{1}{m} \sum_i (x_i - \hat{\mu}_{y_i}^{MLE}) (x_i - \hat{\mu}_{y_i}^{MLE})^T\end{aligned}$$

כאשר m_k הוא מספר הדגימות עם תוויות k .

לאחר שהיחסנו אותם, נוכל בקלות לחשב את ה-BOC באמצעות הנוסחה

$$\hat{y} = \underset{k}{\operatorname{argmax}} \Sigma^{-1} \mu_k + \log \pi_k - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k$$

נצטרך להוכיח שגם הנתונים שלנו אכן עוניים על הדרישות הינ'ל, או הולמד אכן BOC.

ב-QDA עושים אותו הדבר רק שבמקרה להנחת $x_i | y_i$ מתפלגים בשונות זהה אבל תוחלת אחרת, משחררים גם את השונות ולבן ב-BOC. במקרה כזה הכל נשאר אותו הדבר חוץ מזיה שחשונות (לפי MLE) היא בעצם $x_i | k \sim N(\mu_k, \Sigma_k)$

$$\hat{\Sigma}_k^{MLE} = \frac{1}{m_k} \sum_i (x_i - \hat{\mu}_{y_i}^{MLE}) (x_i - \hat{\mu}_{y_i}^{MLE})^T$$

שבוע VII | שיטות מכלול

הרצאה

עד כה ראיינו אלג' סיוג ספציפיים והיום עוסוק במתא-אלג', שיטות שימושísticas באלג' קיימים כדי לשפר ביצועים - והשיפור יכול להיות מאוד שימושתי.

אנחנו עוסקים בבעיות סיוג עם $\{ \pm 1 \} = \mathcal{U}$ (אפע"פ) שקל להכליל הכל למחלקות רבות ולביעות ריגרטייה. נזכר בהתייה ושונות: מחלוקת היפותזות עשרה היא בעלת הטיה מאוד נמוכה (שגיאה מוכללת נמוכה) כי נוכל להתאים לאמת טוב מאוד ולהפק. שונות היה כמה אנחנו מושפעים ממודגס אימון ספציפי, ככל שחלוקת היפותזות יותר עשרה, כך קל יותר להתאים לאמת (הנתונה

לנו) ולהיות מושפעים ו “לפספס את התמונה הכללית”. נראה מטא-אלג' שמוריד את ההטיה ואחד שמוריד את השונות. הפעלה של מטא אלג' אחד על השני תעלה חשיבות הרבה מאוד והשיפור בשלב ההוא לא תמיד ניכר ולכון זה לא מוכנת כס' אינסופית (אבל כן סופית).

ועדות

נניח שיש לנו ועדה עם T חברים שכל אחד צודק בדייבד בהסת' p וטעעה בהסת' $p - 1$. נניח שככל החברים חכמים באותה המידה ונבחר לפי הרוב. פורמלית, יש לנו $X_1, \dots, X_T \stackrel{\text{i.i.d}}{\sim} \text{Ber}(p)$. ההחלטה של הוועדה היא $\bar{X} = \text{sign}\left(\sum_{t=1}^T X_t\right)$. אם כל חבר צודק בדייבד בהסת' p , עדיף כבר שאחד יבחר. אם $0.5 < p$, ההסת' שהרוב יצודק הרבה יותר גבוהה משל יחיד.

תרגיל הסטודנטית המשקיעה תחשב את ההסת' של בחירה נכונה כפ' של p ו- T ומה הגבול הזה כ- T -ושאך לאינסוף.

ניתן להראות כי אם X_i הם i.i.d עם שונות σ^2 , אז השונות של $\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t$ היא $\frac{\sigma^2}{T}$, ולכון נסיק כי הוועדה אכן עיקבית אם היא מתבקשת לבצע החלטה באופן ב"ת כמה פעמים על אותם הנזונים.

תרגיל הסטודנטית המשקיעת תחשב את השונות של מ"מ ברנולי ומשם את השונות של הוועדה כפ' של T ו- p .

במציאות המ"מ אינם ב"ת. נניח כי הקורלציה בין כל שני חברים היא ρ .

טענה אם X_1, \dots, X_t נבדקים מאותו מ"מ ממשי עם שונות σ^2 ושונות משותפת ρ בין כל שני מ"מ, אז השונות של \bar{X} היא $\sigma^2 + (1 - \rho) \cdot \frac{\sigma^2}{T}$

מכאן ניחח השראה ל-ML, ונוכיח שאכן אם ננהל ועדה של לומדים ונקבע לפי החלטת הרוב, קיבל שנות יותר נמוכה (בנחה שלא נאמן את אותו המודל או מודלים עם שונות משותפת גבוהה מדי כי אז לא יהיה שיפור משמעותי ולא נשימוש בלומד עם $0.5 < p$) שדוועכת אקספוננציאלית ב- T .

שיטות ועדה

שיטות ועדה הן מטא אלג' שלוקחים לומדים ומיזינים אל תוכם T מוגני אימון מלאכותיים.

הערה בהרצאה זו אנחנו עם סיוג ביןארי וכן כאמור מטיימים נוכל למשקל את הסכום.

הערה בריירסיה אפשר לבחור $h(x) = \frac{1}{T} \sum_{t=1}^T h_t(x)$ כהחלטה הוועדה.

איך ניצור מידע יש מעין? בהינתן $S = \{(x_i, y_i)\}_{i=1}^m$ ובהמשך גם S^{*2} וכו'.

Bootstrap

נדגים m פעמים מתוך S עם חוזרות. הדגימה ה- i -הנדגמה מסומנת ב- x_i^{*1}, y_i^{*1} ואז מוגן האימון החדש הוא (שאלולי מכיל חוזרות). נחזור על הפעולה זו מספר פעמים ככל שנרצה.

לכארה, דומה שיטתה זו לא באמת תורמת לשום דבר. אם נניח S נדגם מהתפלגות \mathcal{D} על $\mathcal{X} \times \mathcal{Y}$ אז כל בוטסטראפ מקרב דגימות p.i.i-m- \mathcal{D} המקורי שאינו ידוע לנו.

הגדירה התפלגות האמפירית $\hat{\mathcal{D}}_S(C) = \frac{|C \cap S|}{m}$ על $\mathcal{X} \times \mathcal{X}$ הוא התפלגות המושרha מ- S כולם איחוד על S ו-0 מחוץ לו.

בוטסטרואפ הוא למעשה מושג p.i. מtower $\hat{\mathcal{D}}_S$ ואינטואיטיבית לפחות $\hat{\mathcal{D}}_S$ מתכנס ל- \mathcal{D} (הסתודנית המשקיעה תעשה סימולציה ותגלה זאת) ובתקופה ש- S לא רחוק מדי מ- \mathcal{D} , הבוטסטרואפ הוא מוצלח.

טענה בتوزלת, בערך שליש מהדוגמאות נשארות בתוך ה"שך" ולכן אכן יהיו חוזרות.

Bagging

בגיג הוא שימוש של בוטסטרואפ בלמידת מכונה שמשפר את הדיוק של אלג' למידה supervised. מתחילה מאלג' A . ומודגמים אימון S . בוחרים T ובונם T מוגמי אימון של בוטסטרואפ $S^{*1}, S^{*2}, \dots, S^{*T}$ בגודל m כל אחד. נותנים לומד ללמידה כל מודגם אימון בנפרד. מרכיבים ועדה $h_{bag}(x) = \text{sign}\left(\sum_{t=1}^T h_{S^{*t}}(x)\right)$ באמצעות כלל החלטה $x \in h_{S^{*1}}, \dots, h_{S^{*2}}$ אם הלומד שלנו הוא עז החלטה, העצים שנוצרם מכל מודגם אימון הם אחרים.

הערה באיגינג הוא מאוד אפקטיבי עד רמה מסוימת, ואחרי הרף זהה לא ממש משתפרים הביצועים (כי הקורלציה המשותפת כבר גבוהה מדי).

דה-קורלציה

ישנן שיטות שונות להקטנת השונות המשותפת של לומד שאומן באמצעות באיגינג, בראשן עצים מקרים. עז מקרי הוא אלג' לומד שמבצע באיגינג על עצי החלטה שטם מבצע דה קורלציה. הוא משתמש בפרמטר $d \leq k$ שאיתו שמאגדלים כל עז מקרי הוא אלג' לומד שמבצע באיגינג על עצי החלטה שטם מבצע דה קורלציה. הוא משתמש בפרמטר d שמאגדלים כל עץ החלטה, בכל פיצול, נבחר באקראי k מtower d הקורדיינטות וرك עליהם נאפשר לפצל (נזכיר כי עצי ההחלטה בוחרים ציר ונΚודה אחת שכל מה שפניהם בציר הנוכחי הולך לצד אחד וכל מה שאחראית לצד השני).

הণיון הזה מעלה את ההטייה של העץ, אבל הדה-קורלציה מאוד משמעותית כי בכלל שאלג' הפיזול הוא חמדן, פיזול בנקודה אחרת משנה לוגרי את כל העץ וזה מושגים קורלציה נמוכה יותר.

פרמטרים של עז מקרי

- העומק המקסימלי של כל עץ ההחלטה. $R \in \mathbb{N}$.
- m_{min} מספר הדוגמאות המינימלי בכל עלה.
- T מספר מוגמי הbagging (וכך גם מספר העצים).
- k מספר הקורדיינטות שנדרשות עליהם אפשר לפצל.
- פרמטר לגיזום העץ שבוណון בהערכתה הבאה.

פסאודו-קוד לאלג' עז מקרי

• לכל $t = 1 \dots T$

- דוגם מודגס אימון S^{*t} - S .

- אמן עץ החלטה $h_{S^{*t}}$ על S^{*t} ובמהלך הגידול, בכל פיצול:

* בחר k קורדייניות מותoxic $\{1, \dots, d\}$, באופן אחד.

* בחר את הקורדייניטה והנקודה בה הטובים ביותר רק מקורדייניות שנדגמו.

* פצל על הבחירה.

- אל תפצל קופסה אם עומק המקסימלי R או אם המספר המינימלי של דגימות m_{min} הושג.

• החזר את $.h_{S^{*1}}, \dots, h_{S^{*T}}$.

באגינג פוגע לנו רק אם הלומד שלנו גרוע. החסרונות שלו הם שהוא קשה לפרשון, שהוא דורש T מודלים לאחסן וצריך לנבא T פעמים כל פעם, וכמוון העול החישובי.

עם זאת, באגינג משפר דרסטית את השונות ולא מגירע את ההטיה יותר מדי. מבחינת חישוביות - ביגינג הוא **embarrassingly parallelizable** - כלומר קל מאוד למקבל אותו (העצים הם ב"ת').

תרגיל הסטודנטית המשקיפה ת מלא ת"ז על עצים מקרים, כפי שעשינו כאן.

הערה באמצעות הסטת רף הרוב, אפשר לקבוע דרישות יותר נוקשות/מקלות על ההחלטה, וכך גם מקבלים שערוך על הסט' המחלקות השונות.

Boosting

בוסינטג לוקח אלג' לומד חלש (שגיאה הכללה גבוהה) ומsharp אותו עם שיטת ועדה לאלג' עם דיוק גבוהה.

באגינג העמדיינו פנים שיש לנו הרבה דגימות חדשות מאותה התפלגות מגולמת. בעצם, כל חבר ועדה יהיה כל החלטה שנובע מדוגמאות אימון S_t של התפלגות אחרת \mathcal{D}^t . לעומת גיג שאמנו באופן ב"ת את החברים, בוסינטג מלמד באופן סדרתי את הלומדים, כאשר כל מוגם אימונו חדש הוא יותר טוב מהקודם.

הרעיוון החכם שמאחורי בוסינטג הוא שאימנו את h_t בהתבסס על \mathcal{D}^t , אנחנו מעדכנים את התפלגות ומעלים את הסט' ליפול על דגימות ש- h_t טעה בהן, ואז לא- h_{t+1} יהיה קשה להטעם מהשגיאות ויתקן אותן וכך הלאה. בגלל שאנחנו ממשקלים את התפלגות והכל צריך להסתכם ל-1, הסט' ליפול על דגימות שצדקו בהן תהיה קטנה יותר.

לבסוף נבצע החלטת ועדה ממושקלת של $.h_1, \dots, h_T$.

אפשר לאמן את h_t עם בוטסטראף ממושקל בו ההסת' ל- S היא $(x, y) \in S$ אבל לעיתים אפשר יותר טוב.

אם עובדים עם ERM, אפשר לשנות את הסיכון האמפירי להיות $L_{S, \mathcal{D}^t}(h) = \sum_{i=1}^m \mathcal{D}_i^t \mathbf{1}[y_i \neq h(x_i)]$ כאשר $\sum_{i=1}^m \mathcal{D}_i = 1$

Adaboost

יש הרבה מאוד אלג' בוסטינג, כל אחד בונה אחרית את ההתפלגות החדשה \mathcal{D}^t בהינתן \mathcal{D}^{t-1} וכן את המשקלות הסופיו של הועדה. Adaptive Boosting :

$$\mathcal{D}_i^t = \frac{1}{m} \bullet$$

•

$$\mathcal{D}_i^{t+1} = \frac{e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m \mathcal{D}_j^t e^{-w_t y_j h_t(x_j)}} \mathcal{D}_i^t$$

כאשר למעשה אנחנו מבצעים עדכון אקספוננציאלי עם נורמל. אם ($h_t(x_i)$ צדק אז הוא באותו סימן כמו y_i ולכנו ההתפל' תדעך אקספ'). אם טעינו, ההתפלגות תגדל אקספ! נגיד w_i כך ש- $\sum_{i=1}^m \mathcal{D}_i^{t+1} \mathbb{1}[y_i \neq h_t(x_i)] = \frac{1}{2}$ כלומר שhhסת', לפחות דוגמיה שטיענו בה בפעם הקודמת היא בדיק $\frac{1}{2}$, שזה הכח גרווע שיכול להיות. בחרה של $\epsilon_t = \sum_{i=1}^m \mathcal{D}_i^t \mathbb{1}[y_i \neq h(x_i)]$ $w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$ נותרת את התוצאה הרצוייה.

- ההחלטה הסופית היא $h_{boost} = \text{sign}\left(\sum_{t=1}^T w_t h_t(x)\right)$

להלן הפסאודו-קוד של Adaboost, שמקבל מקלט מדגם אימון S בגודל m , לומד \mathcal{A} . שמקבל מדגם אימון והסת' על הנקודות (לדוגמא ERM כאמור או בוטסטראף ממושקל אם אין ברירה) ומספר איטרציות T .

$$\mathcal{D}^1 = \left(\frac{1}{m}, \dots, \frac{1}{m}\right)$$

• לכל $t = 1 : T$ עד :

$h_t = \mathcal{A}(\mathcal{D}^t, S)$ – נקרא לומד

$$w_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right) \text{ ו } \epsilon_t = \sum_{i=1}^m \mathcal{D}_i^t \mathbb{1}[y_i \neq h_t(x_i)]$$

– נחשב – מעדכן

$$\mathcal{D}_i^{t+1} = \frac{\exp(-w_t y_i h_t(x_i))}{\sum_{j=1}^m \exp(-w_t y_j h_t(x_j))} \mathcal{D}_i^t$$

לכל $i \in [m]$

$$h_{boost}(x) = \text{sign}\left(\sum_{t=1}^T w_t h_t(x)\right)$$

הגדירה נאמר כי \mathcal{A} הוא לומד γ -חלש עבור מחלוקת היפוטזות \mathcal{H} אם קיימת $\mathbb{N} \rightarrow (0, 1)$ m כך שלכל $1 < \delta < 0$ ולכל התפלגות \mathcal{D} על \mathcal{X} ולכל כלל החלטה $f : \mathcal{X} \rightarrow \{\pm 1\}$, אם הנחת הרילאייזביליות מתקיימת ביחס ל- $\mathcal{H}, \mathcal{D}, f$ אז בהערכת \mathcal{A} על S מוגדל (δ)

שנדגם d.i.i. מ- \mathcal{D} עם ליבלים מ- f , מתקיים

$$P \left(L_{\mathcal{D},f} (h_S) \leq \frac{1}{2} - \gamma \right) \geq 1 - \delta$$

נאמר כי \mathcal{H} היא γ -חלשה-למידה אם קיים $L_{\mathcal{H}}$ לומד γ -חלש.

הערה הרעיון הוא שהאלג' מבטיח רמת דיקט מסוימת, בהסת' גובה כל שנרצה וככל ש- γ יותר גדול, כך הדיקט גובה יותר. זהה הגדרה חלשא בהרבה לממדות-PAC שכן אכן אנחנו אומרים קיים $\epsilon (\gamma - \frac{1}{2})$ ולא כלל.

ההיסטוריה, בוסטיניג נבע מלמדות PAC. אם \mathcal{H} אומר $L_{\mathcal{H}}$ לומד אותה היבט. אם $ERM_{\mathcal{H}}$ קשה לחישוב, כפי שכבר ראיינו, נוכל למצוא מחלוקת היפותזות פשוטה $ERM_{\mathcal{H}_{base}}$ כך ששיעור של $ERM_{\mathcal{H}_{base}}$ הוא לא קשה מדי ומהווה לומד γ -חלש ל- \mathcal{H} עבור γ כלשהו. ככלומר, אנחנו יכולים חשב ביעילות עם דיקט $\gamma - \frac{1}{2}$ ואולי אי אפשר למצוא בכלל לומד עילם γ יותר גבוה. נוכל לעשות בוסטיניג ל- $ERM_{\mathcal{H}_{base}}$. באופן יעיל חישובית ולהציג ביעילות לומד שմקרב את $ERM_{\mathcal{H}}$.

משפט יהיו S מדגם אימון. נניח שבכל איטרציה, Adaboost משיג שגיאות הכלכלה אמפירית γ הסיכון האמפירי של h_{boost} מקיים

$$L_S (h_{boost}) = \frac{1}{m} \sum_{i=1}^m \mathbb{1} [y_i \neq h_{boost} (x_i)] \leq \frac{1}{2} - \gamma$$

מסקנה הפעלה סדרתית של Adaboost מצמצמת באופן אקספוננציאלי את הסיכון האמפירי.

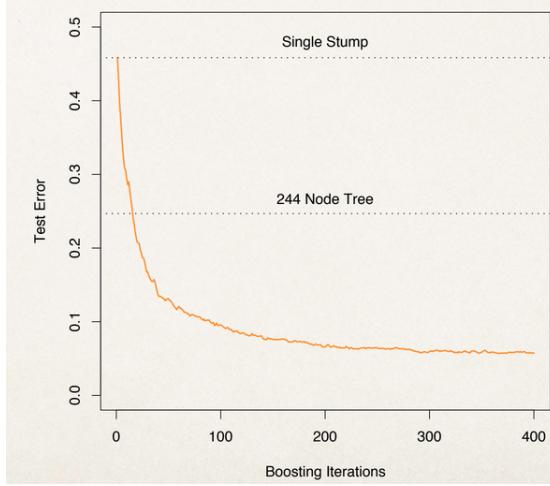
מחזיר כלל החלטה מתוך מחלוקת היפותזות Adaboost

$$\mathcal{H}_T = \left\{ x \mapsto \sum_{t=1}^T w_t h_t (x) : w_t \in [0, \infty), \sum w_t = 1, h_t \in \mathcal{H}_{base} \right\}$$

שזה למעשה אוסף כל הקומבינציות הקמורות בגודל T מתוך \mathcal{H}_T .

יוצא שהשונות עולה כמעט תמיד אבל ההטיה יורדת דרסטית (הרוי אנחנו מותאים כל איטרציה למציאות), ולא מגיעים ל-*overfitting* מדי מהר.

בתנאים מסוימים, $\text{VCdim}(\mathcal{H}_T) = T \cdot \text{VCdim}(\mathcal{H}_{base})$ (כזכור השונות מושפעת מעשירות המחלוקת). מעשה הירידה בהטיה היא אקספוננציאלית (ראו איור) והזוק לא יותר מדי גדול (*overfitting* ב- T -ים גדולים).



לרוב בostoning לומדי ERM פשוטים מוצלח יותר מבostoning לומדי ERM של מחלקות בסיס מורכבות יותר.

בגיניג	בostoning	
במקביל	סדרתי	לומד חברים
בostoneraf	בostoneraf ממושקל או ERM ממושקל	מדגמי האימון של החברים
מומלץ	לא עוזר	זה-קורלציה
אין אובר-פייט	כן, מעט	אוברפייט כ- T -גדול מדי
ירידיה בשונות	ירידיה בהטיה	שייפור
עצים عمוקים	עצים רדודים	בשימוש בעצים מקרים, נבחר
בקלות	אין	IMPLEMENTATION
לא ממושקלת (ממוצע קלאס)	ממושקלת	הצבעת הועדה

לxicom למדנו על bostoneraf שהוא עיקרונו היסוד של בגיניג שיוצר ועדת של הלומד עם עצמו על מדגמי אימון שונים (מתוך bostoneraf). bostoning יוצר ועדת ממושקלת אקספוננציאלית עם חברים שמשתפרים כל פעם והוא עובד היטב עבור לומדים γ -חלשים עם $0 < \gamma$.

תרגול

כרגע אנחנו מסתכלים על $\{\pm 1\}^m$. \mathcal{H} ב-PAC אנחנו מצימים $\mathcal{L}^{\mathcal{X}} \subseteq \mathcal{Y}$ (כי אחרת אי אפשר למוד מ-NFL) ומתייחסים ל蹶ה הריאלייזבילי. לאחר מכן נמצא אלג' לפי איזשהו עקרון מיידה $\mathcal{H} \rightarrow (\mathcal{X} \times \mathcal{Y})^m$. נסתכל על שגיאת ההכללה, קרי $[h_S = \mathcal{A}(S)]$ כאשר $L_{\mathcal{D}} = E_{x \sim \mathcal{D}} [h_S(x) \neq f(x)]$.

מחלקת ההיפותזות היא PAC למידה אם היא עוברת הכללה היבט, כלומר, אם לכל רמת דיווק וביתחון שנרצה, יש מספר דוגימות מינימלי שאחריו אנחנו מכך בדיק גובה בביטחון גובה (אותם פרמטרים שביקשנו).

דוגמה נסתכל על פ' אמיתית שהיא פ' סף מהצורה $I = [a, b]$ כאשר $f(x) = \begin{cases} 1 & x \in I \\ -1 & x \notin I \end{cases}$ ככלומר

$$\mathcal{H}_{interval} = \left\{ h_{a,b}(x) = \begin{cases} 1 & x \in [a, b] \\ -1 & x \notin [a, b] \end{cases} : a < b \in \mathbb{R} \right\}$$

נctrיך לתאר פ' סיבוכיות מודגס ואלג' למידה כדי להוכיח ש- $\mathcal{H}_{interval}$ -למידה.

$$\text{כולם הדוגמיה הראשונה והאחרונה שהם עם לייבל } 1. \hat{a} = \max_{i:y_i=1} x_i \text{ ו-} \hat{b} = \min_{i:y_i=1} x_i \text{ כך } \mathcal{A}(S) = (\hat{a}, \hat{b})$$

טענה $\mathcal{H}_{interval}$ היא PAC-למידה.

הוכחה: יהיו $\epsilon > 0$ ויהי \mathcal{D} התפלגות על \mathcal{X} . נשים לב כי אם ההסת' לקבלת ערכים בתחום האינטראול היה נושא קשה ללמידה ונראה זאת בסיבוכיות המדגם. נוכיח כמה שיותר מהטענה עד שנוכל בקלות להגעה לסיבוכיות המדגם ויחד זה ישלים את ההוכחה. הסיכוי לטיעות הוא הסיכוי לפיה ההתפלגות ליפול בתחום $L_{\mathcal{D}}^- = P([a, \hat{a})), L_{\mathcal{D}}^+ = P([\hat{b}, b])$ או $[a, \hat{a}], [\hat{b}, b]$, כלומר, עבור מתקיים

$$\begin{aligned} L_{\mathcal{D}}(h_S) &= E_{\mathcal{D}}[h(x) \neq I] \\ &= P(x \in [a, \hat{a}] \vee x \in [\hat{b}, b]) \\ &\text{זרים} = P(x \in [a, \hat{a}]) + P(x \in [\hat{b}, b]) \\ &= L_{\mathcal{D}}^- + L_{\mathcal{D}}^+ \end{aligned}$$

נשים לב כי מתקיים $\{L_{\mathcal{D}}(h_S) \leq \epsilon\} \supseteq \{L_{\mathcal{D}}^- \leq \frac{\epsilon}{2} \wedge L_{\mathcal{D}}^+ \leq \frac{\epsilon}{2}\}$

$$\begin{aligned} P(L_{\mathcal{D}}(h_S) \leq \epsilon) &\geq P\left(L_{\mathcal{D}}^- \leq \frac{\epsilon}{2} \wedge L_{\mathcal{D}}^+ \leq \frac{\epsilon}{2}\right) \\ &= 1 - P\left(L_{\mathcal{D}}^- > \frac{\epsilon}{2} \vee L_{\mathcal{D}}^+ > \frac{\epsilon}{2}\right) \\ &\text{בב''כ } L_{\mathcal{D}}^- > \frac{\epsilon}{2} \geq 1 - 2P\left(L_{\mathcal{D}}^-(h_S) > \frac{\epsilon}{2}\right) \\ &= 1 - 2P(x_1 \notin [a, \hat{a}] \wedge \dots \wedge x_m \notin [a, \hat{a}]) \\ &= 1 - 2 \prod_{i=1}^m P(x_i \notin [a, \hat{a}]) \\ (*) &= 1 - 2 \prod_{i=1}^m \left(1 - \frac{\epsilon}{2}\right) \\ &= 1 - 2 \left(1 - \frac{\epsilon}{2}\right)^m \end{aligned}$$

(*) מניחים שההסת' לקבלת בתחום $[a, \hat{a}]$ היא $\frac{\epsilon}{2}$ זו אבstrקציה לא מעניינת כדי להקל על ההוכחה.

לכן עבור נרצה מספיק דוגמאות כדי שיתקיים $m_{\mathcal{H}} = \left(\frac{2 \log \frac{2}{\delta}}{\epsilon}\right)^m$, נתנו לנו, אנחנו בוחרים רק את m , כלומר $\delta = 2 \left(1 - \frac{\epsilon}{2}\right)^m$

הערה אינטואיטיבית, מימד VC הוא גודל הקבוצה המקסימלי שאפשר לקבל את כל האפשרויות ללייבלים בהם, ואם אין הגבלה על גודל קבוצה כזו, הקבוצה כללית מדי ולכון אי אפשר ללמוד אותה.

הגדירה הגבלה של \mathcal{H} על \mathcal{X} $\subseteq \mathcal{Y}^{\mathcal{X}}$ על \mathcal{H} $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ $\{h(x_1), \dots, h(x_m)\} \subseteq \mathcal{H}$ להיות.

דוגמה $\mathcal{H}_C = \{h_C = (h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\}$. כל קבוצה בגודל 1 $\{x\}$ היא מנוטצת כי אפשר לבחור סינגלטון שהוא x (בשביל לייבל 1) או סינגלטון שהוא לא x בשביל לייבל 0.

קבוצה בגודל 2 אי אפשר לנתח. לדוגמה $C = \{x, y\}$ אי אפשר להשיג כי 00 אפשר (סינגלטון שהוא לא x ולא y), 10, 01, 11 אפשר גם אבל אי אפשר 11 כי $y \neq x$ ואין סינגלטון שנית 1 על שני ערכים שונים. לכן מימד ה-VC של $\mathcal{H}_{singleton}$ הוא 1.

דוגמה מה מימד ה-VC של $\mathcal{H}_{interval}$? קבוצה 1 מנוטצת - ברור. קבוצה בגודל 2 (נבחר אינטervalים שכך/לא על הנקודות וכו'). אי אפשר לנתח קבוצה מגודל 3 כי דבר מהצורה 101 אי אפשר להשיג (אינטerval הווה תמיד 010). לכן $\text{VCdim}(\mathcal{H}_{interval}) = 2$. לכן מהמשפט היסודי, PAC-למידה.

דוגמה מהו מימד ה-VC של \mathcal{H} ? קבוצה בגודל 1 אפשר לנתח כי פשוט בוחרים על מישור בהתחם. קבוצה בגודל 2 גם אפשר לנתח. קבוצה בגודל 3 לא ניתן לניתוץ. נראה זאת אינטואיטיבית. לכל w , אם כל הנקודות באותו הצד של העל מישור, אי אפשר לקבל מצב עם לייבלים מעורבבים, אם הנקודות בצדדים שונים, אי אפשר לקבל אלה עם אותו הליבל כולם.

טענה $\text{VCdim}(\mathcal{H}) = d$.

הוכחה: ראשית נוכיח כי $d = \text{VCdim}(\mathcal{H}) \geq |C| = d$. מספיק שנוכיח שקיים \mathcal{X} כך ש- \mathcal{X} מנוטצת. נבחר $C = \{e_1, \dots, e_d\}$. הבסיס הסטנדרטי.

יהי $y = \{\pm 1\}^d$ סדרת לייבלים. נגדיר $w = y$ ונראה כי $h_w(C) = \sum_{i=1}^d h_w(e_i)$.

$$h_w(e_i) = \text{sign}(\langle e_i | w \rangle) = \text{sign}(\langle e_i | y \rangle) = \text{sign}(y_i) = y_i$$

כלומר C מנוטצת ע"י \mathcal{H} .

נוכיח כי $d = \text{VCdim}(\mathcal{H}) \leq |C| = d+1$. יהיו $x_1, \dots, x_{d+1} \in \mathcal{X}$ כך ש- \mathcal{X} קבוצה בגודל 1 + d בתוך מ"ז (\mathbb{R}^d מימד d , הרו שקיים α_i ממשי $\alpha_i \neq 0$ ו $\sum_{i=1}^{d+1} \alpha_i x_i = 0$). בה"כ $\alpha_{d+1} \neq 0$ ולכון

$$x_{d+1} = \sum_{i=1}^d \frac{\alpha_i}{\alpha_{d+1}} x_i = \sum_{i=1}^d b_i x_i$$

$$y_i = \begin{cases} \text{sign}(b_i) & i \in [d] \\ -1 & i = d+1 \end{cases} \quad \text{נניח בשלילה שקיימים על } b_i \text{ הם סימוני}. \text{ נראה כי לא נוכל לקבל את } y \in \{\pm 1\}^{d+1} \text{ המוגדר ע"י}$$

מישור w שמקיימת $y = h_w(C)$

$$\begin{aligned} -1 &= \text{sign}(\langle w | x_{d+1} \rangle) = \text{sign}\left(\left\langle w | \sum b_i x_i \right\rangle\right) \\ &= \text{sign}\left(\sum b_i \langle w | x_i \rangle\right) \stackrel{(*)}{=} \text{sign}(\geq 0) = 1 \end{aligned}$$

(*) הסימן של b_i הוא הסימן של $\langle w | x_i \rangle = y_i$ (השוון מתקיים מההנחה בשלילה).

■ $\text{VCdim}(\mathcal{H}) = d$ לכן

הערה ככלל כדי להוכיח מימד VC, צריך לקבל אינטואיציה למימד (לנחש לצורך העניין), להוכיח שקיימת קבוצה מנותצת בגודל כלשהו ואז לכל קבוצה גדולה ממנה (ובפרט בגודל 1 יותר) היא לא מנותצת.

הגדירה $x \in \mathbb{R}^d$ מונום הוא פ' מהצורה $\prod_{j=1}^d x_j^{n_j}$ כאשר $n_j \in \mathbb{N}_0$ והדרגה שלו היא k .

הגדירה פולינום הוא צירוף לינארי של מונומים.

הערה ניתן לרשום $\langle p(x) | \psi \rangle$ כאשר $p(x)$ הוא הוקטור עם כל המונומים על d ממדים מדרגה לכל היותר k . לכן מדיסקרטית, $\psi \in \mathbb{R}^{\binom{d+k-1}{k}}$ כל האפשרויות בעלי סדרה עם חזרה). יחד עם w , מקבלים צ"ל של מונומים כאשר אם הוא לא נמצא בפולינום יש סקלר 0-ב- $w \in \mathbb{R}^{\binom{d+k-1}{k}}$ במקום הרלוונטי.

דוגמא מההערות הנ"ל, \mathcal{H} היא פשוט אוסף העל מישורים מעל \mathbb{R} ולכן מימד ה-VC שלה הוא $\binom{d+k-1}{k}$ (באמצעות ההוכחה הקודמת).

טענה כל מחלקה היפוטזות סופית היא למידה PAC.

שבוע VII | טריזידוף הטיה-שונות וbossting

תרגול

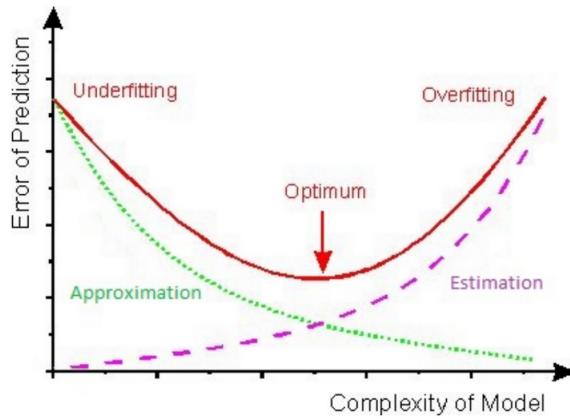
$h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$ ונרצה למצוא את $\mathcal{X} \xrightarrow[\mathcal{H}, \mathcal{D}]{} \mathcal{Y}$ נניח שאנו משתמשים ב-ERM, כלומר $h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_{\mathcal{D}}(h)$

$$L_{\mathcal{D}}(h_S) = \frac{L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*)}{\epsilon_{approx} + \epsilon_{est}}$$

אנחנו לא יודעים לחשב את $\epsilon_{approx}, \epsilon_{est}$ אבל נרצה לモער אותם באמצעות איכשהו.

• נובע מחלוקת היפוטזות ולא קשרו ל- S - היפוטזה הכח טובה ב- \mathcal{H} היא זו שגורמת לו וככל שהמחלקה יותר עשירה כך יהיה יותר נמוך. זהה למעשה ההטיה.

• ϵ_{est} תלוי ב- S והוא מבטא כמה רוחק S הוליך אותנו מהדבר הכח טוב רק כי הוא לא מייצג (אם זה מאוד גדול, L_S מאוד שונה מ- $L_{\mathcal{D}}$). זהה השונות על \mathcal{H} .



זה שונה ממה שראינו עם MSE (שם שגיאת ההכללה הייתה הטעיה + השונות ברכיבים) כי אנחנו לא יודעים מהי שגיאת ההכללה, הנוסחה כאן היא כללית.

דוגמא נניח שאנו רוצים לזרום ספנות בכתב ידי. במקומות החינוך למתמטיקה עם דברים מאוד מורכבים, אפשר לבחור לחס恬ך רק על פיקסל אחד ולפיו לקבוע אם זה 0 או 1 לדוגמה. זה לא לומד יותר מדי טוב אבל יחד עם עוד כמה לומדים מאוד בסיסיים כמו זה, יכול להיות שהוא מוצלח.

חזרנו על ההגדרה של לומד γ -חלש, שהוא לומד כך שלכל רמת ביטחון δ ישנה סיבוכיות מוגבלת כלשהי שאחריה שגיאת ההכללה של (S, \mathcal{A}) היא לכל היותר $\gamma - \frac{1}{2} \ln \delta$.

משפט אם מחלוקת היא γ -חלש-למידה (ע"י אלג' כלשהו) אז היא PAC למידה (אולי עם אלג' אחר).

ועדה עם חברים ב"ת

יהיו X_1, \dots, X_T כארה הרנוולי הוא על $\{0, 1\}$ במקומות $p > 0.5$. נניח שההתשובה הנכונה היא 1, מה הסבירו של X לצדו?

$$P(X) = P(|\{X_i\}| > |\{\text{שווים } X_i\}|) = P(X > 0)$$

נחשבים מלמטה את הסיכוי שצדקו עם פיתוח דומה לא"ש צ'רנוフ.

$$\begin{aligned} P(X \leq 0) &\stackrel{a \geq 0}{=} (-aX \geq 0) = P(e^{-aX} \geq e^0) \\ &\leq E[e^{-aX}] = E[e^{-a(\sum X_i)}] \\ &\stackrel{\text{i.i.d}}{=} E[e^{-aX_1}]^T \end{aligned}$$

מתקיים

$$E[e^{-aX_1}] = pe^{-a} + (1-p)e^a \stackrel{\text{אינפ}}{\leq} e^{a-p-pe^{-2a}}$$

לכן עבור $a = \frac{1}{2} \ln(2p)$ נקבל מתקיים

$$P(X \leq 0) \leq \left(e^{a-p-pe^{-2a}}\right)^T = \dots \leq \exp\left(-\frac{T}{2p}\left(p - \frac{1}{2}\right)^2\right)$$

ולכן $P(X > 0) \geq 1 - \delta = 1 - e^{-\frac{T}{2p}(p - \frac{1}{2})^2}$. כמובן, ככל ש- T גדול יותר לאפס ואז הביטחון שלנו עולה. בנוספ', ככל ש- p גדול, ההסת' לבודוק בועדה גם עולה.

ועדה עם קורלציה בין חברים

$$. X = \frac{1}{T} \sum X_i \text{ ונדיר } X_1, \dots, X_T \stackrel{\text{i.i.d}}{\sim} \text{Ber}(p)$$

$$E[X_i] = p \cdot 1 + (1-p)(-1) = 2p - 1$$

$$\text{var}(X_i) = \dots = 4p(1-p)$$

ולכן $E[X] = 2p - 1$ ו- $\text{var}(X) = \frac{1}{T^2} T 4p(1-p) = \frac{4p(1-p)}{T}$. זה המקרה בו הם ב"ת, שהוא כמובן לא ריאליסטי. הcorr(X_i, X_j) = ρ ו- $\text{var}(X_i) = \sigma^2$ עם $X_1, \dots, X_T \stackrel{\text{i.d}}{\sim} \mathcal{D}$

$$\begin{aligned} \text{var}(X) &= \frac{1}{T^2} \text{var}\left(\sum X_i\right) \\ &= \frac{1}{T^2} \left(\sum \text{var}(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j) \right) \\ &= \frac{1}{T^2} \left(T\sigma^2 + 2 \binom{T}{2} \rho \sigma^2 \right) \\ &= \rho \sigma^2 + \frac{1}{T} (1 - \rho) \sigma^2 \end{aligned}$$

זהה גם מתכנס ל-0 כ- T גדול, כלומר אנסמבל הוא אכן רעיון טוב.

הreasון בבוסטיניג הוא ליצור אנסמבל משיפורים של מודל בסיס כלשהו. הדוגמה הקלאליסטית לבוסטיניג הוא Adaboost, שראינו בהרצאה. הרעיון שלו כאמור הוא לתת משקל יותר גבוה לנוקודות שעבורן הלומד טעה כדי שייכאב לו לטעות עליהן בהמשך.

הreasון בבוטסטראף זה ש- S מייצג בצורה כזו או אחרת את ההתפלגות המוגולמת \mathcal{D} , והגירה של תת-مدגמים מתוך S לאומד תאפשר לנו ללמידה בשיטות שונות את ההתפלגות ובכך להוריד את השונות. יש עדין קורלציה בין הלומדים השונים (שני שלישי מהאיברים זחים) אבל עדין אנחנו מקבלים שיפור. כדי למזער את הקורלציה, נגביל את הלמידה של האלגוריתם פיצ'רים בלבד ואז הוא לא יתנהג כל כך דומה ביחס לבוטסטראפים אחרים.

שבוע III | רגולרייזציה

הרצאה

נזכיר שביעיות ריגרסיה הן בעיות עם תגובה רציפה, $\mathbb{R} = \mathcal{Y}$. נניח בנוספ' שהדומיין שלנו הוא $\mathbb{R}^d = \mathcal{X}$.

רגולרייזציה מאפשרת לנו לבחור מחלקה (רכיפה) של לומדים \mathcal{A}_λ על מחלקה היפותזות אחת \mathcal{H} .

$$\mathcal{A}_0(S) = h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathcal{F}_S(h) \text{ על היפותזה } h, \text{ כלומר } F_S(h) \text{ מינימלי}$$

דוגמה כל אלג' שעבוד לפி ERM מתאים לתבנית כזו אבל אלג' כמו NNk אינו כזה.

הגדרה תהי ℓ פ' עלות כלשהי. נגידר את הסיכון האמפירי המושרחה ע"י ℓ עברו $S = \{(x_i, y_i)\}_{i=1}^m \subseteq (\mathcal{X} \times \mathcal{Y})^m$

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h(x_i), y_i)$$

הבעיה עם אלג' ERM היא שאם \mathcal{H} גדולה מדי אז נוכל לקבל אוברפיט, כלומר לא חלנו היטב. כדי לפתור זאת, אם אנחנו לא רוצים להקטין את \mathcal{H} , נקשרו ל- \mathcal{A}_0 את הידים וכך מורידה את השונות, בתקווה שלא נקבל אוברפיט.

$$h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} (\mathcal{F}_S(h) + \lambda \mathcal{R}(h)) \text{ ע"י } \mathcal{A}_\lambda : S \rightarrow h_S$$

הערה \mathcal{R} היא פ' נפרדת שמייצגת גם מהו שאנו רוצים למוצר, לדוגמה, $\|w\|^2$ ב-SVM.

כל ש- λ יותר גדול ככל הפידלייטי (\mathcal{F}) לא מעניק את הלומד ווא נקבל תוצאות דומות על S -ים שונים, כלומר השונות נמוכה. אם λ קטן השונות תנצל.

הגדרה \mathcal{R} נקרא רכיב הרגולרייזציה.

הערה הערך של \mathcal{R} ימדד את "סיבוכיות" המחלקה, ככל שההיפותזה יותר מורכבת (דבר שיכול להביא לאוברפיט), כך \mathcal{R} גדול יותר. המוצר של $\mathcal{F}_S(h) + \lambda \mathcal{R}(h)$, כמו הרבה דברים אחרים בלמידה חישובית, הוא טריידוף. מצד אחד נרצה h שמתארת את S יותר טוב (\mathcal{F}_S קטן) ומצד שני נרצה h שלא ספציפית מדי ל- S (\mathcal{R} קטן). מטרתנו היא מציאת הנקודת המשולמת שמאזנת בין שיקולים אלה.

כל ש- λ יותר גבוהה, כך המודל יהיה יותר פשוט והbias יהיה יותר גבוהה (אבל השונות יורדת).

הערה באמצעות $(0, \infty]$ אנחנו מקבלים מחלקה אלג' לומדים $\{A_\lambda\}_{\lambda \geq 0}$.

הפידלייטי הcy פופולרי הוא סיכון אמפירי, כלומר L_0 בעיות סיווג, אך משתמש במקביליה לריגרסיה, Squared Loss, כלומר $L_S(h) = \sum_{i=1}^m (h(x_i) - y_i)^2$, Sum of Squares האמפירי הוא

למד שלושה אלג' שביסודות הם ריגרסיה לינארית, אבל עם ביטוי רגולרייזציה אחר כל אחד.

עצי ריגרסיה

למדו איך בונים עץ (באמצעות CART), איך בונים עץ מכמה עצים (באמצעות באגינגד) ועכשו הגיע הזמן למדוד איך גוזמים יער.

$$\text{ע"ז ריגרסיה הוא } \mathbb{R}^d = \bigcup_{j=1}^N B_j \text{ כאשר כל } B_j \text{ הוא קופסה מקבילה לצירים. אם הקופסה } h\text{-}i \text{ קיבלה את התווית } c_i \text{ אז הניבוא של הע"ז הוא} \\ .h(x) = \sum_{j=1}^N c_j \mathbf{1}_{B_j}(x)$$

הערה הבדל החיד בין עצי ריגרסיה לSieog הוא ש- c_i יכולים להיות מספרים סתם ולא רק ±1.

כדי לנצל הע"ז בריגרסיה, נרצה את האלג' CART (חמדן שעובר על כל חלוקה אפשרית) רק שהפעם במקומות סיכון אמפירי לפי ℓ_1 , נשתמש Sum of Squares בשביל העיקרונו החמדן. התווית שנצמיד (בלמידה ובנבואה) תהיה ממוצע התוויות בתוך הקופסה.

הסיבה שאנו נזמין את הע"ז זה כדי להוריד את השונות, ובחרה במצב שיתכן שהבחירה החמדנית הביאה אותנו למצב טוב מאשר היפותזה יותר פשוטה.

$$\text{הסיכון האמפירי על הע"ז ריגרסיה } T \text{ הוא } L_S(T) = \sum_{j=1}^N \sum_{i:x_i \in B_j} (y_i - \hat{y}_S(B_j))^2 \text{ והוא יהיה } \mathcal{F}_S \text{ שלנו.}$$

הערה נסמן ב- T_0 את הע"ז המלא שגודל באמצעות CART. נסמן T אם T הוא תת-ע"ז של T_0 , כלומר אם T מתקבל מ- T_0 "יעי" איחוד של כמה קופסאות בו.

ביטוי הרגולרייזציה שלו יהיה $|T| = |T|(\mathcal{R})$, מספר הקופסאות ב- T (N בסימונים לעיל). קל לראות ש- \mathcal{R} נותן לנו מدد טוב למורכבות הע"ז. لكن בעיית הרגולרייזציה שלנו היא

$$\min_{T \subseteq T_0} L_S(T) + \lambda |T|$$

הסיבה שאנו מזמינים על תתי-עצים של הע"ז ולא על כל העצים האפשריים שוב היא שיש מימוש יעיל למעבר על כל תתי העצים בתוכנה. פרקטית, הגיוזם, בתמורה לסייעות נוספת יחסית קטנה, נותן לנו שימור משמעותי בשונות.

שים לב כי אם הדआת מתנגדת אחרת בחלוקת שונים של המרחב, נוכל לגוזם בחלק אחד יותר כדי להיפטר מהרעש ובחלק אחר להשאיר דומה אם התוויות יותר רגועות.

עצים מקרים הם ארבעה רעונות לא טריויאליים שיחד מניבים אלג' מאוד איקוטי/יעיל - האלג' החמדן לגידול, באגינג עם דה-קורלציה ועתה הגיוזם.

ריגרסיה מודרנית

בריגרסיה לינארית העדפנו $-d \gg m$ כי אם $d \sim m$ או השונות גדולה מאוד. במקרים להסיר פיצ'רים, כי אנחנו לא יודעים, או להקטין את מחלוקת ההיפותזות שזה לא כדאי בהקשר הזה, נקשר כמובן את ידי האלג'.

כיום כל מאד לאסוף הרבה מידע על כל דבר ולכון סביר שיחיה לנו $d \gg m$. על כן קשרת הידיים הזו היא דוגמה מעולה לרגולרייזציה.

Best Subset

בריגרסיה הליינארית הקלאסית היה לנו $L_S(w) = \|w_0 1 + Xw - y\|$ כאשר עכשו אנחנו מוצאים את w_0 מהמכ"פ.

בעיית ה-Sparsity הינה $\|v\|_0 = |\{i : v_i \neq 0\}|$ כאשר v הוא Best Subset והוא $\arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d, s.t. \|w\|_0 \leq t} \|w_0 1 + Xw - y\|^2$ בהגדרה המתמטית אבל היא נותנת לנו אומדן לכמה הוקטור "פשוט".

עתה הולכה למעשה קובע לכמה פיצ'רים אפשר להתייחס - אם לכל היותר t משקלות אינן אפס, זה אומר שאנו מסתכלים בכלל רק על כל היותר t פיצ'רים. השם מגיע מכך שאנו מסתכלים על כל תת-הקבוצות של הפיצ'רים בגודל (לכל היותר) t ובוודים מה הכי טוב.

הבעיה זו היא NP-קשה כי $\|w\|_0$ לא קמור ולכן אי אפשר לפתור את זה כבעיה קמורה.

הבעיה הצמודה (לא דו-אליטית) לו הנו, היא הבעיה $\mathcal{R}(w) + \lambda \mathcal{L}_S(w)$ או $\mathcal{R}(w_0, w)$ כאשר $\mathcal{R}(w_0, w) = \arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \mathcal{L}_S(w_0, w) + \lambda \mathcal{R}(w)$.

ריגרסיות עם רגולרייזציה לפי נורמות שונות

• הבעיה $\arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \|w_0 1 + X^T w - y\|^2 + \lambda \|w\|_2^2$ נקראת Ridge, או ריגרסיה לינארית עם רגולרייזציית- ℓ_2 .

• הבעיה $\arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \|w_0 1 + X^T w - y\|^2 + \lambda \|w\|_1$ נקראת Ridge, או Lasso.

• הבעיה $\arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \|w_0 1 + X^T w - y\|^2 + \lambda \|w\|_0$ נקראת Ridge, או Best Subset.

הערה הסיבה שהזינו את Intercept w_0 היא שלא נרצה שהוא ישפיע על ביטוי הרגולרייזציה כי אין בעיה שההיסט יהיה מאוד רחוק מהראשית, זה לא מה שאנו מנסים להגביל ולהגביל את זה זה מאוד רע.

$w = \lambda \text{ כל הביעות הלומדים הנ"ל}$ הן ריגרסיה לינארית. $w = \lambda \text{ כל הביעות יתנו } 0$.

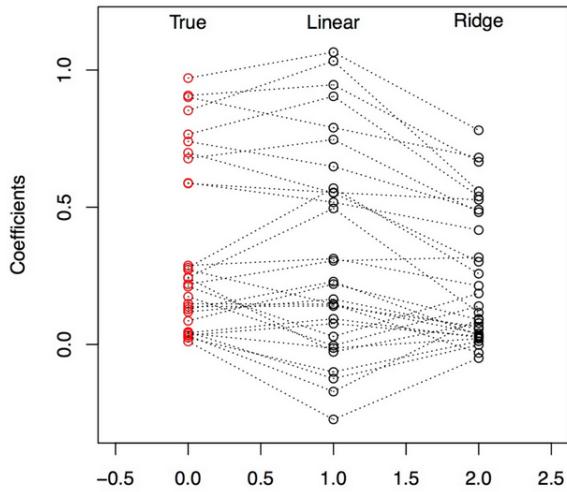
ריגרסיה Best Subset (עם רגולרייזציית ℓ_0)

כאן אנחנו מודדים מרכיבות באמצעות מספר הפיצ'רים שאנו מוגבלים את הריגרסיה אליהם - מدد הגינוי. הבעיה זו היא גם NP-קשה אבל עד $d \approx 40$ החישוב לא כל כך נורא אם חשוב דיוק מאוד גבוהה.

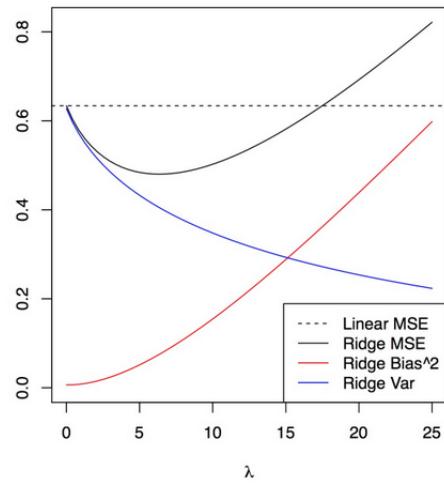
יתרונו שימושי של ריגרסיה עם רגולרייזציית ℓ_0 הוא יכולת הפרשן את הלומד - אם פיצ'ר נבחר זה אומר שהוא משפיע, אם הוא לא נבחר, הוא משפיע פחות.

Ridge

הבעיה $\arg\min_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \|w_0 1 + X^T w - y\|^2 + \lambda \|w\|_2^2$ קמורה ובפרט QP ולכן ניתן לפתור אותה ביעילות (שימוש לב לירובע על הנורמה!). במקרה הזה, הנורמה אמנים לא נראה כמו משווה שמשפיע על מרכיבות היפותזה אבל במצבות היא נותנת ביצועים טובים. ניתן לראות מהאיור הבא, בו הדגימות נדגמו מ- $y = w^T x + noise$ מעדייף משקלות יותר נמוכות על פני גבוהות.



באיזור הבא רואים את הטרידוף הטיה-שונות של ריגרסית רידג' לעומת ריגרסיה רגילה. כמובן ככל ש- λ עולה, השונות יורדת וההטיה עולה.



למעשה ניתן להראות שהוספה של קצת רגולרייזציה ℓ_2 תמיד עוזרת בהכללה.

אם עושים קצת חדו"א מגיעים לכך שהפתרון של בעיית-QP מקיים $w = (X^T X + \lambda I)^{-1} X^T y$. במקרה זה, $X^T X \succeq 0$ ועבור $\lambda > 0$ $X^T X + \lambda I > 0$ ולכן יש פתרון ייחודי נומרי (המטריצה הפיכה) לבעה! לעומת זאת, הוספה רידג' לבעיית ריגרסיה לינארית קלאסית תושיף יציבות נומרית, לא רק תתרום מתודולוגית.

טענה אם σ_i הם הערכים הסינגולריים של $\hat{w}_\lambda^{ridge} = U\Sigma^\lambda V^T y = U\Sigma^\lambda V^T X = U\Sigma^\lambda V^T X + U\Sigma^\lambda V^T \underbrace{\lambda I}_{\text{הטיה}}$ אז $\frac{\sigma_i}{\sigma_i^2 + \lambda}$

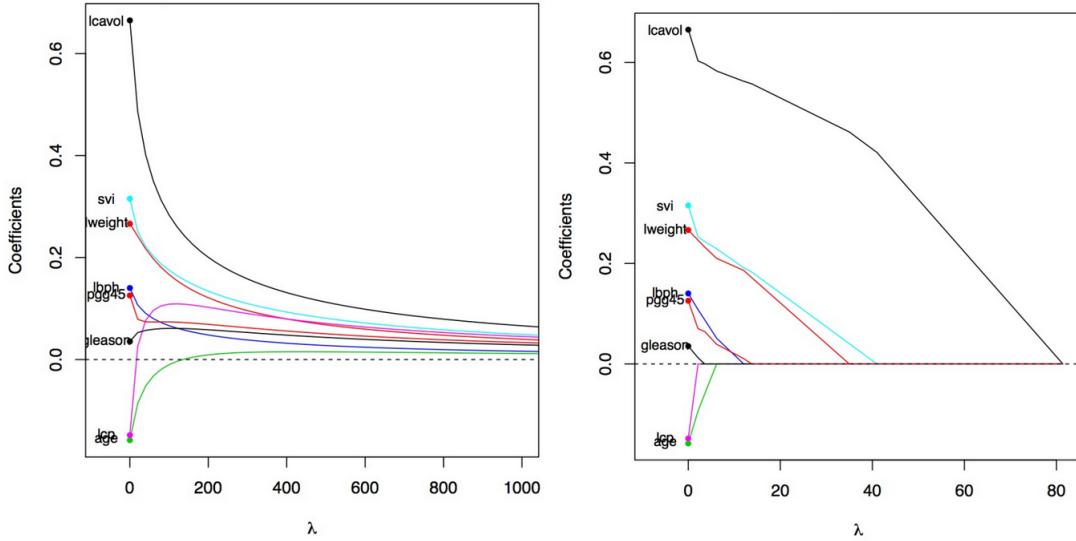
הגדרה מעקב אחר המשקלות (\hat{w}_λ) עבור ערכי λ שונים נקרא מסלול הרגולרייזציה של המודל.

הערכה מסלול רגולרייזציה מאפשר לראות איך כל פיצ'ר משפיע עבורו λ -ים שונים.

ריגרסית Lasso

בעית הריגרסיה הילינארית עם רגולרייזציה היא בעיה קמורה וזו אחת מהשיטות המקובלות לפתרון בעית ריגרסיה מודרנית.

המשkolות של \hat{w}_λ^{lasso} נוטות להיות די Sparse יחסית לridge'. באյור הבא ניתן לראות את מסלול הרגולרייזציה של ridge' מול לאסו.



לאסו מוצאת עצמה מתחקה אחר ℓ_0 ואנחנו מקבלים איזושהי תות-קבוצה של פיצ'רים משפיעים. חשוב להעיר שאפע'פ' שפיצ'רים מתאפסים, הם יכולים "להתעורר" אחרי זה ולחזור להיות רלוונטיים.

אמנם זו בעיה קמורה, אבל בغالל ש- $\|w\|_1$ לא גזיר, זו בעיה לא טריוויאלית קמורה. למעשה השפיץ של ℓ_1 הוא זו שمبיא אותנו ל-sparsity. גם ℓ_1 הוא ניתן לפרשון.

למה הוקטורים נוטים להיות Sparse? בבעיה הצמודה,

$$\underset{w_0 \in \mathbb{R}, w \in \mathbb{R}^d, s.t.}{\operatorname{argmin}} \|w_0 1 + Xw - y\|^2$$

אנחנו מוצאים את עצמנו מזערirs על פוליהידרון שהוא קוביה מוכפלת, ככלומר עם שפיצים. הערך המינימלי מתקיים בחיתוך בין הפוליהידרונים ברדיוס t לבין קו גובה כלשהו על האליפסoid שהוא הפידלייטי (Least Squares), בغالל שהפיניות של הפוליהידרון, בנויגוד לכדור ייחיד ב- ℓ_2 (שאין לו פיניות), מישרות על הציריים, באופן טיפוסי נקבל שהחיתוך הוא בפיניות ולכן חלק מהציריים מתאפסים וקיבלונו פתרון Sparse.

Orthogonal Design אם $X^T X = I$ אז זה אומר שכל הפיצ'רים אורטוני אחד לשני, זה נקרא

$$\hat{w}^{LS}, \hat{w}^{ridge} \text{ אזי אם } \eta_\lambda^{hard}(x) = x \mathbb{1}_{|x| \geq \lambda} \text{ ו } \eta_\lambda^{soft}(x) = \begin{cases} x - \lambda & x \geq \lambda \\ 0 & x \in (-\lambda, \lambda) \\ x + \lambda & -\lambda \geq x \end{cases}$$

טענה נגידר פ' סוף רכה,Subset Selection להיות הפתרונות של Least Squares, Ridge', לאסו ו-lasso אזי:

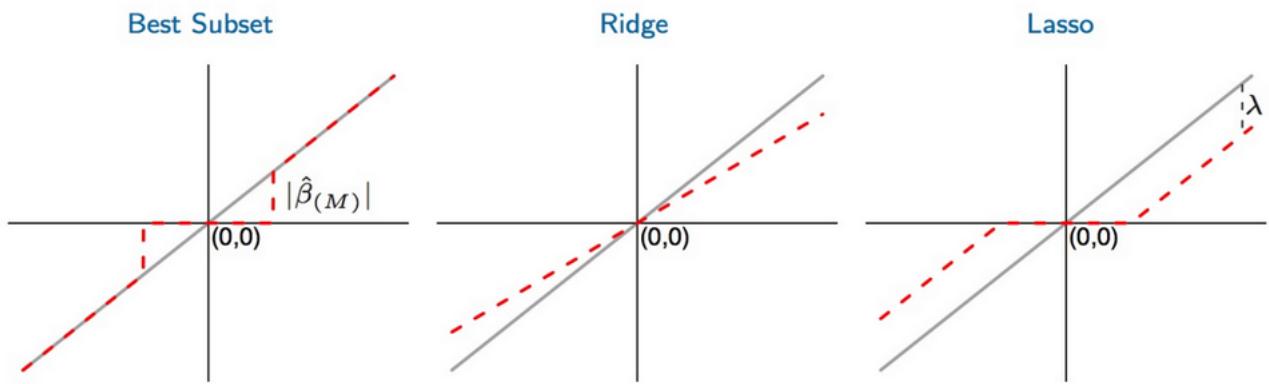
$$\hat{w}_\lambda^{ridge} = \frac{\hat{w}^{LS}}{1+\lambda} \bullet$$

$$\hat{w}_\lambda^{lasso} = \eta_\lambda^{soft}(\hat{w}^{LS}) \cdot$$

$$\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}(\hat{w}^{LS}) \cdot$$

כאשר הפעלת פ' היא על כל קורידינטה בנפרד של וקטורי המשקלות.

באיור הבא ניתן לראות כל וקטורי משקלות ביחס למשמעותם של ריגרסיה לינארית קלאסית, כדי להבין את המשמעות של כל פ' צמצום.



ריגרסיה לוגיסטית עם רגולרייזציה- ℓ_1

נזכיר כי ברגרסיה לינארית, הפידליטי הוא

$$\mathcal{F}_S(w) = \sum_{i=1}^m \log \left(1 + e^{w_0 + \langle x_i | w \rangle} \right) - y_i (w_0 + \langle x_i | w \rangle)$$

ואז עם רגולרייזציה הלומד פותר את הבעה

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^m \log \left(1 + e^{w_0 + \langle x_i | w \rangle} \right) - y_i (w_0 + \langle x_i | w \rangle) + \lambda \|w\|_1$$

הבעיה כאן היא שזה לא גזיר אבל כאמור אפשר להשתמש בעיות קמורות, וספקטיבית לבעיה זו יש פוטרים מאוד ייעילים.

הערה חיבור טובה לומד זה ושאר הרגולרייזציות וניתוח שלhn היא `glmnet` שמקומפלט ב-`fortran` והוא מאוד יעיל, ויש לה גם חיבור לפיתון.

בחירה ושערוך מודלים

בחינתן בעיה, באיזה אלג' נשתמש? ובעור כל לומד אילו היפר-פרמטרים נבחר בשביילו?

לכוארה אפשר לבחור את המודל עם ה-`Training Error` הכי נמוך (כי הרי אין לנו את ה-`Test Data`). זה רעיון גורע כי זה מוביל לאוברפיט וגם מגדיל את מחלוקת ההיפותזות מאוד (אם מאמנים הרבה מודלים) ואז השונות גבוהה גם כן.

הדבר האידאלי הוא להמציא דאטא חדש ולבדוק עליו. אז אפשר לחתה מה-Training Data ולבדוק עליו. הבעה עם זה היא שזה מקטין את כמות הדאטא לאימון שזה מוריד ביצועים.

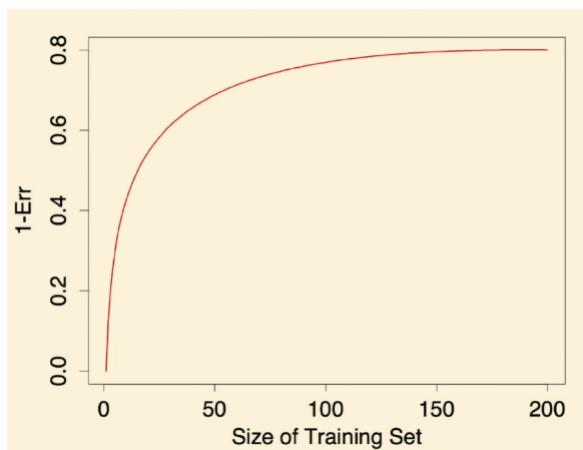
k-fold Cross Validation

נחלק את הדאטא ל- k חלקים ונירץ k איטרציות שבה בכל פעם נוציא $\frac{1}{k}$ מהדאטא ונאמן על שאר הדאטא את המודלים השונים ונחשב את Test Error כשהדאטא לבחן הוא מה ששמרנו בצד. נמצא את השגיאות ונבחר את המודל עם ממוצע שגיאות טוב (קטן) ביותר. כך אנו מדמים המצאה (נאיבית ביוטר) של נתונים חדשים.

הערה הממוצע הוא כדי להפחית ביאס - קרי להתעלם מהרעש.

חסרונות של CV

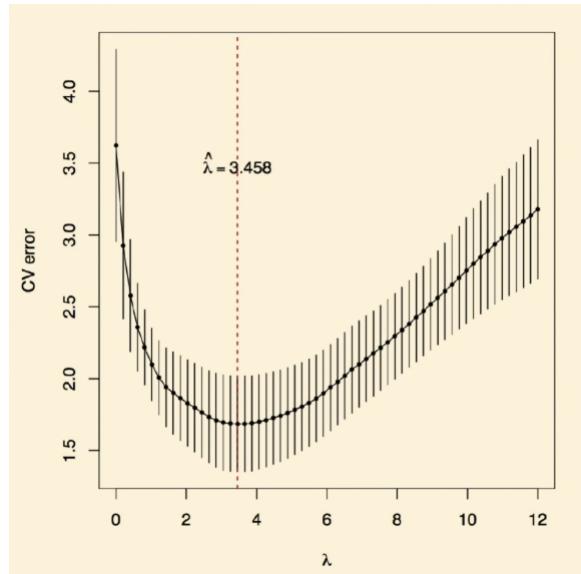
- הממוצע האמפירי מנסה לדמות שגיאת ההכללה שהיא תוחלת על ההסת' לטעות בהינתן התפלגות \mathcal{D} כלשהי. לכן אם קיבלנו מוגם לא מייצג עדין יש סיכון לאובר-פיט (כי אנחנו חושבים ש- \mathcal{D} הוא משווה שהוא לא).
- ההרצתה של כל האלג' מספר פעמיים דורשת כוח חישובי גדול כ- k -fold (לא קטן מאוד).
- הביבוצים לפי CV נחותים ביחס למציאות כי הוא לא מאומן על כל הנתונים - יכול להיות שאנו נוטנים הערכה פסימית סתם.



באյור הנ"ל ניכר כי התרומה של עוד דאטא פוחתת ככל שיש יותר ממנו (עבור בעיה כלשהי). לכן, אם $m = 200$ בדוגמה זו ונשתחמש ב-CV-5-fold נקבל הערכה די טובה כי אנחנו לומדים על 160 דוגמאות שנראה שהשגיאה די דומה עבورو ועבורו 200.

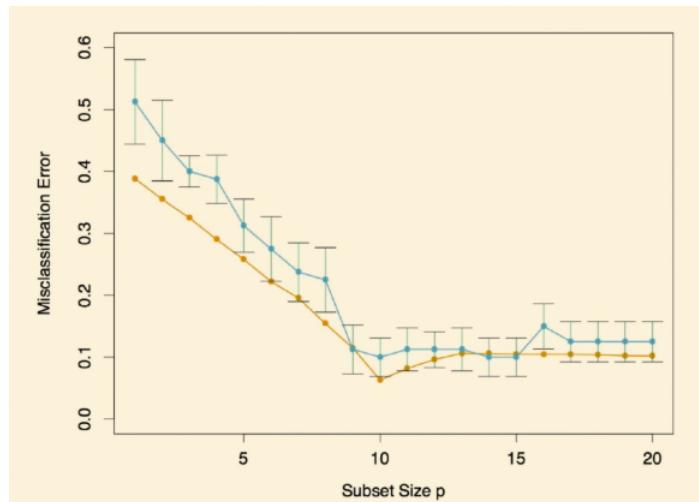
לעומת זאת, אם $m = 50$ מאד לא מצליח כי הבדל הביבוצים עבור $m = 40$ לעומת $m = 50$ הוא מאד דרמטי, והוא אנחנו פסימים מאד ביחד למה שייהי במודל הסופי.

נותן גם k-fold CV אוטומטי, והgraf הבא מציג את תוצאתו בבחירה לבריגסית Lasso עבור בעיה כלשהי.



הסטיית התקן לא מעניינת אותנו יותר מדי כשה מגיע לבחירת מודל, אבל בהערכת מודל זה כן חשוב, כי זה מראה לנו על השונות של המודל.

הערה גם אחרי שאחנו בוחרים מודל עם CV, מאמנים שוב את המודל על כל הדאטא שיש לנו ואוותו מייצאים. כאן, שגיאת ההכללה בכתום ו-CV-fold 10 בכחול, וניכר כי ההערכתה שלנו היא די טובה (ספקטיבית כאן הדגימות דהיינו "שפויות").



בחירה של k היא גם מאוד חשובה. $k = 1$ משמעו לא עשינו דבר ; $k = 2$ נקרא split-sample-CV והוא פשוט חלוקה לשתיים של הדאטא ; k קטן מעלה את הביאס כי-h-Test Error לא מייצג וקטן ; k גדול מדי אומר שה-training Data מאוד דומה בכל האיטרציות ; נקרא $k = m$ leave-one-out CV והוא אומר שבודקים כל פעם על דוגמה אחת.

ככל ש- k יותר גודל ככה החישוב יותר יקר, וככל אצבע בוחרים מספר יחסית נמוך כמו 5 או 20.

בשביל בחירת מודלים Bootstrap

נדגום החוצה דגימות עם החזרה ונשותמש בהן כ-*Test Error*. ככל שנעשה את זה יותר פעמים נקבל הערכה יותר טובה של \mathcal{D} (שאנו מודמיינים שקיים) מאותן הסיבות כמו Bagging. כך בדומה ל-*k-fold CV* נבצע בחירת מודלים וגם הערכות.

ב-*Bootstrap* יש קורלציה גבוהה בין הנתונים ובנוסף מדיסקרטיבית שליש מהדאטא לעולם לא ידגם שזה פער גדול אל מול כל הדאטא ולכן קשה לנו להתקרב אליו לשגיאת החכללה.

מלבד הפסיכיות שהזכרנו לעיל, יש בעיה חמורה יותר של אופטימיות יתר שרבים שוגים בה. במהלך אימון המודל אנחנו מנקים את הנתונים, מנרמלים וכל מיני שינויים אחרים. השינויים אלה לא חלים על דאטא חדש שמניעו. כדי לטפל בזה, ראוי לעשות את הדברים הבאים:

- לשים חלק מהדאטא לצד ולא לגעת בו עד שיש מודל מוכן כדי לראות שאנו המודל מבצע על נתונים שלא עיבדנו.
- לבצע pre-processing רק על חלק קטן מהנתונים ולהיות מודעים לזה שאנו מודע אופטימיים לבבו.
- לכתוב בקוד את תהליך ה-pre-processing ולהריץ אותו בכל לולאט CV על הדאטא (גם האימון וגם הבדיקה).
- להפעיל את ה-pre-processing גם על נתונים חדשים שאנו מקבלים בפודקשן.

תרגול

שאלות חזרה

1. איזו מהטענות הבאות אינה נכונה?

- (א) החזקה השלישי של פ' קמורה אי-שלילית היא קמורה.
 (ב) איחוד קבועות קמורות הינו קבועה קמורה.
 (ג) ריבוע של פ' קמורה הוא פ' קמורה.
 (ד) פ' העולות MSE היא קמורה על משקלות הריגרסיה הלינארית.

תשובה ד' נכון כי ידוע לנו שהיא קמורה ואפילו שיש לה מינימום יחיד. א' נכון כי זוג הרכבה של פ' קמורה ופ' קמורה ומונוטונית עולה (חזקת השלישי על ערכים אי-שליליים היא קמורה ומונוטונית עולה). ג' גם נכון מאותה הסיבה ולכן לא נכון.

ב' לא נכון כי עבור $y \in S_1 = \{x\}, S_2 = \{y\}$ שון קמורות באופן ריק, האיחוד שלן כמובן אינו קמור (עבור $y \neq x$) כי $\frac{x+y}{2} \notin S_1 \cup S_2$ לדוגמה.

2. ברוצנו למדוד את הדאטאסט הבא (ראו טבלה) $S = \{(x_i, y_i)\}_{i=1}^m$ ולמדו כפולינום עם שגיאות ℓ_2 . בחרנו להשתמש ברגression לינארית כדי למצוא את הפולינום, יהיה מדרגה 3 לפחות. מה תהיה מטריצת הדגימות שלנו $X \in \mathbb{R}^{m \times d}$ (אל תשכחו פולינומיים מדרגה 0)?

x_i	2-	1	3	4	6
y_i	5	7	1-	6	6

$$\text{תשובה} \quad X = \begin{pmatrix} \vdots & & & \vdots \\ x^0 & \dots & x^3 \\ \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{m \times 4}$$

3. איזו מהטונות הבאות אינה נכונה בהקשר של בעיית ריגרסיה לינארית $w = ?Xy$?

(א) אלג'-ה-*Least Squares* פותר את בעיית הריגרסיה הליינארית בעזרת עיקרונו-*ERM* על מחלוקת ההיפותזות

$$\mathcal{H} = \{x \mapsto \langle w | x \rangle + b\}$$

(ב) אם מספר הפיצ'רים גדול בהרבה ממספר הדוגמאות יש סבירות גבוהה ל-*Overfit*.

(ג) אם $X^T X$ הפיכה אז למשוואות הנורמליות, $y = X^T Xw$ יש פתרון יחיד.

(ד) אם $X^T X$ לא הפיכה אז אין למשוואות הנורמליות פתרון.

תשובה א' נכון כי זה בדיקות מה שהראנו בתרגול. למעשה את אותו הפתרון של Least Squares פותר גם לומד MLE כאשר הנחנו

$$\text{שחרוש מתפלג } (\hat{w}, 0, \sigma^2 I_m).$$

4. נניח שאנו רוצים ללמידה מודל ריגרסיה לינארית על דאטאסט כלשהו. אחרי שאיםנו אותו שמו לב שמו פיצ'ר אחד פעמיים. במקומות להסיר את הפיצ'ר ולאמן את המודל מחדש, החלטנו להתעלם מהפיצ'ר הכפול ולהוסיף את המשקלות המתאימה לו למשקלת של הפיצ'ר שנשאר. נסמן \hat{w}_a את המודל הערוך וב- \hat{w} שהוא מתקיים אילו היו מאמנים בלי הפיצ'ר הכפול.

איזו מהטונות הבאות נכונה?

(א) \hat{w}_a ו- \hat{w} יש מקדים שונים והם נתונים ניבואים שונים על דוגמה חדשה.

(ב) \hat{w}_a ו- \hat{w} אולי יש מקדים שונים ובמקרה זה הם יתנו ניבוא שונה.

(ג) \hat{w}_a ו- \hat{w} אולי יש מקדים שונים אבל הם נתונים ניבוא זהה על דוגמה חדשה.

(ד) \hat{w}_a ו- \hat{w} אולי יש את אותם המקדים הם נתונים ניבוא זהה על דוגמה חדשה.

תשובה הסירה של פיצ'ר משמעה הסרת העמודה X . אנחנו ממזערים MSE, כלומר אם האינדקסים של הפיצ'רים זהים הם α, β

א'

$$\begin{aligned} \sum_i (Xw - y)^2 &= \sum_i \left(\sum_{j \neq \alpha, \beta} x_{ij}w_j + x_\alpha w_\alpha + x_\beta w_\beta \right)^2 \\ &= \sum_i \left(\sum_{j \neq \alpha, \beta} x_{ij}w_j + x_\alpha (w_\alpha + w_\beta) \right)^2 \end{aligned}$$

לכן סכימה של המקדים שקופה לחילוטין להסרת העמודה, ככלומר ד' היא התשובה הנכונה. אינטואיטיבית, בגלל שריגרסיה לינארית עשויה בסה"כ הטלה אורותוג' על $(\text{Im } X, \text{sp } X)$ לא השתנה בהוספת עמודה זהה, הניבוא יהיה זהה.

5. יש לנו דאטאסט לסיוג y , X ולמדנו אותו עם Soft-SVM וקיבלו $\hat{w} = (3, -4, 5)^T$, $\hat{b} = 5.5$. קיבלנו דוגמה חדשה או חלק מהמדגם כדי זהה יהיה מוגדר היטב אבל לא קרייטי, $x = (2, -6, -7)^T$, $y = 1$. ונסמן ב- ξ את המשתנה Slack של הדוגמיה, ככלומר זה שמקיים את הדרישה $\xi - \|w\|_1 \geq 1$.

נכונות?

(א) המודל שלנו מסוווג נכון את הדגימה.

(ב) המודל שלנו מסוווג את הדגימה לא נכון.

(ג) $\xi = 0$.

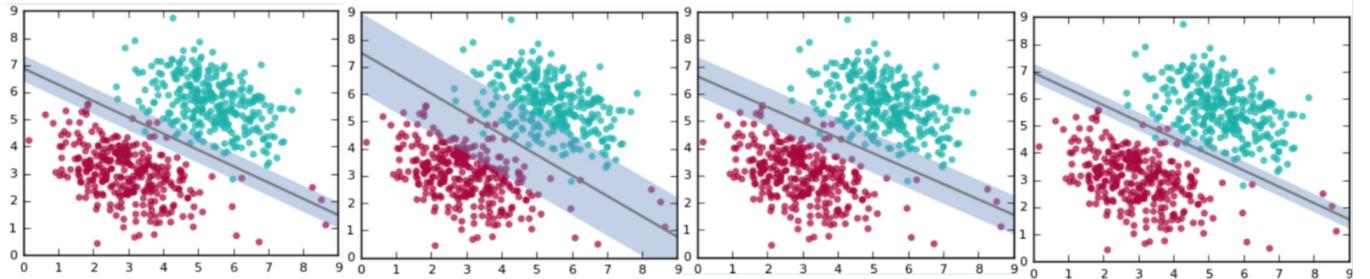
(ד) $\xi < 0$.

תשובה ראשית אנחנו מסוווגים נכון כי

$$\hat{y}(x) = \text{sign}(\langle \hat{w}^T | x \rangle + b) = \text{sign}(3 \cdot 2 + (-4)(-6) + 5 \cdot (-7) + 5.5) = \text{sign}(0.5) = 1$$

לכן א' נכון וב' לא נכון. ה- ξ הוא מה שחסיר כדי שהדגימה לא רק תצדוק, אלא גם תהיה מחוץ למרג'ין (שמדובר ע"י הנקודה הכי רחוקה שישווגנו לא נכון). לכן במקורה הזאת $y \geq w^T x + b - 0.5 = 0.5 \geq 1 - 0.5 = 0.5$. חשוב לציין שב-SVM חשוב לנו שהמדגם האימונו לא רק יהיה בצד הנכון של העל-מישור, אלא גם יהיה בצד הנכון (מחוץ ל-)מרג'ין (באמצעות ξ -ים שמתקנים). לכן גם ג' וגם ד' אינם נכוןים.

6. נניח שפתרנו את בעיית האופטימיזציה של Soft-SVM $\arg\min_{w,b} \|w\|^2 + \frac{C}{m} \sum_i \ell_{hinge}(y_i \langle w | x_i \rangle)$ עם ערכי C שונים.Soft-SVM מאריך מרג'ין מאשר טוויות סיווג. את $C = 0.01, 0.1, 1, 10$ לגרפים של המודלים שהći סביר שהשתמשו בהם.



תשובה C הוא פרמטר הרולגייזציה, ובמקרה הזה ככל שהוא יותר גדול כך יותר חשוב לו המרחק מהמרג'ין מאשר טוויות סיווג.

לכן בהתאם לאינטואיציה זו, המרג'ין יהיה צר יותר ככל ש- C גדול יותר. מכאן קל להתאים:

C	0.01	0.1	1	10
גרף	2	3	1	4

7. תהי \mathcal{H} מחלקה היפותזות על \mathbb{R}^2 כך ש- $h(x) = \mathbb{1}_S(x)$ כאשר S היא איחוד של שני מלבנים המקבילים לצירים. איזו מהטענות הבאות נכון?

(א) $\text{VCdim}(\mathcal{H}) = 2$

(ב) $\text{VCdim}(\mathcal{H}) = 4$

(ג) $\text{VCdim}(\mathcal{H}) = 6$

(ד) $\infty > \text{VCdim}(\mathcal{H}) \geq 8$

$$\text{VCdim}(\mathcal{H}) = \infty$$

תשובה אינטואיטיבית, מלבד זה שני אינטואולים וראינו שאיטרול זה מミיד 2. אך זה כנראה לא 2 וגם לא 4 כי יש לנו לפחות שני מבנים לכך זה לא א' ולא ב'. בנוסף, בגלל שהוא למעשה מקרה פרטי של עצים וראינו שעצם זה למיד עם ERM, הרי שמייד ה-VC הוא סופי ולכן זה לא ה'. אינטואיטיבית (בגלל שלא צריך להסביר את הפתרון שלנו) אפשר לנתח קבוצה בגודל 7 ול.then התשובה היא ד'.

שבוע 2 | למידה לא מפוקחת

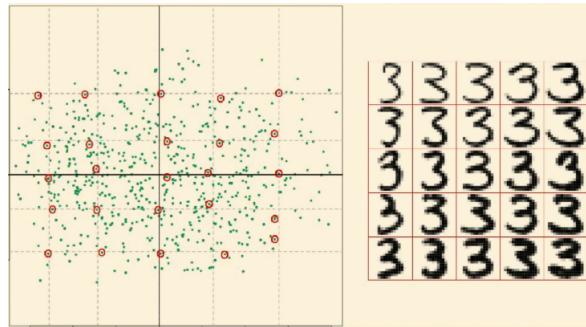
הרצאה

עד עכשיו רק על למידת batch מפוקחת, כשיש לנו א' ו-ב' ואנו מוחשים כלל החלטה. עתה אין לנו לנו לייבלים בכלל, כלומר הדאטה הנתון לנו הוא $\{x_i\}_{i=1}^m$.

דוגמא נניח שיש לנו דאטה ב- \mathbb{R}^d כאשר d מאד גדול, לדוגמה תמונה 1024×1024 . עם זאת, ידוע לנו שלמעשה הדאטה האמתי הוא כמעט הרבה יותר נמוך. לדוגמה להבדיל בין חתול וכבל זה רק כמה עשרות פיצ'רים (האף וכו'). הורדת d היא משמעותית כי זה מוריד זמן אימון; מוגר אוברפיט ווד.

דוגמא נתונות תמונות 1024×1024 של בן אדם מסוימים שונים ואנו רוצים לדעת לאיזה כיוון אנחנו מסתכלים. במקום $\sim d$, המימד האמתי הוא הרבה יותר נמוך כי כל מה שאנו רוצים את זווית ההסתכלות. 1,000,000

דוגמא נסתכל על תמונות של הספרה 3 בעל גרען דו ממד. אם נתאר כל אйור של הספרה בתווך סיבוב וכיום של אירור מקורי של הספרה, נוכל להציג את כל התמונות במקום בגרף ממימד 28×28 לגרף דו ממדי.



המשימה הזו נקראת הפחיתת מימד (dimension reduction). אנחנו מוצאים העתקה $W(x)$ שמעתיקה $x_i \in \mathbb{R}^d$ כאשר d גבוהה ל- d' כאשר d' נמוך. שיטה זו עוזרת להבנת המבנה של הדאטה, ויזואליזציה, פרוי-פרוססינג ועוד.

Clustering

דוגמה נניח שיש לנו דאטא של תכונות של ספרות ואנחנו רוצים לדעת כמה ספרות יש לנו - בלי תוויות. לחולופין קיבלנו אוסף תכונות של פנים של בני אדם ונרצה לדעת כמה אנשים שונים יש לנו.

כדי לבצע את המשימה זו, נוכל להוריד מימד באמצעות הקטנת המימד שעליה בדיק דיברנו ואז לבצע קלאסטרינג על מימד הרבה יותר נמוך.

המטרה המוצחרת של קלאסטרינג היא לחלק את הדאטא לצברים כך שכל צבר מכיל דגימות עם משווה, בעוד אין לנו ליibiliים.

Anomaly Detection

נרצה בהינתן דגימות זהות מה היא התנהגות נורמלית ומה הוא חריג.

דוגמה בהינתן סטוריים של פתיחת גללים של מטוס, נרצה להדיל נורה אדומה אם לדעתנו משווה השתבש בתהליך הפתיחה.

דוגמה זיהוי מתקפות סייבר, שגיאות תוכנה, זיהוי הונאה בקריטיסי אשראי וכו'.

הערה אנחנו לא צריכים להגיד מה לא עובד, אלא רק שזה לא עובד.

Principal Component Analysis

פורמלית, בעיית הורדת המימד היא בהינתן $\mathbb{R}^d \rightarrow \mathbb{R}^k$ שבו $x_1, \dots, x_m \in \mathbb{R}^d$ שהוא מבחן האימון שלנו, נרצה $d < k$ שבערו כל דבר שאפשר לעשות עם x_i אפשר לעשות גם עם (x_i) (בנפנוף ידים).

הערה אם W לינארית זה נקרא הקטנת מימד לינארית ואם לא אז הקטנת מימד לא לינארית. אנחנו עוסקים רק בהקטנות לינאריות.

הערה כאמור היתרונות של הקטנת מימד כוללים בין היתר את העובדה שרוב האלגוריתם שלנו עובדים הרבה יותר טוב עם $m > d$ (מלבד k NN, שלא נכון לו).

הערה הטרנס' עלולה לגרום לאובדן משמעות של הפיצ'רים, למשל אין לנו גובה, משקל וכו' אלא איזשהו איחוי בין כל מיני פיצ'רים.

ניח עתה שהדאטא $\{x_i\} \in \mathbb{R}^d$ לא נמצא בתוך כל \mathbb{R}^d אלא רק "מ" בגודל k של \mathbb{R} (או משווה מאוד קרוב לזה). נוכל בקלות להוריד את המימד באופן הבא: נמצא בסיס אורותוני "لت" מ- \mathbb{R}^d והקטנת המימד תהיה לקטור הקוורדיינטות של כל דגימה לפי הבסיס הזה.

הבעיה היא שאנו צריכים למצוא את הת"מ הזה, שם כבר הוכחנו.

נבחר k כלשהו ונחש $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$ ובאותו הזמן $U : \mathbb{R}^d \rightarrow \mathbb{R}^d$ שמהווה "הופכית" ל- W (מחזירה את הדגימות במימד הנמוך למימד הגבוה). הסיבה להטעסקות זו היא שהפעלת $(W(x_i))$ למשה בודקת כמה הרחקנו את x_i מהמקום המקורי שלו, כאשר אם x_i בת"מ זה יהיה 0 ואחרת לא, כי נctrיך להזיז את הנקודה כדי שתתאים לת"מ.

כדי למדוד את איקות הביצועים שלנו, נזעיר RSS, פורמלית הבעיה שאותה PCA פוטר היא בהינתן $\sum_{i=1}^m \|x_i - U(W(x_i))\|^2$ ולכן פורמלית הבעיה שאותה PCA פוטר היא בהינתן $\{x_i\} \in \mathbb{R}^d$

$$\operatorname{argmin}_{W \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}} \sum_{i=1}^m \|x_i - UWx_i\|^2$$

משפט (הפתרון לביעית PCA נסמן $A = \sum_{i=1}^m x_i x_i^T$). זהה מטריצה ריבועית, סימטרית ו-PSD והיא נקראת S .
תהי $U \in \mathbb{R}^{d \times k}$ המטריצה שעמודותיה הן u_1, u_2, \dots, u_k , הוי $W = U^T$. אזי W פותרת את בעיית ה-PCA עבור $\{x_i\}$.

הערה לכsoon של כל המטריצה הוא מיותר ביחס אם אנחנו רק צריכים את k ו"ע שהוא נמוך משמעותית ממספר ה"ע. נרצה רק לחלק את ה"ע ובאופן יציב נומריה (יש אלג' כאה).

הוכחה: UW היא מטריצה $d \times d$ מדרגה k ולכן $S = \text{Im } UW$ הוא מミיד לכל היותר k . לכן $x \mapsto UWx$ היא העתקה $\mathbb{R}^d \rightarrow \mathbb{R}^d$. לכן $\|x_i - UWx_i\|^2$ מינימלי אם x מקיים ש- UW הינה הטלה אורתוגונלית (הטלה בלינארית, הינה הדבר כי קרובה על הת"ם לכל וקטור). לכן נristol (בלינארית 2) $UW = VV^T$ כאשר V היא מטריצה שעמודותיה הן בסיס אורתוגונלי של S ולכן בה"כ מתקיים $V = U^T$ ועמודות U הן אורתוגונליות.

מתקיים

$$\begin{aligned} \|x - UU^T x\|^2 &= \|x\|^2 - 2x^T UU^T x + x^T UU^T UU^T x \\ &= \|x\|^2 - x^T UU^T x \\ &= \|x\|^2 - \text{tr}(U^T x x^T U) \end{aligned}$$

ולכן בעיית ה-PCA שcola לביעיה $\underset{U \in \mathbb{R}^{d \times k}: U^T U = I}{\text{argmax}} \text{tr} \left(U^T \left(\sum_{i=1}^m x_i x_i^T \right) U \right)$

כולם אנחנו מנסים למקסם את סכום איברי האלכסון של $U^T AU$.
כיצד נמקסם את $\frac{x^T Ax}{\|x\|^2}$ כאשר x וקטור ייחידה? הפתרון הוא בעצם ה"ע המובייל (A לסכינה) כי הוא זה שמקסם את $x^T Ax$ שבמקרה הזה הוא הנורמה כפול ה"ע (הגדל ביחס).

נסמן $D = VDV^T$ הפרקטרלי של A כאשר D אלכסוני עם ע"ע בסדר יורד ו- V עם עמודות ו"ע מתאימים. ניתן להוכיח כי השווין הזה מתקיים כאשר U מכיל את k ה"ע המוביילים של A (זהו הערך של הביטוי מימין) ולכן כמובן ש- U כזה מקסם את הביטוי המקורי. U ו- W כפי שהגדכנו אותם בניסוח המשפט בדוק מקיימים את התיאור הזה ומכאן נובע שהם הפתרון לבעיה. ■

הערה ה"ע המוביילים של A הם אלו שקובעים לאילו כיוונים אנחנו מכוצחים את המרחב לת"ם המוצומצם.

אין סיבה ש- W תהיה רק לינארית ולא אפינית, ככלומר נחשף $(x - \mu) \in \mathbb{R}^d \rightarrow \mathbb{R}^{k-1}$ כאשר $W(x) = \tilde{W}(x - \mu)$ הינה העתקה לינארית. כך הת"ם לא חייב לעבור דרך הראשית (הגדודתית הוא כן, אבל אנחנו מעתיקים למקום שהוא ת"ם מזוז). באופן כללי, מחייב $\bar{x} = \sum_{i=1}^m x_i$

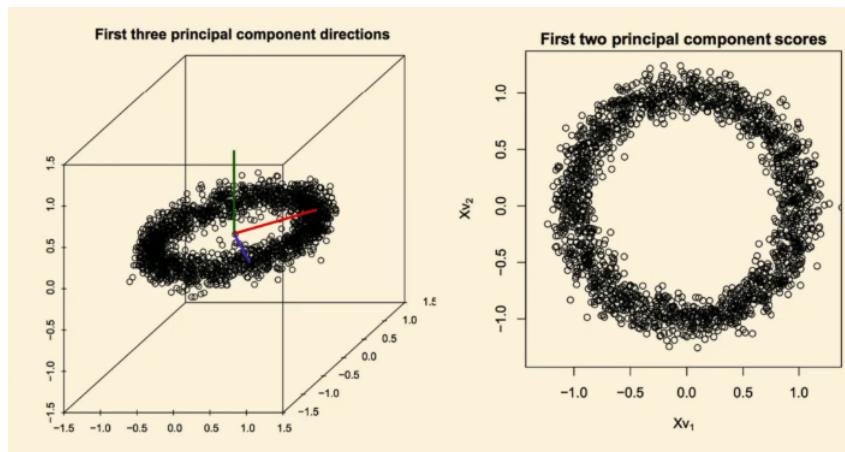
לכן, פתרון ה-PCA עם האינטראספט מתקבל על ידי לכsoon/ו"ע של $A = \sum_{i=1}^m x_i x_i^T$ במקומות $(x_i - \bar{x})(x_i - \bar{x})^T$ של $S = \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^T$ אשר להסביר את הקשר בין PCA למטריצת ה-Covariance. Sample Covariance-הו"ע שבעורם השונות של הדadata היא מקסימלית.

הגדרה תהיה $S = \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^T$ מטריצת ה-Covariance של $\{x_i\}$. יהיו u_1, \dots, u_d המובילים המתאים לע"ע $\lambda_1 \geq \dots \geq \lambda_d \geq 0$ נקרא u_i Principal Vector של x_i .

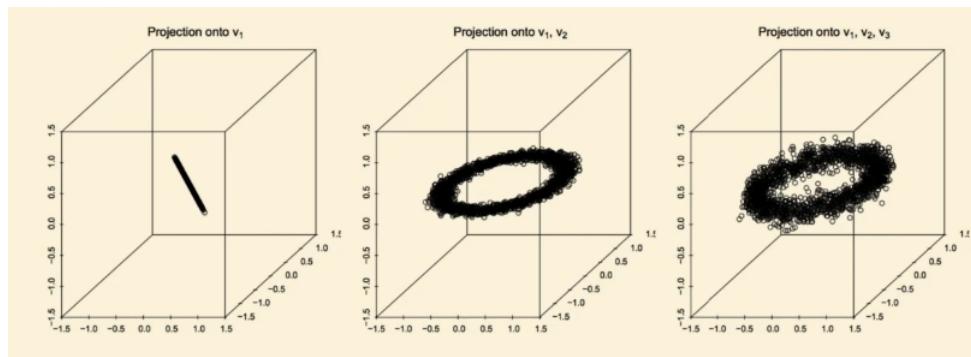
הגדרה יהיו k הוודת מימד מסווג PCA מעתקה $U^T x_i$ כאשר עמודות U של $\{x_i\}$ Principal Vectors.

דוגמאות ל-PCA

דוגמא נניח $d = 3$ ו- $k = 2$. לא כל כך מעניין אותנו ה-"גובה", בتوزוק הביגול והאלג' גם בין את זה, כפי שניתן לראות מההעתקה שעוברת הע"ע המתאים לוקטור הירוק קטן מהאחרים ולכן הוא לא ייחשב. Principal Subspace-הו"ע $\{v_1, v_2\} \subseteq \mathbb{R}^k$ נקרא $\text{sp}\{\text{Intrinsic Subspace}$ (מה שרואים מימין (מה שרואים נקרא ה-Subspace).



נסתכל על הפחתת מימד לממדים שונים, ניכר כי אם k קטן מדי אנחנו מאבדים משמעות לגמרי בהטלה לת"מ (הלא היה WU).



דוגמא נחזור לתמונות 50×50 של פרצופים של בני אדם



אין יותר מדי שוני מבחן התוכן כי כולם גברים, כולם מסתכלים למטה ואנחנו לא צריכים כל כך הרבה דרגות חופש. עם $k = 10$ מקבלים את התוצאות הבאות לאחר הפעלת היטלה (כאן אנחנו רואים את UW כולם אחורי שהחזרנו לו- d , כאמור הטלה אורתוגונלית).



כולם אנחנו מצלחים לשמור היטב את התמונות ועכשו לעובוד עם הדאטא זה הרבה יותר קל.

הערה ביצענו סוג של דחיסה, במקום לשמר כל תמונה צריך לשומר רק את ה-“ U ” ואת הפיצ’רים המופחתים.

תרגיל הסטודנטית המשקיעה תראה שהדאטא יושב על ת-“ V ” מימיד k של \mathbb{R}^d אם “ S ” הוא בדרגה לכל יותר k . במקרה כזה, היא תבין כיצד למצוא בסיס אורתוגונלי V באמצעות PCA?

הערה למעשה גם ה-“ U ” הם וקטורים ב- \mathbb{R}^d ולכן אפשר להסתכל עליהם. עד כה ראיינו את x_i וה- UWx_i הווקטורים כתמונות ב- \mathbb{R}^d אבל אפשר גם לראות איך התמונות של x_i נראים. בגלל שהם פורשים את הת-“ V ” הפרינסיפאלי, נצפה שהם יראו כמו פנים “טיפוסיות” בהתבסס על מה שראינו עד כה.



משמעות לשים לב שהשחור והלבן משלימים אחד את השני כי הוקטוריהם האלה חייבים להיות ניצבים אחד לשני.

מחינת חישוב מעשי של PCA צריך לכסן מטריצה $\mathbb{R}^{d \times d}$. כאשר $m \gg d$ האלג' הבא מחשב PCA ב- $(m^2 d)$ וכלל אנחנו יכולים לחשב PCA ב- $(small^2 \cdot large)$.

פסאודו-קוד לאלג' לפתרון PCA

$m > d$ אם •

- נסמן $A = X^T X$ והוא u_1, \dots, u_n המובילים של A .

$m \leq d$ אם •

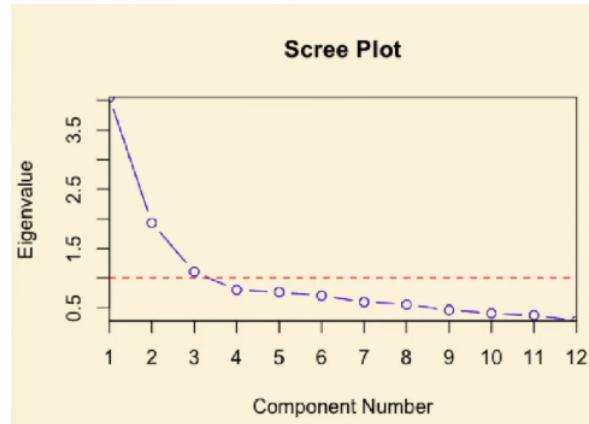
- נסמן $B = X X^T$ והוא v_1, \dots, v_n המובילים של B .

- לכל i נגדיר $u_i = \frac{1}{\|X^T v_i\|} X^T v_i$

• נחזיר u_1, \dots, u_n .

איך נבחר את k ? אם k נזק נקבל השחתה של הנתונים ושגיאה גבוהה. אם k גדול מדי לא נקבל שיפור אל מול k "האינטרינזי" ולכן נכוון k -האופטימלי שאחריו אין שיפור ממשמעותי. אם הדאטה יושב בדיק על ת"מ ממימד k , אז לפחות היותר עד k השגיאה תהיה 0.

ה策ורה המקובלת לקביעת k היא לחשב את ה-Sample Covariance ולקציר את הע"ע על גוף ולבחור את הנקודה של דעתנו אחוריה הע"ע מפסיקים קטנים (משמעותית), ראו איור (שנקרא Scree Plot, כאשר Scree Plot משמעו מדרון עם אבני קטנות).



נשים לב כי יש קשר אדווק בין גודל הע"ע לבין השגיאה שכן ע"ע נזק מאוד גם אם מתעלמים ממנו לא משפייע מאוד (זהו גם עיקנון הדחיסה של JPEG, לפחות בעבר, לא קשור בכלל).

הסתכלות על ה-Scree Plot ובחירה k האופטימלי היא תחילה אנושי ולא אלג' פורמלי אבל בימינו יש כל מיני משפטים וחסמים בהתבסס על גוף כזה.

תרגיל הסטודנטית המשקיעה תבחר k , תייצר $x_1, \dots, x_m \in \mathbb{R}^d$ שלמעשה כולם על ת"מ (פרינסיפיאלי) מממד k . דרך אחת לעשות זאת היא לאפס כמה קוודינטות ואז לסובב (באמצעות מטריצה אורתוג')

כדי לג'רט X לפניו סיבוב, נגריל מטריצה $k \times m$ שמתפלגת נורמלית (כל ערך), נצמיד לה מטריצת אפסית $(d - k) \times m$ כדי לקבל X מממד $m \times d$.

כדי להגריל מטריצת סיבוב נרגיל מטריצת סיבוב באופן אחד באמצעות הרצת פירוק QR על מטריצה מתפלגת נורמלית. ולקיחת Q . לsicום ניתן ל-PCA את X כ- $\{x_i\}$. נחקרו את ה-Scree Plot. עתה נוסיף רעש נורמלי Z ועם דרגת רעש σ ניתן ל-PCA את $X + \sigma Z$. עתה אחרי k השגיאה לא מתאפסת אבל עדין יש ירידה בשגיאה.

Clustering

בහינתן x_1, \dots, x_m , נרצה לחלק אותם ל- k קבוצות, או צברים. לעיתים ברור לנו כמה k אמור להיות אבל לעיתים לא. למה זו בעיה מעניינת? כדי לחקור את הדטא שלנו, מעוניין לראות אילו מאפיינים מסווגים חלקיים בדטא וממה מאפיין אותו. לחלופין, אם אנחנו רוצים לאמת חלוקה שימושה נתן לנו ולברר האם יש קשר בין הגאותייה של הבעה לבין המהות של הדגימות. בغالל שאין לנו ליבלים, אנחנו לא יודעים מה באמת נכון. ככלומר למעשה הבעה בכלל לא מוגדרת היטב. ננסה לפרק את הבעה.

הגדרה נניח שיש לנו מטריקה d (לדוגמה נובעת מנורמה). פונקציית המחויר עבור קלאסטרינג היא $\{x_1, \dots, x_m\} = \bigcup_{ij=1}^k C_j$

$$G(C_1, \dots, C_k) = \min_{\mu_1, \dots, \mu_k \in \mathbb{R}^d} \sum_{j=1}^k \sum_{x \in C_j} d(x, \mu_j)$$

כלומר סכום המרחקים המינימלי של $\{x_i\}$ כאשר המרחק הוא $m-x$ לוקטור שמזעור את RSS של כל הקלאסטרים ביחד.

הערה בהינתן C, μ הוא יחיד ומהו מרכז מסה למעשה centroid .

טענה נניח כי $\mathcal{X} = \mathbb{R}^d$ ו- d היא המטריקה האוקלידית, הראו כי הצנטרOID של C_j הוא המוצע האמפירי, $\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$.

כיצד נזער את G ? יהי G היא $\sum_{i=1}^m d(x_i, \mu_j)$ על כל החלוקות של $\{x_i\}$ ומזעור שלו הוא (נאיבית לפחות) בעיה קומבינטורית. הבעיה זו היא NP קשה ולכן נדרש לשימוש בהיוריסטיות.

k-Means Clustering

כל חלוקה משרה צנטרואידים כלשהם ובהתאם צנטרואידים מסוימים חלוקה - נתאים כל וקטור לחלוקת המתאימה לצנטרOID הקרוב ביותר לו.

נתחיל מחלוקת כלשהי, זו תשרה צנטרואידים, ואלו משרים חלוקה וכו' וכו', זה נקרא האלג' של Lloyd.

- נאותל μ_1, \dots, μ_k צנטרואידים (לדוגמה דגימות מסוימות, או אקרים, וכו').

- חזרה עד להתקנסות:

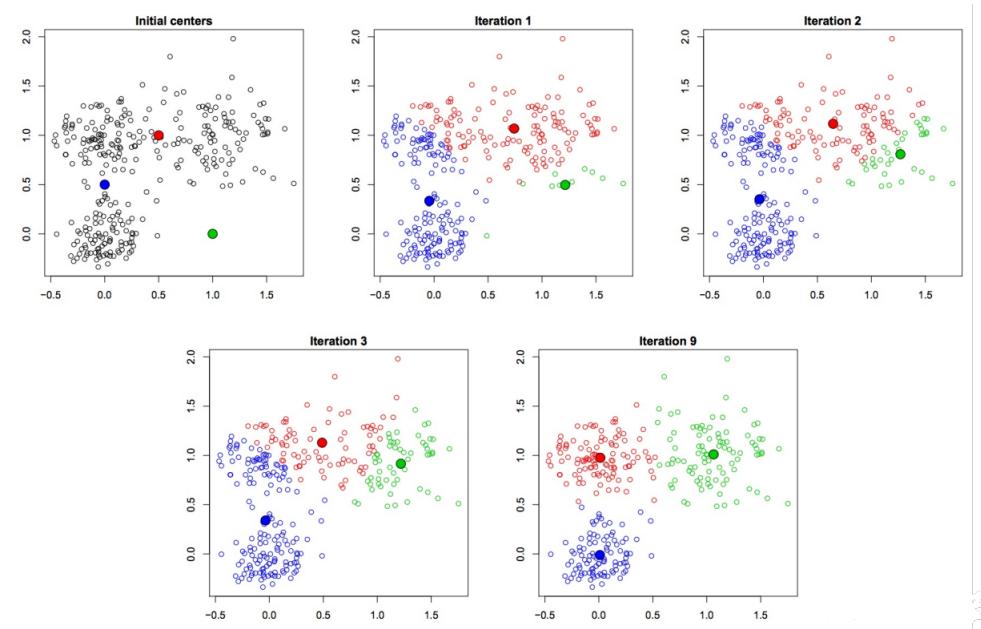
- קבע את C_j להיות כל הנקודות שקרובות ל- μ_j יותר מכל צנטרoid אחר, לכל j .

$$\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x, \text{ כפי שראינו זהו}$$

הערה זהו אלג' איטרטיבי מתחלף (בכל איטרציה הולכים הלוך ושוב).

הערה חלוקה שמושרת מצנטרואידים נקראת תאי Vornoi.

דוגמה עבור $d = 2$, $m = 300$ ו- $k = 2$, אנחנו מחלקיםכאן ל-3 צברים. ניתן לראות כי לאט אנחנו מזוהים הפרדה כלשהי יחסית מוגדרת בדתא.



תכונות של האlg' של Lloyd

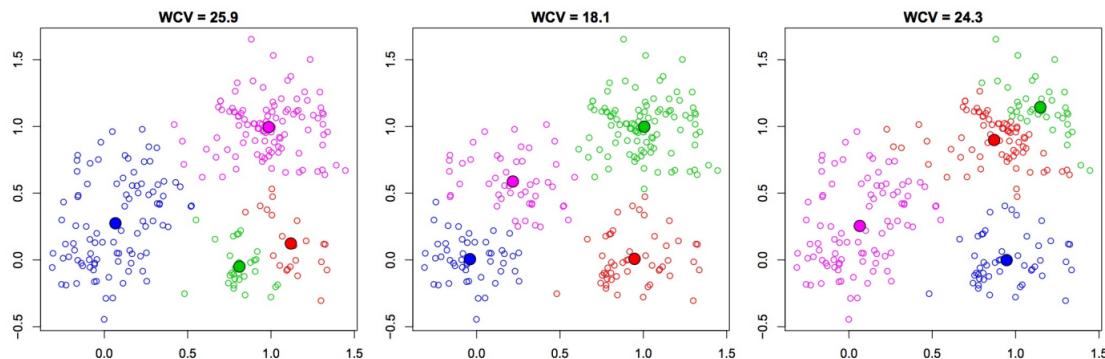
- השונות בכל קלאסטר יורדת בכל איטרציה.

- האlg' מתכנס גם אם מאוד לאט.

- התוצאות הסופיות תלויות מאוד באתחול הצנטרואידים.

דוגמה עבור $d = 2$, $k = 4$, עבור שלושה אתחולים שונים קיבלנו תוצאות שונות למלי. כאן אתחלנו במקומות שונים והתקנסנו לתוצאות שונות (למרות שיש מאפיינים משותפים). כדי לאמוד את התלות הזה אפשר להריץ הרבה פעמים ולהסתכם על

השונות של הקלאסטרים.



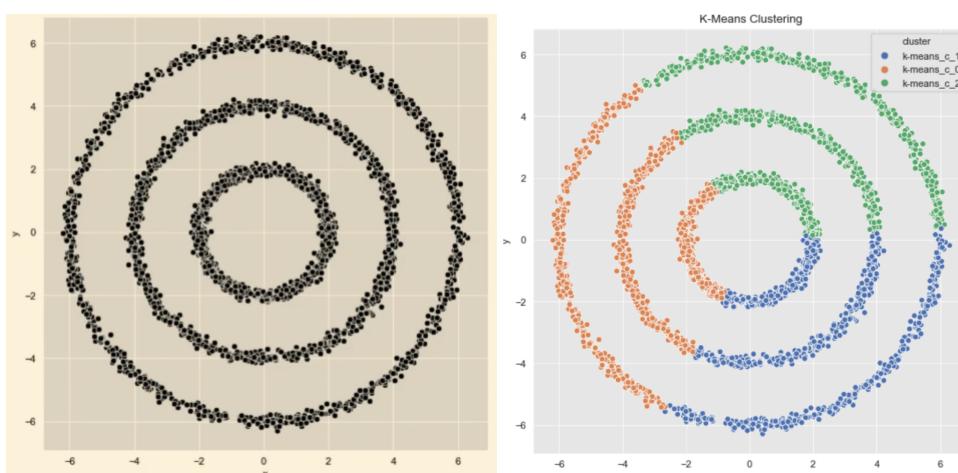
הבחירה של k גם כאן אינה מושכלת במיוחד פשטוט מסתכלים על הגרף של k -Means וקובעים אונשיית גבול. אם k גדול מדי אנחנו מקבלים אוברפיט (בטח בהתחשב בשונות הגבואה בהינתן תנאי ההתחליה). אם השיפור בין שני ערכי k לא גדול מדי, זה פשוט אומר שהייה לנו קלאסטר בסדר שהפרידנו לשני חלקים ולא הייתה סיבה (ולמעשה זה עושה נזק מבחינות שונות).

קלאסטרינג ספקטורי

k -Means מכיל כמה בעיות, השתיים שבהן נטפל הום האם ניון להטעלים ממתקדים גדולים בקלאסטרינג והאם ניתן להתבסס רק על מרחוקים בין כל שני איברים, במקום מטריקה על כל המרחב.

הערה האחרון חשוב כי ברשות חברתית לדוגמה מדובר בגרף עם קשיות בין אנשים ואין משמעות למרחב \mathbb{R}^d .

דוגמה עבור הדטא הבא (משמאלו), ברור לנו שיש מרחק גדול בין כל טבעת וברור שהחלוקת היא פר-טבעת, אבל k -Means יתן את התוצאה מימין,



במקרה הזה אנחנו רוצים לקחת את המידע המקומי (כל דגימה קרובה לדגימות אחרות במעגל אבל מאוד רחוקה מהאחרות) ולהסיק דברים בעורטה. זה בדיקת מה שאינטגרל עושה - הוא לוקח את הנגזרת, שיפוע/שינוי מקומי, ומסיק איך הפ' הכלליות נראה.

דוגמה בהינתן דטא על תМОונות של פנים של אנשים נרצה ליזוח אילו תМОונות הן של אותו האדם. המרחק האוקלידי (תאורתית) בין תМОונה של אדם מסתכל שמאליה לתМОונה של אותו אדם מסתכל ימינה יכול להיות מאוד גדול וזה יחרוס לנו את הקלאסטרינג. לעומת זאת,

אם נתונים לנו מרחקים בין כל שתי דוגמאות נוכל “לאנטגרל” ולקבל מרחקים בין נקודות שונות, כאשר נתעלם לחלוטין מרחוקים גודולים כי אין להם שום משמעות במציאות.

כל אסטריניג ספקטרלי מגיע מאיחוד של שלושה ריעונות:

1. נסתכל רק על מרחקים קטנים בין דוגמאות.

2. לבנות גרף ממושקל של מרחקים בין הדוגמאות.

3. להסתכל על k ה” α ” המובילים של מטריצת הסמיוכיות של הגרף.

הגדרה נגדיר מטריצת קשרים (Affinity Matrix) $[A]_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\epsilon}\right)$ כאשר $0 < \epsilon < \infty$ הוא היפר-פרמטר.

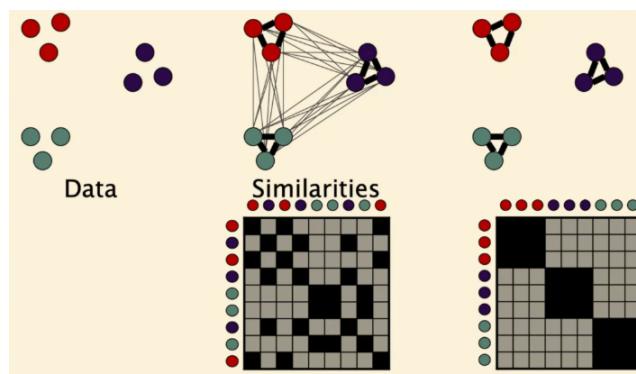
הערה הנורמה כאן שולחה למטריקה כלשהי, זה לא משנה. למעשה זו הפעם הראשונה שבה אנחנו יכולים ללמידה על $\mathbb{R}^d \neq \mathcal{X}$ אם רק יש לנו מרחקים בין כל שתי דוגמאות.

הערה האינטואיזיה לערכים של A היא שככל שהמרחב יותר גדול, ככח אקספוננציאלית נתעלם ממנו. רק מרחקים ממש קטנים יקבלו משמעות. איברים מאוד קרוביים מקבלים ערך קרוב-ל-1 ודוגמאות מאד רחוקות יקבלו ערך קרוב-ל-0.

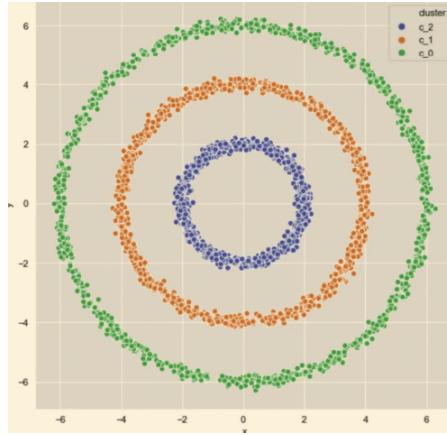
A היא מטריצת מרחקים סימטרית ונגדיר $L = D^{-1}A$ כאשר $D_{ii} = \sum_{j=1}^m [A]_{ij}$ והוא מכילה ערכים אי שליליים שעבורו סכום כל שורה הוא 1, זו נקראת מטריצת סטוטוכסטית.

הגדרה L נקראת גרף לפלייאן מנורמל על הגרף המוגדר ע”י A .

דוגמה באירור הבא רואים את הנתונים, ואז את המטריצה A (מעוגלת) ואז את המטריצה L שמסודרת לפי בלוקים כאשר כל בלוק מתאים לכל אסטר.



דוגמה במקרה של הטבעות, כל אסטריניג ספקטרלי מניב את התוצאה (הרצויה) הבאה,



האלג' של קלאסטרינג ספקטראלי

הקלט הוא \mathbb{R}^l סט $S = \{s_1, \dots, s_n\} \subseteq \mathbb{R}^l$ ואנחנו רוצים k קלאסטרים.

1. נבנה את המטריצה $A \in \mathbb{R}^{n \times n}$ כפי שהוגדרה לעיל.

2. נגדיר מטריצה אלכסונית D שאיברה ה- i -הו סכום השורה ה- i של A . ונבנה $L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

3. נמצא את $x_1, \dots, x_k \in \mathbb{R}^{n \times k}$ המובילים של L (אורותון אחד לשני) ונגדיר $X \in \mathbb{R}^{n \times k}$ שמודותיה x_i .

4. נורמל מחודש את X ל- Y .

5. נרים k -means על Y עם $y_1, \dots, y_n \in \mathbb{R}^k$ שחן שורותיה של Y .

6. נתאים את s_i לקלאסטר ה- j אם y_i הותאם לקלאסטר ה- j .

גם כאן הבחירה של k היא "לפי העין". לסיום, הרעיון של קלאסטרינג ספקטראלי הוא k -means לא על הדadata המקורי אלא הדadata מסודר במרחב חדש, מסודר לפי (בנפנו ידיהם) הcyoionics שימושים מבחינת מרחוקים קטנים בין איברים (בסיומו של דבר בלי כל הנרטומולים אנחנו עושים k -means על A שהיא מתנית המרחוקים בין כל שתי נקודות).

תרגול

אנחנו עוסקים עתה בניסיון למצוא את האופטימום בגרף הביאס-וריאנס המוכר. גם הפעם נשתמש בעיקרו של פישוט מחלוקת ההיפותזות או יותר נכון בחירה מושכלת של מודלים מורכבים מותוכה) ובניה מעלה של משחו יותר מורכב, וזאת מתוך להפחית את השונות, שכן אובייקטיב זה הדבר הכי גרווע שיכול לקרווט.

הערה כשלומד שלנו הוא בלתי-מורט, שניiat ההכללה היא אך ורק השונות. אם כן, למה שנרצה להשתמש באומד מوطה? לעתים השיפור בשונות יהיה כל כך שימושי שנדיף לקבל ביאס כלשהו בתמורה לשיפור בשונות.

כעת במקומות למזער רק שגיאה אמפירית, אנחנו גם נרצה למזער רכיב נוסף שיבטיח שאם נבחר מודל עם שגיאה אמפירית נמוכה (כלומר חSSH לאוברפייט), הוא יהיה "שווה" את זה מבחינת הביצועים שלו בכלל. נפמל זאת.

הلومד שלנו הוא אלג' שפותר את הביעת $\underset{h \in \mathcal{H}}{\operatorname{argmin}} \mathcal{F}_S(h) + \lambda \mathcal{R}(h)$ כאשר $\mathcal{F}_S(h)$ היא השגיאה האמפירית הקלאסית שאנו מכירים - RSS ו- \mathcal{R} تعיד על המרכיבות של h (פולינום ממעלה נמוך ופולינום ממעלה גבוהה יקבל ערך גבוה, לדוגמה). λ הוא הפרמטר ששולט בטריידוף בין השגיאה האמפירית והרכיבות של המחלקה.

אם λ גדול אנחנו מכונים למודל מאוד פשוט ודיקוק לא ממש חשוב לנו ואילו אם λ קטן בעיקר חשוב לנו השגיאה האמפירית ופחות כמה מרכיבות המחלקה (ואוברפייט פחות נמנע במקרה הזה).

דוגמה ב-Soft-SVM הבעיה שלנו הייתה

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \lambda \|w\|^2 + \frac{1}{m} \sum_i \ell^{hinge}((w, b) | x_i) y_i$$

כאן אנחנו מצד אחד ממקסימים מרג'ין (raiyo שמקסום מרג'ין שקול למזער הנורמה) ומצד שני מנסים לסתוב נכון כמה שייותר נקודות.

כל-ש-א יותר גדול ככל פחות אכפת לנו לטעות ויוטר מרג'ין גדול, כלומר אכפת לנו יותר מההתמונה הכללית ולא מהנתונים הספציפיים כאן, כלומר שונות נמוכה יותר.

איןטאטייבית, מזעור של $\|w\|^2$ מעיד על "רכיבות" ההיפותזה כי אם לדוגמה יש לנו קוודינטות שמתאפסות אז זה יותר פשוט מאשר היפותזה שמתחשבת בכל פיצ'ר ופיצ'ר.

Ridge Regression

ברידג' אנחנו פותרים את הבעיה

$$\hat{w}^{ridge} = \underset{w, b}{\operatorname{argmin}} \|Xw + b - y\|^2 + \lambda \|w\|^2$$

כל רכיב בנפרד (h -RSS והנורמה של w) הם קמורים ולכן יחד זה קמור ולכן כדי לפתרור את הבעיה נגורר ונשווה לאפס.

$$\frac{1}{2} \frac{\partial f}{\partial w} = X^T X w - X^T y + \lambda w = 0$$

כאשר הגזירה מותבسطת על הגזירה של RSS ברגression לינארית קלאסית. לכן $y = X^T X + \lambda I$ כלומר

$$\hat{w}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

$X^T X$ ומולכsoon אורטורן של מטריצות סימטריות נוכל לרשום $X^T X = UDU^T$ כאשר D היא אלכסונית עם ע"ע א"ש (כי X

היא PSD). אם v ו"ע של $I + \lambda X^T X$ אז

$$X^T X + \lambda I = U D U^T + \lambda U U^T$$

$$\begin{aligned} &= U(D + \lambda I)U^T \\ &= U \begin{pmatrix} \lambda_1 + 1 & & \\ & \ddots & \\ & & \lambda_d + 1 \end{pmatrix} U^T \end{aligned}$$

כאשר $0 \geq \lambda_i$ ולכן $\lambda X^T X + \lambda I$ הפיכה ולכן הביטוי מוגדר היטב. כאן, גם אם λ קטן מאוד, הפתרון הוא עם מטריצה הפיכה שיש לה פתרון יחיד ולכן יש לנו יציבות נומרלית במימוש במכונה.

$$\begin{aligned} \hat{w}^{ridge} &= (X^T X + \lambda I)^{-1} X^T y \\ SVD \text{ פיתוח} &= \left(V \Sigma^T U^T U \Sigma V^T + \lambda V V^T \right)^{-1} V \Sigma^T U^T y \\ &= V (\Sigma^T \Sigma + \lambda I) \Sigma^{-1} U^T y \\ &= X^\dagger y \end{aligned}$$

כאשר כרגיל $\Sigma_i^\dagger = \frac{\sigma_i}{\sigma_i^2 + \lambda}$ רק שהפעם $X^\dagger = V \Sigma^\dagger U^T$.

עכשו אפשר לראות שאם $\infty \rightarrow \lambda$ או $w^\dagger \Sigma$ מכיל ערכים מאד קטנים ואו \hat{w}^{ridge} קטן יותר ואיילו אם $0 \rightarrow \lambda$ אנחנו מתקרבים יותר ויותר לפתרון של ריגריסה לינארית קלאסית, כמובן.

Lasso ריגריסית

Lasso מגיע כשאנו רוצים לפשט את המודל באמצעות אי-שימוש בחלק מהפיצ'רים. במקום למצוא RSS מינימלי על מודל המוגבל במספר פיצ'רים לכל תת קבוצה של פיצ'רים, נרצה נורמה קטנה למשקלות הפיצ'רים, ככלומר במקום

$$\underset{w,b: \|w\|_0=c}{\operatorname{argmin}} \quad RSS(w,b)$$

נפתרו את

$$\underset{w,b}{\operatorname{argmin}} \quad RSS(w,b) + \lambda \|w\|_1$$

כאשר פרמטר הרגוליזציה כאן קבוע כמו אנחנו רוצים להגביל את המודל במספר פיצ'רים קטן יותר (בגלל שמדובר בהדפס איברים).

Sparsity

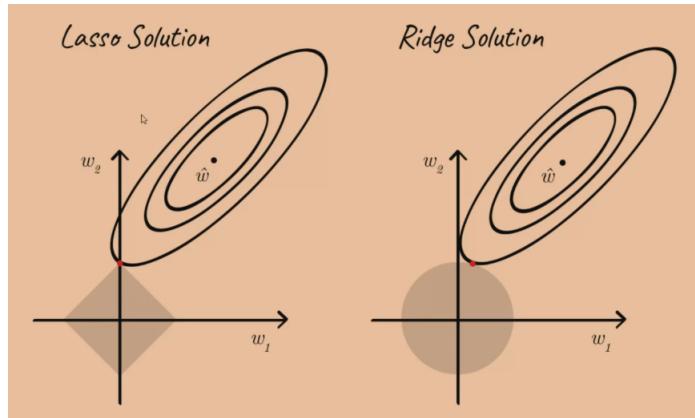
נסה להבין מדוע Lasso נוטה לאחסן איברים ו-ridge יכול אבל לא לעשות זאת זה באופן טיפוסי.

הגדרה גדרת נורמת ℓ_q היא $\|x\|_q = (\sum |x_i|^q)^{\frac{1}{q}}$.

הערכה עבור $1 \geq q$ זו נורמה ופ' קמורה. עבור $1 < q$ זו אינה נורמה אבל הפ' עצמה מקבלת ערכים גבוהים עבור וקטורים שהם sparse, כלומר עם הרבה קוורדינטות שונות.

אם נסתכל על בעיה דומה לרגוליזציה, $\underset{w: \mathcal{R}(w) \leq C}{\operatorname{argmin}} \mathcal{F}_S(w)$ אז כל הפתרונות צריכים להיות איפה שהוא בתוך כדור היחידה המוגדר ע"י ה"נורמה" $\mathcal{R}(w) = \|w\|_q$. ואילו בغالל שהפידלייטי קמור הרி שככל שמתחרקים מנקודת המינימום הערך תמיד גדול ולכן הפתרון לבעה זו הוא הנקודה הכי קרובה למינימום שהיא גם בתוך מעגל היחידה.

כאן נמצא ההבדל בין Ridge ו-Lasso - נקודות רוחקות מהראשית בתוך מעגל היחידה של נורמת ℓ_1 (קרובות יותר למינימום הפידלייטי) הןsparse ולכן התוצאה היא sparse ואילו ל-Ridge יש יותר אפשרויות, ביניהן לא sparse (ראו אייר).



הערכה במקרה הספציפי שבו $X^T X = I_d$ (כלומר שכל הפיצ'רים אורתוגונליים אחד לשני) יש לנו פתרון סגור לכל בעיות הריגסיה המודרנית שראינו, בניגוד למקרה הכללי שבו אין לנו פתרון סגור אלא צריך להשתמש בפתרני בעיות קמורות.

בחירה מודל

נניח שאנו עושים בוגינג בעיצים מקרים, נctrיך לבחור את עומק העץ המקסימלי - ככלומר זהו היפרפרמטר. כדי לבחור את היפרפרמטר האידיאלי, נחלק את הדאטא לדאטא אימון, דאטא ולידציה ודאטא מבחן, נאמן על האימון את המודל עם כל אפשרות של היפר-פרמטר, נבחר את האחד שמבצע הכי טוב על הולידציה ונדוח את התוצאה שלו על הטסט. ככה יש אי-תלות בשגיאת הכלכלה כי אנחנו לא מודוחים על האחד שמבצע על הטסט הכי טוב, אלא האחד שהכי טוב על הולידציה, עם שגיאת הכלכלה לפי הטסט.

הבעיה עם השיטה הזו היא שהולידציה תופס לנו הרבה דאטא שאנו רוצים לאמן עליו ולכן נעדיף להשתמש בשיטה אחרת CV.k-fold.

- נוציא דאטא טסט שלא ניגע בו עד שאנו מסיימים את בחירת היפר-פרמטר.

- ממה שנשאר, נחלק את הדאטא ל- k חלקים.

- לכל אפשרות להיפר-פרמטר:

- נוציא כל חלק פעם אחד בתור "ולידציה":

* נאמן על כל השאר ונחשב את השגיאה ונשמר את ממוצע השגיאות.

- נבחר את ההיפר-פרמטר עם ממוצע שגיאה הכי נמוך.

• נאמן על כל דאטא האימון (כל k החלקים)

- נדוח את השגיאה על דאטא הטסט ששמרנו בהתחלה.

שבוע II | שיטות גירעון

הרצאה

שיטות קernal הן לומדים לינאריים על סטראודים, שאפשר להפעיל על כל לומד סיוג או ריגריסה שפועל לפי כלל החלטה לינארי.

מחלקת ההיפותזות הלינארית היא כזו: $\mathcal{H}_{lin} = \left\{ x \mapsto w_0 + \sum_{i=1}^d w_i x_i : w_j \in \mathbb{R} \right\}$

הגדרה תהי $\mathcal{H}_\psi = \left\{ x \mapsto \sum_{i=1}^k w_i \psi(x)_i : w_j \in \mathbb{R} \right\}$, $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$. נגדיר את מחלקת ההיפותזות הבאה, $k > d$. נגידיר את מחלקת פולינומים מדרגה 3, הייתהה לנו דוגמה x והשתמשנו

הערה מדובר בלומד לינארי ב- (x) ψ במקום ב- x .

דוגמה כבר ראיינו שיטות קernal בלי להגיד את זה - התאמות פולינומיים. אם השתמשנו בפולינומים מדרגה 3, הייתהה לנו דוגמה x והשתמשנו

ב- $\psi(x) = (1, x, x^2)$ ואז התאמת פולינומיים הייתהה מודל לינארית ב- (x) .

כלומר, polynomial fitting, מושתמש בריגרסיה לינארית כדי ללמידה פ' מאווד לא לינארית באמצעות קernal פולינמיAli מדרגה n .

דוגמה כיצד נכלי זאת לממדים יותר גבוהים של \mathcal{X} (לדוגמה $\mathcal{X} = \mathbb{R}^2$)? נשתמש במולטיינום, בדרגה 2, שהוא פ' מהצורה

$$p(x_1, x_2) = w_{(0,0)} + w_{(1,0)}x_1 + w_{(0,1)}x_2 + w_{(2,0)}x_1^2 + w_{(0,2)}x_2^2 + w_{(1,1)}x_1x_2$$

$$\text{ואז } \psi(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$$

הגדרה פולינום כללי מדרגה n מ- \mathbb{R}^d ל- \mathbb{R}^d הוא פ', קלומר $\langle w, p \rangle$ כאשר $p(x) = \sum_{a \in \mathbb{N}_0^d : \sum a_i \leq n} w_a \prod_{i=1}^d x_i^{a_i}$ (כלומר $p(x) = \sum_{a \in \mathbb{N}_0^d : \sum a_i \leq n} w_a x^{a}$). ניתן לחישוב מדיסקרטית).

עבור \mathbb{N}_0 הינה הטבעיים עם 0 (ניתן לחישוב מדיסקרטית).

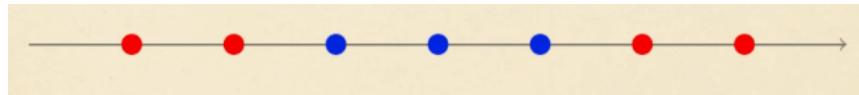
מוטיבציות לשיטות קernal

- העשרה מחלקת ההיפותזות הלינארית: רק באמצעות ψ הפיטה לפולינום כללי שהגנו מחלקת היפותזות הרבה יותר עשירה.

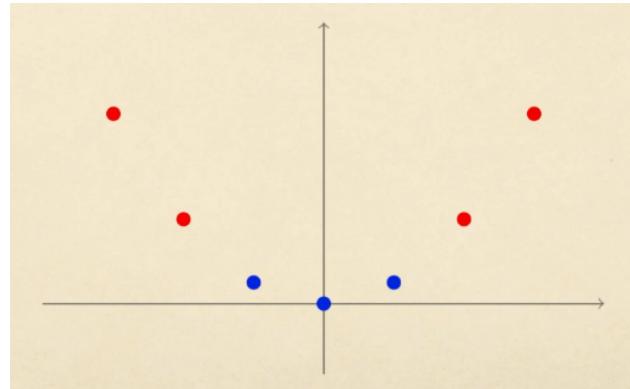
- יעילות חישובית: אפשר למצוא היפוטזה בעלי להסתכל על \mathbb{R}^k .

- למידה מטרית: כשאין פיצ'רים, מספיק מרחקים בין כל שני איברים כדי לאמן את המודל.

דוגמה עבור $\mathcal{X} = \mathcal{A}$, ראו נתוני הדמה הבאים.



ברור שהיעיות שלנו של המרחב באמצעות הKernel $(x, x^2) \mapsto x$ מאפשר הפרדה מושלמת עם חצאי מרחב (ראו איור).

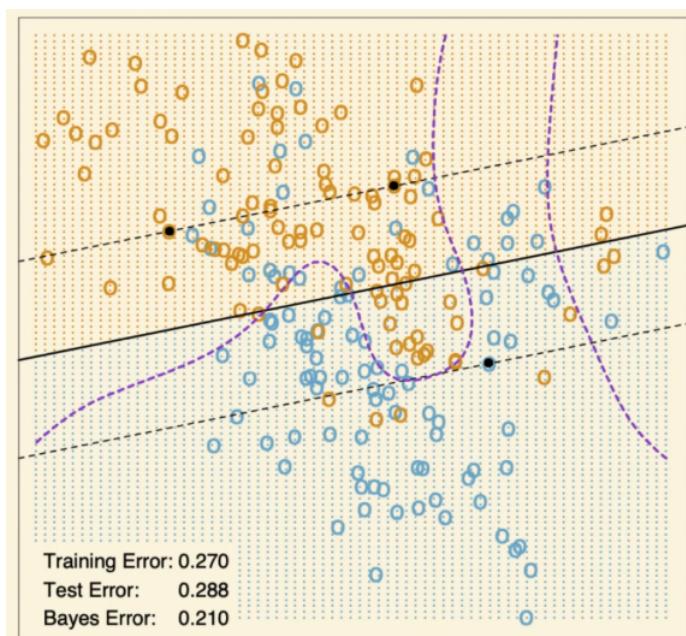


דוגמה החוק האוניברסלי של הכבידה אומר שכוח הכבידה בין שני גורמים הוא $F(m, M, r) = G \frac{mM}{r^2}$. מודל לינארי לא יכול ללמוד את החוק הזה, אבל אם נבחר $\psi(m, M, r) = (\ln m, \ln M, \ln r)$ אז הפ'

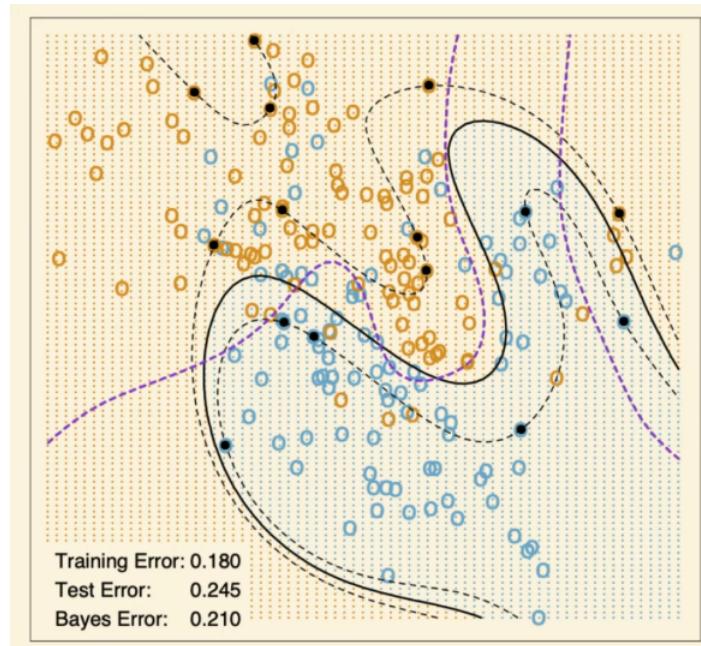
$$\ln F(m_1, m_2, r) = \ln G + \ln m + \ln M - 2 \ln r$$

היא לינארית ביצוג החדש ומתאימה בדיקות' שאנחנו מנסים ללמידה (ביאס 0).

דוגמה נתונים הנתונים הבאים, כאשר הקו השחור המקווקו הוא השול, השחור הוא הגבול SVM קבע והסגול הוא מה שהעלת מימד עם Kernel נוتنת - כל החלטה הרבה יותר טוב!



דוגמה גם תחת הנסיבות הבאים, עם קרNEL פולינומיAli (מדרגות שונות, שחור וסגול הם קרNELים שונים), אנחנו מצליחים לקבל מסוג הרבה יותר טוב,



הערה כדי להזכיר את הניבוא למרחב המקורי, צריך ש- ψ תהיה הפיכה (מקומית לפחות), אבל זה הרבה נכון כי גם העלת המימד לרוב תיינן לנו פ' חח'ע.

תבנית הkernel ולמידה עם kerNL

בזמנו אמרנו שמייד גבוה מדי במודל לינארי זה לא טוב. עם זאת, במקרה של שיטות קריל, באמצעות Kernel Trick נפתרת הבעיה.

נניח שיש לנו $S = \{(x_i, y_i)\}_{i=1}^m$ ונסתכל על לומדים שבוחרים $w_S \in \mathbb{R}^k$, $h_S \in \mathcal{H}_\psi$ (שcoil ל- k) לפי התבנית

$$w_S = \underset{w}{\operatorname{argmin}} f(\langle w | \psi(x_1) \rangle, \dots, \langle w | \psi(x_m) \rangle) + \lambda \|w\|^2$$

כאשר $f : \mathbb{R}^m \rightarrow \mathbb{R}$.

ראינו כבר הרבה לומדים שעוניים על תבנית זו, לדוגמה Maximum Margin (f הוא סכום ה-hinge-ים), ריגרסיה לינארית (f סוכמת על המרחק בריבוע בין y_i למשתנה ה- i), ERM, ריגרסיה לינארית עם רגולרייזציה, SVM, ריגרסיה לוגיסטיבית, MLE ועוד ועוד.

משפט (The Kernel Trick) נניח שיש לנו בעיית אופטימיזציה 1 מהצורה

$$w_S = \underset{w}{\operatorname{argmin}} f(\langle w | \psi(x_1) \rangle, \dots, \langle w | \psi(x_m) \rangle) + \lambda \|w\|^2$$

כאשר $x_i \in \mathbb{R}^d$ ו- $\lambda \in \mathbb{R}$, $f : \mathbb{R}^m \rightarrow \mathbb{R}$. עתה נביט בעיה אופטימיזציה 2 מהצורה

$$\alpha_S = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} f(G\alpha) + \lambda \alpha^T G\alpha$$

כאשר f, λ -ו $[G]_{ij} = \langle \psi(x_i) | \psi(x_j) \rangle$.

אזי הפתרון α_S לעיה 2 מקיים את הקשר הבא לפתרון w_S לעיה 1: $w_S = \sum_{i=1}^m [\alpha_S]_i \psi(x_i)$ (**לחותה**).

הערה המשפט מוכיח פה שהבעיות 1 ו-2 הן דואליות. למעשה אנחנו מראים ש כדי למצאו וקטור משקלות אופטימלי, כל מה שצריך לעשות זה לפטור את הבעיה של המקדמים שלו, כאשר הוקטורים עליהם אנחנו עושים "ל' ידועים כבר (להלן $(x_i) \psi$)".

הערה המשפט מראה גם שבמקרה ש- $m \gg k$ (שכיח) אז w לא צריך לחפש בכל \mathbb{R}^k אלא מספיק בת"ם ממימד m (α_S נמצא במרחב ממימד m).

הגדרה פונקציית קרנל K היא $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$: ψ ומוגדרת ע"י $\langle \psi(x) | \psi(x') \rangle = K(x, x')$

הערה אם אפשר לפתור את בעיה 2 רק עם K , אז אנחנו לא צריכים לדעת את ψ , את הפיצ'רים וגם לא את \mathbb{R}^k , אלא רק את מטריצת הגראם $[G]_{ij} = K(x_i, x_j)$ G המוגדרת ע"י

לכן בזמננו אימונו, נctrיך רק את $K(x_i, x_j)$ ולא את $\psi(x_i)$ ו- $K(x_i, x_j)$

בහינתן מודגמים אימוני S , אם אנחנו יכולים לחשב את $K(x, x')$ לכל x, x' , נחשב את G ונפתר את הבעיה $\alpha_S = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} f(G\alpha) + \lambda \alpha^T G\alpha$ להיפוטזה שנבחר מותוק H_ψ היא $\sum_{i=1}^m [\alpha_S]_i \psi(x_i)$ בבהינתן דגימה חדשה, נחשב

$$\begin{aligned} \langle w | \psi(x) \rangle &= \left\langle \sum_{i=1}^m [\alpha_S]_i \psi(x_i) | \psi(x) \right\rangle \\ &= \sum_{i=1}^m [\alpha_S]_i K(x_i, x) \end{aligned}$$

כלומר למעשה אנחנו משתמשים ב- α_S ולא משתמשים בערכי ψ בכלל, אלא רק K .

הערה כדי לשמר את המודל צריך לשמר את $\alpha_S \in \mathbb{R}^m$ איכשהו ועבור m גדול זה יכול להיות ממשמעוני (יקר).

הערה מבחינת המוטיבציות שהזכרנו, ברור שקיבלו את כולם: יש לנו מחלוקת היפוטזות מאוד עשרה עבור ψ בלבד; אם אנחנו יכולים לחשב את $K(x, x')$ ביעילות, לא צריך לחשב את ψ ולבצע ניבאים ולמידה ביעילות; ולא צריך פיצ'רים, רק את ערך הקרן על כל שתי דגימות.

הוכחה: (של **משפט הקרן טרייק**) נכתוב $w^* = \sum_{i=1}^m \alpha_i \psi(x_i)$ הפתרון לעיה כאשר $\psi(x_i) \perp u$ לכל i (ההטלה האורתוג'אל תוקן). $\langle w | \psi(x_i) \rangle = \langle w^* | \psi(x_i) \rangle$ ו- $\|w^*\|^2 = \|w\|^2 + \|u\|^2$. מתקיים $w = w^* - u$.

לכן הביטוי בתוך ה- argmin שווה עבור w ו- w^* , פחות $\|w\|^2$ לעומת w^* מזער לא פחות מ- w^* ולכן מואופטימליות w^* ו- $w = 0$.

כלומר הפתרון האופטימלי הוא בתוך $\{\psi(x_i)\}$, ולכן ניתן לבטא אותו כמשקל של $\{\psi(x_i)\}$ - זה נקרא המשפט המיצג.

מתקיים

$$\langle w | \psi(x_i) \rangle = \left\langle \sum_j \alpha_j \psi(x_j) | \psi(x_i) \right\rangle = \sum_{j=1}^m \alpha_j \langle \psi(x_j) | \psi(x_i) \rangle = [G\alpha]_i$$

$$\|w\|^2 = \left\langle \sum_j \alpha_j \psi(x_j) | \sum_j \alpha_j \psi(x_j) \right\rangle = \sum_{i,j=1}^m \alpha_i \alpha_j \langle \psi(x_i) | \psi(x_j) \rangle = \alpha^T G \alpha$$

ולכן אפשר פשוט לפתור את בעיה 1, כאמור

$$\underset{\alpha \in \mathbb{R}^m}{\text{argmin}} f(G\alpha) + \lambda \alpha^T G \alpha$$

כיאחרי שגילינו את המבנה של w^* (ק"ל של $\{\psi(x_i)\}$, הביטויים ב- argmin הראשון והשני שוים הגדրתי).

הערה k במקורו יכול להיות ∞ , כי כל מה שחשוב לנו כאמור זה (x, x') ולא ממש ψ לערכים ספציפיים.

אלגוריתמי למידה מקורנים

קרנול הוא לkit לומד קיים והפעלת קרנול עלי.

Hard-SVM מוקובל: במקרה החומני, הוא בעית המינימיזציה מהמשפט המיצג, נכון לרשום .
 $\underset{w: y_i \langle w | x \rangle \geq 1, \forall i}{\text{argmin}} \|w\|^2$

$$\text{ולכן הבעה המקורנת היא } \langle w | x_i \rangle = [G\alpha]_i \text{ ו- } \|w\|^2 = \alpha^T G \alpha$$

$$\underset{\alpha \in \mathbb{R}^m: y_i [G\alpha]_i \geq 1}{\text{argmin}} \alpha^T G \alpha$$

כאשר $[G]_{ij} = \langle \psi(x_i) | \psi(x_j) \rangle$ לכל ψ שرك נרצה.

Soft-SVM מוקובל: במקרה החומני, הבעה המקורית היא .
 $\underset{w, \xi: y_i \langle w | x_i \rangle \geq 1 - \xi_i, \xi_i \geq 0}{\text{argmin}} \frac{1}{m} \sum_{i=1}^m \xi_i + \lambda \|w\|^2$

از עבר פ' hinge המוגדרת ע"י $\ell^{hinge}(w, (x, y)) = \max \{0, 1 - y \langle w | x \rangle\}$ ופ' הסיכון האמפירי שהוא מושך Soft-SVM

$$\underset{w}{\text{argmin}} L_S^{hinge}(w) + \lambda \|w\|^2$$

שזה הרבה יותר נוח לkernel. אך הגרסה המקורנלית היא

$$\operatorname{argmin}_w \frac{1}{m} \sum_{i=1}^m \max \{0, 1 - y_i [G\alpha]_i\} + \lambda \alpha^T G\alpha$$

שזה אלג' מאד חזק כי יש לו גם רכיב רגולרייזציה (מוריד שונות) וגם יכולות קירנוול (יכול להעלות שונות).

3. Ridge: ריגרסית רידג' במקורה החומגי היא

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^m |y_i - \langle w | x_i \rangle|^2 + \lambda \|w\|_2^2$$

ואם נעביר את x ל- $\psi(x)$ נקבל

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^m |y_i - \langle w | \psi(x_i) \rangle|^2 + \lambda \|w\|_2^2$$

ובכתיב מטריציוני (מקורנוול על מלא) נקבל

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \|G\alpha - y\|^2 + \lambda \alpha^T G\alpha$$

לא זו בלבד, הסטודנטית המשקיעה תוכיח באמצעות פיתוח חדש של המשוואות הנורמליות (בגרסה המקורנלית) שהנוסחה הסגורה לפתרון בעיית הרידג' המקורנוול היא $\hat{\alpha} = (G + \lambda I)^{-1} y$.

הערה בקרוב נראה שהסיבוכיות של החישוב (d) (x, x') $\mathcal{O}(K)$ ומכאן הסטודנטית המשקיעת תחשב את סיבוכיות הזמן לaimon וניבוא על נקודות חדשות.

4. ריגרסיה לוגיסטיבית מקורנוול: ריגרים לוגיסטיים עם רגולרייזציית ℓ_1 בלי אינטראפטים היא

$$\operatorname{argmin}_{w \in \mathbb{R}^d} \sum_{i=1}^m \log \left(1 + e^{\langle x_i | w \rangle} - \sum_{i=1}^m y_i \langle x_i | w \rangle \right) + \lambda \|w\|_2^2$$

והסטודנטית המשקיעת תחשב את הגרסה המקורנלית של זה.

קרנלים מפורטים

קרנלים טובים הם קרנלים שניתן לחישוב ביעילות את (x, x') (K (אנחנו מחשבים את זה m^2 פעמים) ולכן צריך איזושהי סימטריה או טריק כלשהו בפ' שיעזר לחישוב יעיל - לא כל פ' עובדת.

1. **הקרナル הפולינומיAli:** $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ המוגדרת ע"י $\psi(x)_a = \prod_{i=1}^d x_i^{a_i}$ (כל סדרות החזקות שמביאות אותנו למונומים מדרגה לכל היותר n).

לכואורה החישוב כאן הוא $\mathcal{O}(k) = \mathcal{O}(d^n)$ אבל בכלל שאנחנו מדברים עליו, נראה שיש טרייק.
אם נגיד ψ עם עוד כמה קבועים $(\sqrt{2})$ וכאללה, לא מעניין אותנו בעצם כי בסופו של דבר אכפת לנו רק M - K ולא $M-\psi$, אז בסופו של דבר נקבל ש- d שווה חישוב לינארי ב- d :

תרגיל הסטודנטית המשקיפה ראה איך עברו $d = 1$ הנוסחה אכן מתקינה (פושט $(1 + xx')^n$, כפל מספרים) ובעור $d = 2$ והשורה היא

$$\psi(x_1, x_2) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$$

$$\text{ומתקיים } \langle \psi(x_1) | \psi(x_2) \rangle = (1 + \langle x_1 | x_2 \rangle)^2$$

דוגמה על מולטיינום עם גורלייזיצית ℓ_2 כדי להתאים פולינום מדרגה n , בנינו מטריצת או-דר-מונייה $n \times m$ והשתמשנו ביריגסיה לינארית. נניח $\mathcal{X} = \mathbb{R}^d$ ואנחנו רוצים ללמידה $f : \mathbb{R}^d \rightarrow \mathbb{R}$ כאשר המודגש שלו הוא $y_i = f(x_i)$ לכל $i \in [m]$. נבחר ללמידה את f באמצעות פולינום רב-משתני מדרגה n .

שימוש ב-polyfit קלאסי היה מביא עליינו מטריצה עם $\mathcal{O}(d^n)$ עמודות שווה לא פרקטוי. עם זאת, עם ריגרסית רידג' מוקורן, נפתרו את

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \|G\alpha - y\|^2 + \lambda \alpha^T G(\alpha)$$

ואז ננба באמצעות G כפי שהראנו לעיל.

2. הקרנל הגאוסיאני (RBF): עבור \mathcal{X} , נגדיר

$$\psi(x) = \left(1, e^{-\frac{x^2}{2}}, \frac{1}{\sqrt{2}}e^{-\frac{x^2}{2}}x^2, \dots\right)$$

(איינסוף) ופורמלית $\psi(x)_n = \frac{1}{\sqrt{n!}}e^{-\frac{x^2}{2}}x^n$. כלומר ψ היא העתקה מ- \mathbb{R}^d ל- \mathbb{R}^∞ , שהוא מרחב הילברט - מ"ז עם מימד איינסוף.

הגדרה מרחבי הילברט H הוא מ"ז מעל \mathbb{R} עם מכ"פ $\langle \cdot | \cdot \rangle$ המקיים שכל סדרת קושי לפי הנורמה $\|x\| = \sqrt{\langle x | x \rangle}$.
דוגמה מרחב ℓ_2 מכיל את כל הסדרות האניטופיות ($x = (x_1, x_2, \dots)$ שמקיימות $\sum_{i=1}^{\infty} x_i^2 < \infty$). עם הנורמה $\langle x | y \rangle = \sum_{i=1}^{\infty} x_i y_i$ יש גבול-ב- H .
זהו מרחב הילברט ממשי מימד איינסוף.

הגדרה אם לכל סדרת קושי יש גבול, המרחב נקרא שלם.

הערה שלמות המרחב מבטיחה קיום הטלה אורתוג' ייחידה על כל ת"מ (סגור) של H .

מתקינים

$$\begin{aligned} \langle \psi(x) | \psi(x') \rangle &= \sum_{n=0}^{\infty} \left(\frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \right) \left(\frac{1}{\sqrt{n!}} e^{-\frac{x'^2}{2}} x'^n \right) \\ &= e^{-\frac{x^2+x'^2}{2}} \sum_{n=0}^{\infty} \frac{(xx')^n}{n!} = e^{-\frac{(x-x')^2}{2}} \end{aligned}$$

ולכן החישוב של K הוא בזמן קבוע.

עבור $\mathcal{X} = \mathbb{R}^d$ פשוט נחליף את $x - x'$ ב- $\|\cdot\|$, ופורמלית, $\psi : \mathbb{R}^d \rightarrow \ell_2$ היא העתקה שלכל $a \in \mathbb{N}_0^d$ מוגדרת ע"י

$$\psi(x)_a = \frac{1}{\sqrt{n!}} e^{-\frac{\|x\|^2}{2}} \prod_{i=1}^d x_i^{a_i}$$

ואז $K(x, x') = e^{-\frac{\|x-x'\|^2}{2}}$. כדי למשם Soft-SVM לדוגמה, נחשב את K לכל הדגימות באמצעות האקספוננט הפשטוני הנ"ל ונרייצ' פוטר בעיות קמורות על הביטוי עם G זהה.

3. $K(x, x') = \|x - x'\|^2 \log(\|x - x'\|^2)$ כאשר לא נתיחס בכלל לא- ψ . כמובן, יכול להיות שהוא אפילו לא קיים וזה הכל אשליה אבל לא אכפת לנו כי יש לנו קרナル. עם זאת, כן יש לנו לפחות אפיון של מה ψ יכול להיות אם הוא קיים.

משפט תהי $[G]_{ij} = K(x_i, x_j)$ אזי K מכ'פ' במרחב הילברט, אם "ס" לכל G, x_1, \dots, x_m המוגדרת ע"י $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ היא PSD.

$$\text{כasher } a, b \text{ ממשיים.} .4$$

למידה מטרית

כאמור, בلمידה עם קרナル לא צריך פיצ'רים, אלא רק את היכולת לחשב את K בין איברים שונים, בלי x -ים אפילו. מספיק שיש לנו מטריצה נניח שיש לנו ליבלים על \mathcal{A} . נקבע אורך מחרוזות כלשהו n . נגיד $\mathcal{A}^n \rightarrow \mathbb{R}^{|A|^n}$ באופן הבא: לכל מחרוזות $\mathcal{X} \in s$ ולכל מחרוזת \mathcal{A} מוגדר $\psi(s)$ שהוא מספר המופיעים של a ב- s בסאב-סטריניג.

דוגמה תהי $\mathcal{A} = \{a, b\}$ (חומרות אמינו לדוגמה) ו- \mathcal{X} מרחב כל המחרוזות הסופיות של הא"ב. נניח שיש לנו ליבלים על \mathcal{A} . נקבע אורך מחרוזות כלשהו n . נגיד $\mathcal{A}^n \rightarrow \mathbb{R}^{|A|^n}$ באופן הבא: לכל מחרוזות $\mathcal{X} \in s$ ולכל מחרוזת \mathcal{A} מוגדר $\psi(s)$ שהוא מספר המופיעים של a ב- s בסאב-סטריניג. ה الكرナル במקרה הזה הוא $K(s, s') = \sum_{a \in \mathcal{A}^n} \psi(s)_a \psi(s')_a$.

יש אלג'יעיל שמחשב את K שלא דרוש מעבר על כל $\mathbb{R}^{|A|^n}$, ומשם ללמידה עם כל לומד מקורナル שראינו לעלה, בלי שימוש פיצ'רים. **הערה** בכל זאת, גם בלי פיצ'רים, אנחנו צריכים עדיין להיות יכולים לחשב את K בין כל דוגמה במדגם האימון לדוגמה חדשה כדי לנבא. **דוגמה** אם אנחנו בודקים דמיון בין סרטים (m, m') או רמת הדמיון בין שני סרטים) אז גם בלי פיצ'רים, אפשר לחשב על דוגמה חדשה דמיון באמצעות חישוב מספר השחקנים המשותפים, מרחק בין שנות ההוצאה וכו' וכו'.

הערה אפשר לעשות PCA מקורナル (בלי פיצ'רים בכלל).

תרגול

למה הורדת מידע זה חשוב?

- בשביל ויזואלית, אנחנו צריכים מימד מאוד נמוך כדי לייצג דברים (לכל היותר 4)

- בשביל לשפר זמן ריצה (יותר פיצ'רים דורשים יותר זמן)

- בשביל להתמודד עם רעש (מסתכלים רק על המgomות המרכזיות בפיצ'רים).

- בשביל לא להטעכ卜 על תלות בין פיצ'רים (יכול להיות שניים קשורים והורדת המימד "תאחד" אותם).

דוגמא נגיד שיש לנו מעגל דו ממדי במרחב תלת ממדי עם רעש. נוכל להטיל את הדאטא על ת"מ מימייד 2 (הטלה) ולקבל בלי אובדן רב את הדגימות בדו מימייד (שיכון).

האלג' המركזי שנעסק בו שמאוד פופולרי SMB צע הורדת מימייד הוא PCA.

PCA

PCA מחשב את $A = \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^T$ הגודלים ביוטר (הו"ע המוביילים).

הערה בהרצאה ראיינו שמה ש-PCA בעצם עשו זה למצוא את

$$W^*, U^* = \underset{W \in \mathbb{R}^{k \times d}, U \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} \sum_i \|x_i - UWx_i\|^2$$

כאשר הביטוי שאנו מזעירים נקרא שגיאת השחזר (הLINEARITY). גילינו ש- $(U^*)^T$ ו גם שעמודות U^* הן u_1, \dots, u_k שהוגדרו לעלה.

נוכיח כי PCA לא פותר רק את הבעיה הזו, אלא גם ממקסם את השונות על הנתונים המשוכנים.

הערה בכללי שונות זה משחו שאנו לא אוהבים אבל במקרה הזה, שונות היא דבר טוב כי היא מייצגת את הממציאות (שאנו מכירים) ולכן אם היא נשמרת (מקסימלית) זה אומר שלא הזינו יותר מדי, שהוא דומה למזעיר שגיאת השחזר.

טענה PCA ממקסם את (אומד) השונות של $X^T U$.

הוכחה: נתחילה במקרה שבו $k = 1$. ההטלה שלנו היא פשוט $x^T v$ כאשר v וקטור ייחידה בלבד. לכן

$$\begin{aligned} E_x[v^T x] &= \frac{1}{m} \sum v^T x_i = v^T \bar{x} \\ \text{var}(v^T x) &= E_x[(v^T x - E_x[v^T x])^2] \\ &= \frac{1}{m} \sum (v^T x_i - v^T \bar{x})^2 \\ &= \frac{1}{m} \sum (v^T (x_i - \bar{x}))^2 \\ &= \frac{1}{m} \sum (v^T (x_i - \bar{x})) (v^T (x_i - \bar{x}))^T \\ &= v^T \frac{1}{m} \sum (x_i - \bar{x})(x_i - \bar{x}) v \\ &= v^T S v \end{aligned}$$

ולכן נרצה את $S = UDU^T$ (פירוק ספקטרלי) ולכן $\hat{v} = \underset{\|v\|=1}{\operatorname{argmax}} v^T S v$

$$\begin{aligned} S &= UDU^T = \sum_{i=1}^d \lambda_i u_i u_i^T \\ v^T S v &= v^T \left(\sum \lambda_i u_i u_i^T \right) v = \sum \lambda_i v^T u_i u_i^T v_i = \sum \lambda_i \langle v_i | u \rangle^2 \end{aligned}$$

כאשר v_i היא העמודה ה- i של U ובגלל ש- U אורתוג'

$$\sum \langle v_i | u \rangle^2 = \|U^T v\|^2 = \|v\|^2 = 1$$

ולכן $\sum \lambda_i \langle v | u_i \rangle^2$ הוא סכום קמור ביחס ל- $\{u_i\}$. עבור סכום קמור כזה, המקסימום מתקבל כאשר יש משקלות ייחודית שאינה אפשרה, והיא מתאימה לערך הגבוה ביותר, כלומר, ככלומר $\lambda_1 = \max_{v: \|v\|=1} \sum \lambda_i \langle v | u_i \rangle^2$ ולכן

$$u_1 = \underset{v: \|v\|=1}{\operatorname{argmax}} \sum \lambda_i \langle v | u_i \rangle^2$$

ולכן במקרה הכללי,

$$v_i = \underset{\|v_i\|=1 \wedge v_i \perp v_1, \dots, v_{i-1}}{\operatorname{argmax}} \sum \lambda_i \langle v | u_i \rangle^2 = u_i$$

■ **כלומר מטיריצת ההטלה שלנו אכן עמידותיה הן u_i .**

עתה נעסק בקלאסטרינג. האלגוריתם אינטואטיבי לכך הוא k-means $\underset{\mu_i, S_i}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2$ שמשמער את k-means כאשר μ_i הוא המרכזoid של S_i והוא הקלאסטר ה- i . האלגוריתם מצליח לתפוס קרבה, אבל לא מרכבות גאותטרית, לדוגמה מעגליים. לשם כך הומצא קלאסטרינג

ספקטראלי שכן מתאפיין לנתחונים באופן יותר עדין חוץ מקרבה.

קלאסטרינג ספקטראלי

אין משמעות לאינדקסציה של הפסקאות, זה ניסיון להתחקות אחר הפורמט של חווות דעת (משפטיות), שמסודרות לפי פסקאות ממושפרות לשם נוחות הקריאה/הפרדה לוגית.

1. נגידר G כאשר $A \in \mathbb{R}^{m \times m}$ ונגדיר $E = \{(i, j) : \|x_i - x_j\| < \epsilon, i \neq j\}$ ו- $V = \{1, \dots, m\}$ מטריצת שכנות ע"י

$$[A]_{ij} = \mathbb{1}_E(\{i, j\}) = \begin{cases} 1 & \|x_i - x_j\| < \epsilon, i \neq j \\ 0 & \text{אחרת} \end{cases}$$

2. נגידר $[D]_{ii} = \sum_{j=1}^m a_{ij}$ ע"י $D \in \mathbb{R}^{m \times m}$ שבה עוזר לנרטול A .

3. נחשב A , זה הלפלסיאן של הגף G .

L היא PSD והע"ע הנמוך ביותר של L הוא 0 (זהו קיים) והרכיבי הגאומטרי שלו נקבע על פי מספר רכיבי הקשרות של G .
בנוספ', המ"ע של הע"ע 0 נפרש ע"י $\mathbb{1}_{[C_1]}, \dots, \mathbb{1}_{[C_k]}$ כאשר $\mathbb{1}_{[C_i]}$ הוא וקטור שבוקור' $-j$ שלו יש 1 אם $j \in C_i$ והוא 0 אחרת.

בפועל לא נשמש בלפלסיאן ה"טהור" אלא בגרסה המנוורמלת שלו, $L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ (שהיא יותר יציבה נומרית).

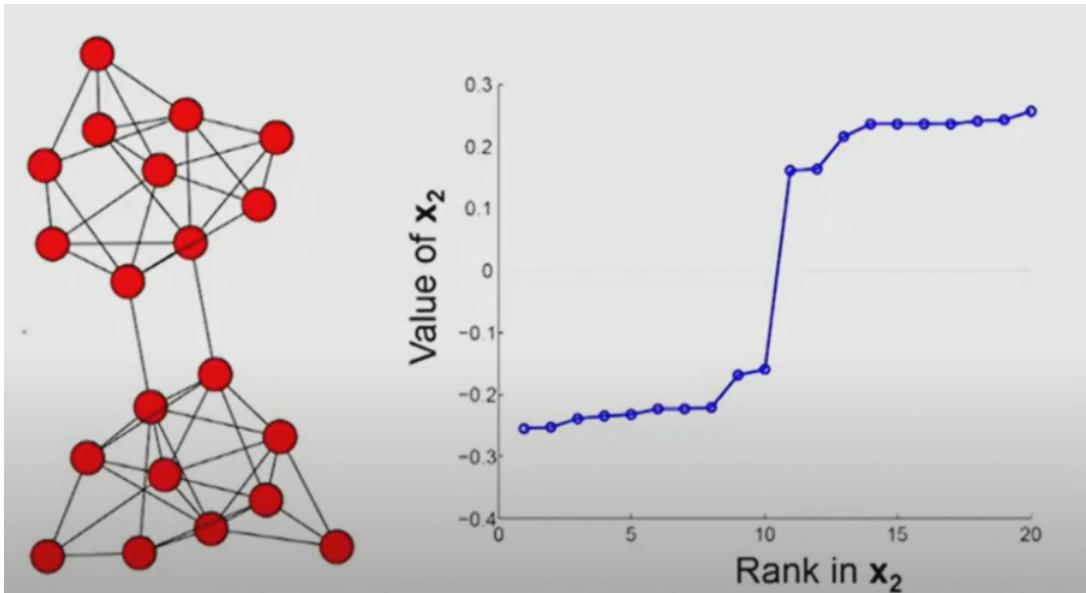
4. בה"כ נניח כי $A = \begin{pmatrix} A_{C_1} & & 0 \\ & \ddots & \\ 0 & & A_{C_k} \end{pmatrix}$ כאשר A_{C_i} היא מטריצת השכנות עבור הקודקודים ברכיב הקשרות ה- i .

$$. L \vec{1} = (D - A) \vec{1} = D \vec{1} - A \vec{1} \text{ כולם } [A \cdot \vec{1}]_i = \sum_{j=1}^m [A]_{ij} = d_i$$

נכיסן $L = UPU^T$ ובגלל ש- $\vec{1}$ י"ע עם ע"ע 0 (ראינו לעלה), בה"כ נניח כי $\vec{1}$ מתקיים $i < m$. לכן לכל $u_m = \frac{1}{\sqrt{m}} \vec{1}$ (סכום כל עמודה הוא 0) כי U אורטור (ולכן

$$\left\langle u_i \mid \frac{1}{\sqrt{m}} \vec{1} \right\rangle = \frac{1}{\sqrt{m}} \sum [u_i]_j = 0$$

5. נניח כי G גוף עם שני רכיבי קשרות (או גוף עם שתי קבועות כמעט כמעט-נפרדות), לכן u_m וניתן להוכיח כי קיימים $0 < c <$ ש- $u_{m-1} = \begin{pmatrix} \mathbb{1}_{[C_1]} \\ -c \mathbb{1}_{[C_2]} \end{pmatrix}$ והוא הערך באינדקס i ב- u_{m-1} (שמייצג קודקוד) וכי הערך באינדקס i ב-



עבור k כללי, נגדיר $U_{-k} = \begin{pmatrix} \vdots & \vdots \\ u_{m-k+1} & u_m \\ \vdots & \vdots \end{pmatrix} \in \mathbb{R}^{m \times k}$ ונוירץ k-means על שורות המטריצה הזו (השורה ה- i היא הדגימה ה- i). ומחזירים את ההתאמה של הדגימה ה- i המקורית למה שבחרנו לשורה ה- i של U_{-k} .

הערה בغالל שההפרדה היא לא תמיד ברורה, במקרה לבחור לפי סימן, נבחר לפחות ϵ קטן ככלו.

במקום A כפי שהגדכנו אותה במקור, אפשר להגיד שערך המרכיבים בין כל שתי דוגמאות זהה מאפשר הבנה יותר מורכבת מאשר סתם מבחון סף.

שבוע XIII | איך לפתור בעיה בלמידה

הרצאה

מערכות לומדות דורשות שני קישורים דינמיים: הבנה מתמטיקה ופורמליקה של רעיונות; ומומחיות בניתוח נתונים, תכונות והציגת תוצאות.

lagst לבעיה חדשה

עד כה התعلמנו איך הגיעו לבעה אלינו, והנחנו שהיא שם, ומשם למדנו מה לעשות אליה.

- הגדרת סט-אפ: כל הרצאה התיכילה כשהיינו שאנחנו מתחשים \hat{w} שמצויר איזושהו ביטוי, אבל זה מסתיר הרבה שלבים במאיצע.

- צריך לבדוק אם בכלל צריך ML בשביל הבעיה, כי יכול להיות שאפשר לפתור rule based, כלומר בלי ML ו-AI, שהוא טוב יותר readability, דטרמיניסטיות וכו'.

דוגמה (הדוגמאות בהרצאה הן מביעות שד"ר סטנובסקי עסק בהם) לנבה כמה חודשי מסר ינתנו בהינתן פסק דין שנתון בשפה טבעית. הגישה שפתרה היבט את הבעיה הייתה לזהות מה המשפט של גור הדין היה, ומשם לעשות **supervised learning** כדי לזהות מה גור דין עצמו. רוב העבודה הייתה על התשתית הזו.

– מה הופך בעיה טובה ל-ML? אם קשה לחשב על חוקים קונקרטיים לפתרון, אם הבעיה עצמה לא דטרמיניסטית ואם יש רוש שקשה למצל. עם זאת, אם יש יותר מדי רוש או שאין סימנים מקרים בין הנתונים לפתרון, זו לא בעיה טובה ל-ML. בנוסף, פיצ'רים אינדיקטיביים הם חשובים. חשוב גם שתיהיה יכולה להכליל את הפתרון, כך שלא נגע תמיד לאובר-פייט. חשוב לחשב על הקשר של המודול – איפה ישמשו בו? איך מודדים את הביצועים שלו? האם הדאטה מייצג (אם אכן אמת) המודל בישראל אבל נשמש בו בנסיבות לא עשוין כלום)? כל השאלות האלה ישפיעו מאוד על בחירות שלנו.

דוגמה מודל שבודק אם תרופה מסוימת שונה מאוד ממודל שמנבא מותי האוטובוס יגיע (מדד הביצועים מאוד שונה). – צריך לדעת הרבה מאוד על תוכן שבו נמצא הטעיה, והרבה מהעבודה היא לקשר בין עולם ה-ML (פורמליקה מתמטית של תופעות) לבין עולם התוכן של הבעיה.

• להבין את הפיצ'רים:

– להסתכל על \mathcal{X} סתם בתור מספרים לא יעזר, צריך להבין את מה הפיצ'רים אומרים:
 * האם הם קטגוריים? אם כן האם יש סדר בין ערכיו ואילו ערכיהם יכולים להיות לו?
 * האם הם רציפים? מה הטווח החוקי שלהם?
 – האם יש לפיצ'רים שם עם משמעות? מה היחידות של הפיצ'ר? אפשר לבקש עוד פיצ'רים? מי דוגם את הפיצ'רים? כל אלה שאלות חשובות שישפיעו על הדרך בה נתעסק עם הדאטה.

• להטיל ספק בליבלים: עד כה סמכנו על הליבלים בעניינים עצומות. יכול להיות שהlivelims מאוד רועשים, וצריך להסתכל על הדאטה (אולי עם מומחה) ולבדוק האם הליבל באמת הגיוני. רוש ב- X הוא שונה מרוש ב- Y . אם יש רוש, יכול להיות שאף אלג' לא צלח ללמידה כמו שצדך.

אפשר להשוות בין מתייגים שונים כדי לראות האם יש הסכמה על מה הליבל אמר להיות – לעיתים יש שטח אפור.

דוגמה לזהות האם ביקורת על סרט היא חיובית או שלילית. הליבלים ביןaries – ביקורת חיובית או שלילית. אם מישחו לנו את הביקורת "השחקן הראשי היה טוב אבל הבמאי היה גרווע" – האם זו ביקורת חיובית או שלילית? לשם כך אפשר לשאול כמה אנשים שיתיינו את הדאטה ואיזו להתייחס לlivelims לפי ההסכמה שלהם, ולהתעלם מalto שאינו עליים הסכמה.

האם 80% הסכמה על הדאטה הוא מרשים? Cohen's Kappa הוא ממד שקובע את רמות ההתארשות (כמה נסמך על ההסכמה) להיות $\kappa = \frac{p_0 - p_e}{1 - p_e}$ כאשר p_0 הוא ההסכמה הנכetta (נגיד 80%) ו- p_e הוא התוחלת של הסכמה על הטלת מטב.

דוגמה במקרה של 80% עם שתי מחלקות, נקבל $\kappa = 0.5$ ואיזו $0.6 = \kappa$ שווה משמעותית נמוך מ-0.8 אבל אם יש 1000 מחלקות, נקבל p_e הרבה יותר גבוה ואיזו $0.799 = \kappa$.

כרגע צריך להיזהר עם בני אדם כי יש להם הטויות.

דוגמה בוגול טרנסלייט doctoranganlit מתורגם לרופא גם אם זה בהכרח רופאה, כי מזעור השגיאה על הדאטה יציג את מה שהמודל צפה על פי רוב - אכן רוב התוכן בעולם הוא של רופאים גברים ולא נשים, אבל זה לא אומר שבמקרה הספציפי הזה זה גם גבר.

הבעיה היא לא בעיה חברתית, אלא פשוט שנבנה לא נכון בניסיון למזער כמה שיוטר את השגיאה, لكن כדי לדוגמה במקרה הקודם לאזן בין מספר המשפטים עם רפואיים ורופאות (יש עוד שיטות).

– חלוקת הדאטה באופן הגיוני: במציאות, לרוב נקלט דאטה הומוגני עם לייבלים, ואנחנו אמורים על חלוקה לaimon, test וco'. לצורךחלוקת הדאטה כך שתמיד יוכל לבדוק הכללה, ולהכליל היטב. כפי שאמרנו בעבר, כדאי להסתכל על דאטה מצומצם מתוך הדאטה הנתון כי אחראית להרשות את היכולת להכליל. בנוסף, חשוב לדאוג שהחלוקת תהיה מואצת מבחינת הליבלים.

דוגמה רוב האנשים לא חוטפים התקף לב, ולכן אין סתמי נחלק את הדאטה נמצא עתה עצמנו עם חוסר איזון חמור של לייבלים, וכך שמחקרים צריכים לשמר על יחס המחלקות בעלייבלים.

פיתוח מודל

המודל המקורי לפיתוח מודל מוצלח מורכב מארבעה חלקים:

Preprocessing → EDA → Baseline → Model Selection

• פרי-פרוססינג: DATA שמנגין מקור אמיתי בסדר גודל גדול גיע עם שגיאות. פרי-פרוססינג הוא שלב עיבוד המידע לפני שימושים אותו למודל. נדרש להחליף ערכים חסרים ומושחתים, לנורמל וליצור פיצ'רים חדשים מפיצ'רים קיימים.

– שלב זה משפייע רבות על ההצלחה וראוי להשיקע בו.

– כਮון שעושים פרי-פרוססינג רק על האימון ולא על הטסט כי אז אנחנו לא מכילים במודל, אלא פשוט משתמשים על דברים שכבר ראיינו.

דוגמה יש כמה סיבות למוחסור מידע - טעות בתהיליך הזנת הנתונים, טעות בתכנות. נדרש לזרות איך נראה ערך שגוי (נempt, או 0 או – או משהו אחר).

– בלי טיפול בערכים חסרים המודל יכול להיות גרוע. טיפול בערכים חסרים נקרא **data imputation** ואפשר לעשות את זה או על ידי השמת הממוצע של הפיצ'ר בערכים החסרים, אקסטרפולציה של הערך מפיצ'רים אחרים או השמת קבוע אחר וכו'.

– גם ערכים חשובים אך לא חסרים הם בעיה, לדוגמה אדם בן 3. כדי לטפל במקרים כאלה, לעיתים נדרש ידע בעולם התוכן של הבעיה כדי לטפל בסיטuatיות יותר מתאגרות.

דוגמה זרים עם בעיות בשחולות היא לא תופעה סבירה. יש מודלים שמאוד רובייטים כלפי דגימות שסוטות באופן חריג, והודות לשונות נזוכה. לעיתים יש פיצ'רים עם פורמט מסוים, ודגימות שלא עונות על הפורטט הזה.

- דוגמה** בשדה בינהי, תשובות כמו 0,1, negative (שלילי בעברית על אנגלית) צריכה לעבור נרמול לפורמט.
- את כל הפרי-פרוססינג מרים בקוד, כך שכל דגימה חדשה שנראה (גם הטסט), לפני שנעביר אותה למודל, נרץ אותה דרך הפרי-פרוססינג. אם לא נעשה זאת זה, הטסט לא יקבל את הטיפול שעשינו לאיומן ונקבל תוצאות גרוועות.
 - יצירת פיצ'רים חדשים מתוך DATA שאינו ב- \mathbb{R}^d היא שלב חשוב, למשל שיכון מילים כוקטוריים ב- \mathbb{R}^d .

• **EDA** (Exploratory Data Analysis) – שמיintended את תבניות בלי היפותזה ספציפית, כדי ללמידה הדאטא יותר טוב.

משתמשים בכל הכלים הסטטיסטיים שיש לנו – סטיטית תקון, חציוון, ממוצע, ערך נפוץ, שונות משותפת בין פיצ'רים. בנוסף שיטות אלו יאפשרו לנו לזהות אם הדאטא לא מאוזן.

כדי לבדוק מה המידע האפקטיבי אפשר להשתמש ב-PCA וקלאסטרינג ספקטורי.

• **Baseline** – מציאת מודל פשוט ונאייבי שמשיג ביצועים טובים, שימושו מושתת אותו מול מודלים מורכבים יותר ולראות אם הסיבוך שלהם באמת שווה את זה.

דוגמה עצי החלטה עמוקים מול רזדים.

האלג' הזה צריך להיות פשוט ולמיושן ולא להשיקע בו יותר מדי זמן, מטרתו לשערך כמה קשה לפתור את הבעיה. גם כישיש מודל מורכב, שווה להוריד כל פעם רכיב אחד מהסיבוך ולראות כמה כל סיבוך מוסיף לנו ביצועים.

. Baseline : שיפור ביצועי **Model Selection**

להוסיף, קרנלים, רגולרייזציה, כמה מהשיטות הנ'ל וכו'. בנוסף, בשלב הזה צריך למצוא היפר-פרמטרים טובים באמצעות CV או שיטת שערוך ביצועים אחרים.

שיקולים חישוביים

את כל השלבים הקודמים צריך לעשות עם כמות דוגימות מאוד גדולה לפחות (נגיד מיליון דוגימות) וזה מביא אליו קשיים. הרבה כלים ופרדיגמות עוזרים לפתור את זה אבל לא לומדים אותם בשום מקום. כמה דוקטוררים ב-MIT שמו לב בעיה זו וייצרו קורס שמכיל כל מה שצריך לדעת, הוא נקרא *The Missing Semester*.

דוגמה לשות פרי-פרוססינג שעובר על כל שורה בקובץ ומנקה אותו ושם אותו בראשימה. זה רעיון גרווע כי זה אומר שאנו שומרים את כל הדאטא בזיכרון אבל אין מספיק מקום. במקרה זאת עדיף לעשות lazy evaluation – לחשב את הערך רק כשצריך אותו, ואז ליצר אותו בפייתון שמחזיק רק שורה אחת בכל פעם כי לא צריך כמה בכל נקודת זמן (ואז אפשר לכתוב כל ערך לקובץ).

ביצוע פעולות מתמטיות באופן ממוקבל היא קריטית לביצועים טובים, שכן אין סיבה לעשות אותם באופן סדרתי באמת.

דוגמה פרדיגמת map-reduce (שלומדים עליה בין היתר ב-OS, Hadoop ו-Spark) עוזרת למקבול (וחבילות רבות משתמשות בכך).

כשבוצעים את החישובים האלה, חשוב למסס את המיקבול. בمبرדים מספר הפעולות במקביל חסום ע"י מספר לכל היותר דו-ספרתי של ליבות, אבל שימוש בקרטיסי מסך מאפשר פעולה בו זמנית (קרטיסי מסך במקור שומש לגרפיקה).

مبرדים חדשים גם מפותחים כדי להיות טובים ל-ML ואפשר بكلות כתוב קוד ל-GPU-ים באמצעות חבילות כמו `pytorch`, `pandas`, `sklearn`, `matplotlib`, `numpy`, `matplotlib` לא לכתוב אף פעם לומדים בעצמו - מישחו כבר כתוב את הקוד טוב יותר ויעיל יותר באחת מהספריות.

`tqdm`, `logging`.

דיווח תוצאות

חשוב להבין איפה המודל מצליח ואייפה המודל לא מצליח - אנחנו אף פעם לא פותרים את הבעיה במלואה. העבודה שהוא לא מצליח באזורי מסויימים היא לא בחרחה רעה - היא פתח למחקר עתידי.

יש כמה דרכים להציג תוצאות:

- לצייר בגרף את השגיאות של מודלים שונים תחת גודל מודם שונה, סיבוכיות מודל שונה, היפר-פרמטרים שונים.
- לצייר `confusion matrix` שומרה כמה מובלבים בין ליibiliים שונים (לדוגמא כמה ניחשו "כחול" אבל בעצם היה "ירוק").
- לדוגם דוגמאות שטיענו עליהם ולראות איפה המודל נכשל ולהסתכל עליהם ידנית.
- ליצור `demo` אינטראקטיבי של המודל שמאפשר בדיקה של בן אדם, נגיד גם עם היפר-פרמטרים שונים. אפשרות ליצור דמו כזה מאוד بكلות.

שחזור תוצאות

כדי לאפשר אחרים לשחזר את אותן התוצאות, חשוב לטע את היפר-פרמטרים. ב-`k-means` לדוגמה התוצאות תלויות ב-`seed` (שגם אותו אפשר לחת), וגם מעבר לזה באקרניות של GPU שבה אי אפשר לשנות. לכן חשוב גם למצער ביצועים בדיווח התוצאות כדי שנitin יהיה לשחזר התנהגות טיפוסית של המודל.

מומלץ להשתמש CLI שמקבל היפר-פרמטרים וקבצים בשבייל קוד יותר מודולרי.

לא להשתמש ב-`Jupyter Notebooks` לפיתוח מודלים - כל דבר מעבר להוראה וויזואליזציה ראוי לעשות ב-IDE IDE לכשה כי קשה מאוד לשחזר תוצאות ממש. לא להשתמש ב-IDE כבד כי הוא עושה הרבה עבודה בשביבנו שקשה אחרי זה להסביר אחרים איך הגיעו לאותו המצב (ניחול חבילות וכו'), מקום זאת, דמיינו שאתם בשנות ה-80' ותשימושו ב-`tomz`.

מודלים גדולים נקיים רק לחברות גדולות עם הרבה משאבים וمبرדים קטנות לא יכולות להרשות לעצמן. יש תחום שלם שנקרא AI Green AI שמנסה להנגיש מודלים חזקים לאנשים עם תקציב קטן.

תרגול

התרגול הוא חזרה אינטנסיבית על מה שעשינו בהרצאה זו שלפניה ולכן לא כולל את כל הדוגמאות וההypothesis שוב.

בשיטת קרNEL אנחנו מנסים ללמוד בעיה במרקח אחר שבתקווה יאפשר למידה יותר איקוית של הדאטא, כלומר במקומות שכלל ההחלטה יהיה

$$.k \gg d \text{ כאשר } \mathbb{R}^d \xrightarrow{\psi} \mathcal{F} \xrightarrow{\text{קרנייזיה}} \mathbb{R}^m, \mathcal{X} \rightarrow \mathbb{R}^m$$

משפט הkernel טרייק מאפשר לנו לפתור את הבעיה

$$w^* = \underset{w \in \mathcal{F} = \mathbb{R}^k}{\operatorname{argmin}} f((\langle w | \psi(x_i) \rangle)_{i=1}^n) + \lambda \|w\|_2^2$$

(שהיא מאוד קשה כ- k -גadol, אבל רצואה כי \mathcal{F} טוב למידה לטענתנו) באמצעות פתרון הבעיה

$$\alpha^* = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} h(G\alpha) + \lambda \alpha^T G\alpha$$

והצבת w^* . את הבעיה של α^* קל לנו לפתור כי זו בעיה קמורה וברטQP.

דוגמה בירידג' למדנו ע"י פתרון הבעיה (עם המעבר כבר ל- $\mathcal{F} = \mathbb{R}^k$) וזו אכן בעיה $\underset{w \in \mathbb{R}^k}{\operatorname{argmin}} \sum (y_i - \langle w_i | \psi(x) \rangle)^2 + \lambda \|w\|_2^2$ בצורה של המשפט הנ'ל ולכן נגרען,

$$\begin{aligned} \sum_i (y_i - \langle w_i | \psi(x_i) \rangle)^2 + \lambda \|w\|_2^2 &= \sum_i \left(y_i - \left\langle \sum_j \alpha_j \psi(x_j) | \psi(x_i) \right\rangle \right)^2 + \lambda \langle w | w \rangle \\ &= \sum_i \left(y_i - \sum_j \alpha_j \langle \psi(x_j) | \psi(x_i) \rangle \right)^2 + \lambda \langle w | w \rangle \\ &= \sum_i \left(y_i^2 - 2y_i \sum_j \alpha_j \langle \psi(x_j) | \psi(x_i) \rangle + \left(\sum_j \alpha_j \langle \psi(x_j) | \psi(x_i) \rangle \right)^2 \right) + \lambda \langle w | w \rangle \\ &= \|y\|^2 - 2y^T G\alpha + \alpha^T G^2 \alpha + \lambda \alpha^T G\alpha \end{aligned}$$

כאשר $[G]_{ij} = \langle \psi(x_i) | \psi(x_j) \rangle$. נמצא פתרון סגור,

$$\nabla_\alpha f(\alpha) = -2Gy + 2G^2\alpha + 2\lambda G\alpha = 0$$

$$G^2\alpha + \lambda G\alpha = Gy$$

$$G(G + \lambda I)\alpha = Gy$$

$$(G + \lambda I)\alpha = y$$

ולכן עבור \tilde{X} מטריצה ששורתה $\psi(x_i)$ מתקיים $G = \tilde{X}\tilde{X}^T$ שהוא PD ולכן y מוגדר ע"י $\alpha^* = (G + \lambda I)^{-1}$

$$\hat{y}(x) = \langle w^* | \psi(x) \rangle = \left\langle \sum_i \alpha_i^* \psi(x_i) | \psi(x) \right\rangle = \sum_i \alpha_i^* \langle \psi(x_i) | \psi(x) \rangle$$

כלומר צריך רק את המכ' "באים של ψ על דוגמאות, ולא ψ , ואנו אנחנו לומדים באמצעות הדמיון בין ה- ψ על הדוגמאות.

דוגמה נקרנל PCA. לכארה אנחנו מעריכים מינימום כדי להוריד מינימום, אבל זה עוזר כי את הורדת המינימום אנחנו עושים רק לאחר שעיווותנו את הדואט אל תוך ת"מ לינארי, שזה די מגביל. עם קרナル, העיוות הוא לא רק אל תוך ת"מ לינארי אלא משחו יותר כלילי ולכון אפשר לתאר תופעות יותר טוב.

ב-PCA יהיה לנו $C = \frac{1}{m} X^T X = \frac{1}{m} \sum x_i x_i^T$ (נניח שמדובר ב- $X \in \mathbb{R}^{m \times d}$) והוא ניקח את k ה"עומדים" של C ואלו יהיו בסיס לת"מ שלנו, משם פשוט נטיל אורתוגונליות.

יהי $0 \neq \lambda$ העומדים ל- v -ב- C . כלומר $v \in \text{Im } C = \text{sp} \{x_1, \dots, x_m\}$ ו- λv שקול לממצא v כך ש- λ מושג (זה ליניארית 2, לא כזו מעניין).

$$v = \sum \alpha_i x_i \quad (\text{lכון } \alpha_i = \frac{1}{\lambda} \langle x_i | v \rangle) \quad . \quad \text{nסמן } v = \frac{1}{\lambda} \sum x_i \langle x_i | v \rangle \quad \text{ולכן } \lambda v = Cv = X^T X v = \sum x_i x_i^T v$$

$$\begin{aligned} \lambda \left\langle x_i \mid \sum_j \alpha_j x_j \right\rangle &= \left\langle x_i \mid C \sum \alpha_j x_j \right\rangle \\ &= \sum \alpha_j \langle x_i | C x_j \rangle \\ &= \sum \alpha_j \left\langle x_i \mid \sum_l x_l x_l^T x_j \right\rangle \\ &= \sum \alpha_j \sum_l x_i^T x_l x_l^T x_j \end{aligned}$$

ולכן $\langle x_i \mid x_j \rangle = G \alpha = G^2 \alpha = G \lambda \sum \alpha_j \langle x_i | x_j \rangle = \sum_j \alpha_j \sum_l \langle x_i | x_l \rangle \langle x_l | x_j \rangle$ כאמור $\lambda \alpha = G \alpha$ כאשר G היא מטריצת PD, לכן כדי לפטור PCA צריך לממצא v על-ל- G .

ב-PCA מצאנו ל- $X X^T$ את $C = X^T X$ את העומדים $\lambda_1, \dots, \lambda_d, v_1, \dots, v_d$ בהתאם למקרה הכלורי. נטיל את x באופן הבא

כדי לעשות PCA צריך להטיל ולשכן את הדוגימות. נטיל את x באופן הבא

$$\tilde{x}_l = \langle v^{(l)} | x \rangle = \sum_{i=1}^m \alpha_i^{(l)} \langle x_i | x \rangle$$

כasher $v^{(l)}$ הוא העומד של C (המוצג על- G).

הגדרה נאמר כי $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ הוא קרナル PSD כאשר $[K]_{ij} = k(x_i, x_j)$ והו $m \in \mathbb{N}$ כל $i, j \in \mathcal{X}$ 。

משפט (תנאי מרסר) תהי $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}$ כasher ψ הוא מרחב הילברט. אז קיימת פ' סימטרית $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ שהיא מכ' ב- \mathcal{F} אם \mathcal{F} PSD.

היא קרナル PSD, ובפרט G היא PSD כאשר $[G]_{ij} = k(x_i, x_j) = \langle \psi(x_i) | \psi(x_j) \rangle$

דוגמה הקרナル הגaussיאני הוא קרナル PSD, המוגדר ע"י $k(x, x') = \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right)$ והוא מושרחה ע"י

$$\psi(x)_n = \frac{1}{\sqrt{n!}} \exp\left(-\frac{x^2}{2\sigma^2}\right) x^n$$

לכל $n \in \mathbb{N}$

דוגמה נשים לב שלא $k = \binom{d+n}{n}$ ואז $k(x, x') = (1 + \langle x | x' \rangle)^k$ היא פונקציית שמשרת $x \mapsto \left(1, \dots, x_i, x_i x_j, \dots, \prod_{i \in J: J \subseteq [d], |J|=k} x_i\right)$. אceptת לנו בעצם מה ψ כי הצלחנו לחשב את k ביעילות.

שבוע Gradient Descent | XIV

הרצאה

למד איך עובדים הפורנים של בעיות קמורות, אבל לפניכן למד על פ' קמורות ובעיות קמורות.

דוגמה הלמידה במזעור LS, מרחק אבסולוטי (שניהם בריגרסיה לינארית), SVM, ריגרסיה לוגיסטיבית ועוד ועוד הן כולן בעיות קמורות.

פונקציות קמורות

הגדירה נאמר כי C היא קבוצה קמורה אם לכל $\alpha \in [0, 1]$ ומתקיים $\alpha u + (1 - \alpha) v \in C$ וכל $u, v \in C$.

הערה לא אתן אינטואיציה לקבוצות קמורות ופ' קמורות, כבר שבוע 13, ראיינו מספיק.

דוגמה כדור היחידה ביחס לכל נורמה הוא קמור. על-מישור ועל-מרחב הם קמורים, ת"מ אפיניים $\{x : Ax = b\}$ קמורים.

אוסף כל מטריצות ה-PSD היא קמורה. $\mathbb{S}_+^n \subseteq \mathbb{R}^{n^2}$

משפט יהיו $C, D \subseteq \mathbb{R}^d$ קבוצות זרות. אזי קיימים על מישור שמספריד ביניהם, ובמפורש, קיימים כך ש-

$\{a^T x \geq b\}$ וגם

טענה יהיו $C, D \subseteq \mathbb{R}^d$ קבוצות קמורות אזי הקבוצות הבאות גם קמורות:

$$C \cap D \quad .1$$

$$\{ax + b : a \in \mathbb{R}, b \in \mathbb{R}^d, x \in C\} \quad .2$$

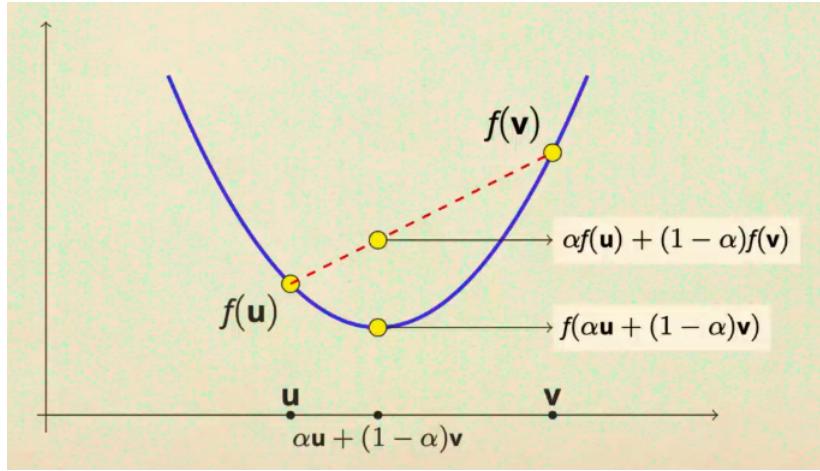
$$f(C), f^{-1}(D) \quad .3$$

$$f(x) = \frac{Ax+b}{c^T x+d} \text{ כאשר } f(C), f^{-1}(D) \quad .4$$

הגדירה תהי $C \subseteq \mathbb{R}^d$ קבוצה קמורה. נאמר כי קמורה אם $\forall u, v \in C$ $f : C \rightarrow \mathbb{R}$

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v)$$

הערה ראו ויזולאייזציה,



טענה f היא קמורה אם האפי-גרף שלה קמור “מעל” גרף הפ'.

דוגמאות

הפ' הבאות מ- $\mathbb{R}^d \rightarrow \mathbb{R}$ הן קמורים.

$$x \mapsto e^{ax}, x \mapsto x^\alpha, x \mapsto -\log x .1$$

$$.2. \text{ פ' אפינית } x \mapsto w^T x + b$$

$$.3. \text{ פ' ריבועית } x \mapsto x^T A x + w^T x + \alpha$$

$$.x \mapsto \|y - Ax\|_2^2, \text{Sum of Squares } .4$$

$$.x \mapsto \|x\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{\frac{1}{p}} .5. \text{ כל נורמות } \ell_p \text{ המוגדרות ע"י}$$

$$.x \mapsto \|x\|_\infty = \max_i |x_i| .6. \text{ נורמת מקסימום}$$

$$.7. \text{ פ' אינדיקטור על קבוצה קמורה.}$$

טענה תהי f גזירה. אז f קמורה אם $\text{dom}(f) \subseteq \mathbb{R}^d$ הוא קמור ולכל

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1)$$

(כלומר f מעלה הקירוב בלינארי שלה מכל נקודה אחרת).

טענה תהי f גזירה פעמיים. אז f קמורה אם $\text{dom}(f) \subseteq \mathbb{R}^d$ הוא קמור וכן $\nabla^2 f(x) \in \mathbb{S}_+^n$ (ההסיאן הוא PSD לכל

פעולות על פ' קמורים

1. אם f_1, \dots, f_n קמורות אז $\sum_i \alpha_i f_i$ קמורה וגם $\max_i f_i$ קמורה.

2. תהי $g : \mathbb{R}^{d+k} \rightarrow \mathbb{R}$ המקיימת $g(x_1, x_2) \mapsto g(x_1, x_2)$ היא קמורה אז $(x_1, x_2) \mapsto g(x_1, x_2)$ היא קמורה.

3. אם f קמורה אז $x \mapsto f(Ax + b)$ קמורה.

4. אם g קמורה ו- h קמורה ומונוטונית עולה אז $x \mapsto h(g(x))$ קמורה.

5. אם h קמורה ומונוטונית עולה בכל אחת מהמשתנים שלה ו- g_i קמורות אז $f(x) = h(g_1(x), \dots, g_k(x))$ קמורה.

זוגמה נגידר $f(x) = \log\left(\sum_{i=1}^k e^{w_i^T x + b_i}\right)$ זו נראית log-sum-exp . הסטודנטית המשקיפה תוכיה שזו פ' קעורה (באמצעות הרכבה אפינית על קמורה וכן קרייטריון קמירות לפ' גזירות פעמיים).

משמעותו כי $f(w) = \sum_{i=1}^m y_i \langle x_i | w \rangle + \log\left(1 + e^{w^T x_i}\right)$

תרגיל הוכיחו כי $f(w)$ היא קעורה (כלומר $-f$ היא קמורה).

טענה אם f קמורה אז כל מינימום מקומי של f הוא גם מינימום גלובלי (זה גרעין היתרון של בעיות קמורות).

הוכחה: נגידר $B(u, r) = \{v : \|v - u\| \leq r\}$ והוא u מינימום מקומי של f , כלומר קיימים $0 < r < \text{שלכל}$ $v \in B(u, r)$ כך $f(v) \geq f(u)$. לכן $v \in B(u, r)$ ו- $f(v) > f(u)$.

$$f(u) \leq f(u + \alpha(v - u))$$

אם f קמורה, אז

$$f(u + \alpha(v - u)) = f(\alpha v + (1 - \alpha)u) \leq (1 - \alpha)f(u) + \alpha f(v)$$

ושה"כ קיבלנו $f(v) \leq f(u)$ וכן u מינימום גלובלי של f .

מסקנה אם אנחנו מצליחים לרדת למינימום מקומי כלשהו בפ' קמורה, הגענו למינימום הגלובלי.

תזכורת קירוב הטילור מסדר 1 של f ב- w ביחס ל- u הוא

$$f(u) \approx f(w) + \langle \nabla f(w) | u - w \rangle$$

טענה אם f קמורה וגירה אז $f(u) \geq f(w) + \langle \nabla f(w) | u - w \rangle$ לכל u (לא רק בסביבה).

תת-גרדיינטאים

הערה אם f קמורה אבל לא גזירה, אין לנו גרדיאנט אבל עדיין נרצה משזה שיישמש כאחד כזה.

הגדרה נאמר כי v הוא תת-גרדיינט של f ב- w אם לכל u

הערה v הוא תת-גרדיינט אם הוא מותח לגרף הפ' בכל נקודה.

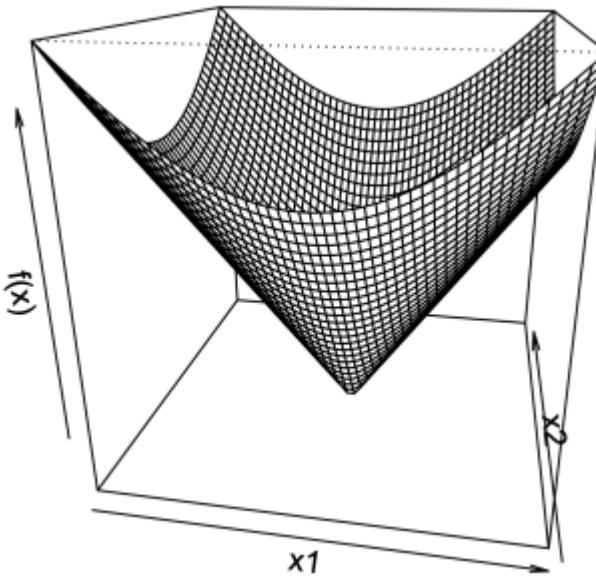
דוגמה עבור $x \in \mathbb{R}$, $f(x) = |x|$ בנקודת $0 = x$ אין גרדיאנט אבל כל ישר עם שיפוע שהוא מותח לגרף הפ' הוא תת-גרדיינט, בפרט

$$\partial f(x \neq 0) = \nabla f(x \neq 0) = \text{sign}(x)$$

הגדרה התת-דיפרנציאל (m) של f הוא אוסף כל התת-גרדיינטים ב- w של f .

טענה f קמורה אם ו רק אם $\partial f(w) \neq \emptyset$.

דוגמה נורמת ℓ_2 היא לא גזירה ב-0 (ראו אייר).



בכל נקודה שאינה $(0,0)$ מתקיים $\partial f(x) = \{\nabla f(x)\} = \left\{ \frac{x}{\|x\|_2} \right\}$,

הסתודנטית המשקיפה תוכיח זאת ותבין זאת לעומק.

דוגמה נורמת ℓ_1 היא לא גזירה ב-0 (כמו התמונה הנ'ן, רק לא עגול אלא מישורים ישרים).

$$\partial f(x) = \begin{cases} \{\text{sign}(x_1), \dots, \text{sign}(x_d)\} & x \neq 0 \\ [-1, 1]^d & x = 0 \end{cases}$$

תרגיל יהיו $f_1, f_2 : \mathbb{R}^d \rightarrow \mathbb{R}$ קמורות וגזירות ו- $f(x) = \max\{f_1(x), f_2(x)\}$. חשבו את $\nabla f(x)$ כאשר $f_1(x) > f_2(x)$.

$$f_1(x) = f_2(x) \quad \text{ולפ' } f_1(x) < f_2(x)$$

טענה $\partial f(x)$ היא קבוצה סגורה וקמורה (גם לפ' לא קמורות).

טענה אם f גזירה אז $\{\partial f(x)\} = \{\nabla f(x)\}$ ואם $\partial f(x)$ הוא סינגולטון לכל x אז f גזירה.

הערה יש תורה אינפי שלמה לתת-גרדייאנטים בדיק כmo שיש לגרדייאנטים (נגזרות).

טענה (ארכיטקטיקה של תת-גרדייאנטים) תהי f פ' אזי:

$$\alpha > 0 \text{ עבור } \partial(\alpha f) = \alpha \partial(f) .1$$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2 .2$$

$$.3 \text{ עבור } \partial g(x) = A^T \partial f(Ax + b) \text{ מתקיים } g(x) = f(Ax + b)$$

כל השוויונות הנ'ל הם שוויונות של קבוצות.

דוגמה ה loss של Soft-SVM הוא

$$f(w) = \lambda \|w\|_2^2 + \sum_{i=1}^m \ell^{hinge}(w, (x_i, y_i))$$

כאשר $\ell^{hinge}(w, (x, y)) = \max\{0, 1 - y \langle w | x \rangle\}$.

תרגיל יהי (x, y) ונדריך פ' w כך ש- $y \langle w | x \rangle \geq 1$ כאשר $-yx \in \partial f(w)$.

$$.y \langle x | x \rangle < 1$$

$$.v = 2\lambda w - \sum_{i:y_i \langle w | x_i \rangle < 1} y_i x_i \text{ כאשר } v \in \partial f(w)$$

אופטימיזציה קמורה

הגדרה בעיה קמורה היא בעיה מהצורה $Ax = b$ כאשר f_i קמורות ו-

$$D = \text{dom}(f) \cap \left(\bigcap_{i=1}^n \text{dom}(f_i) \right)$$

נקראת המטרה, f_i נקראים פ' אילים.

נקודה פיזibilite (feasible) היא נקודה $x \in D$ כך ש- $f_i(x) \leq b_i, \forall i$ (עונה על כל האילים).

אם x היא פיזibilite ומוגה $f(x) = f^*$ כאשר f^* היא הערך האופטמלי של פ' המטרה, היא תקרא נקודה אופטימלית או ממצעתה.

טענה אוסף כל הפתרונות של f הוא קבוצה קמורה.

הגדרה פוטר בעיות קמורות הוא אלג' (shmomes בתוכנה) שמקבל כקלט בעיה קמורה ופלט פתרון אופטימלי.

פתרונות מחולקים לשיטות מדרגה ראשונה (שuttlems רק ב-(תת-)גרדייאנט, ביגנים Gradient Descent) ושיטות מדרגה שנייה (שuttlems ביחסיאן).

הערה לרוב הבעיות שראינו יש פותרנים ייעודיים ולא גנריים שעובדים אפילו יותר טוב על האילוצים הספרטיפיים.

האם היהות בעיית למידה שקופה לבעה קמורה מספיק כדי שהיא תהיה למידה PAC?

דוגמה כ- \mathcal{H} היא מחלקת ה

על מישורים לסיוג,
 $HS_d = \{x \mapsto \text{sign}(\langle w | x \rangle)\}$, פתרנו בעיה קמורה כדי ללמידה.

הגדרה תהי $\mathcal{Y} \times \mathcal{X} = Z$. בעית למידה (\mathcal{H}, Z, ℓ) נקראת קמורה אם \mathcal{H} היא קבוצה קמורה ולכל $(x, y) \in Z$ הינה פ' קמורה על \mathcal{H} .

הערה פתרנו בעית למידה עם ERM הוא פתרון בעיה קמורה, $\min_{w \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \ell(w, (x_i, y_i))$.

דוגמה ברגירסיה לינארית, $\ell(w, (x, y)) = (\langle w | x \rangle - y)^2$.

טענה לא כל בעיות הלמידה הקמורות הן למידות PAC. אבל אם \mathcal{H} חסומה ופ' loss היא לפישיצית אז הבעיה היא PAC-למידה.

הגדרה נאמר כי $f : C \rightarrow \mathbb{R}$ היא ρ -ლיפשיצית אם לכל $w_1, w_2 \in C$ מתקיים $|f(w_1) - f(w_2)| \leq \rho \|w_1 - w_2\|$.

תרגיל תהי f קמורה. הראו כי f היא ρ -ლיפשיצית אם "נורמת כל תתי-הגרדיאנטים היא לכל היותר ρ (אינטואטיבית קצב השינוי לא עולה על ρ ואז f לא יכולה להתרחק יותר מ- ρ כפול הפרש ה- x -ים).

הגדרה בעית למידה (\mathcal{H}, Z, ℓ) נקראת קמורה-ლיפשיצית-חסומה עם פרמטרים B, ρ אם מתקיים:

• \mathcal{H} היא קבוצה קמורה ולכל $w \in \mathcal{H}$ מתקיים $\|w\| \leq B$.

• לכל $(x, y) \in Z$ היא קמורה ו- ρ -ლיפשיצית.

משפט כל בעית למידה קמורה-ლיפשיצית-חסומה היא PAC למידה עם סיבוכיות מדגם שתלויה ב- δ, ϵ, ρ .

Gradient Descent

נראה שיטה פשוטה מסדר ראשון לפתרון בעיות קמורות.

סיבות למידת GD

1. אם אנחנו עובדים עם מטרה מوزה, נצטרך לכתוב בעצמנו פתרון בעיה ו-GD הוא גישה פשוטה לכך.

2. אם אנחנו עובדים עם דאטא-סט ענק, יכול להיות שנצטרך לכתוב פתרון בעיות קמורות מבזבז או GD/SGD זה גישות פשוטות.

כש- $d = 10^9$ ו- $m = 10^{12}$ נדרש לכתוב תשתיות עצמן (יש בעיות כאלה) ואין יותר מדי Open Source לוזה כי זה תלוי בחומרה שיש לנו והרשת וכו'.

3. במקרה של למידת אונליין נדרש לכתוב פתרון עצמן, SGD הוא דרך טובה לוזה.

4. כל Deep Learning מבוסס על SGD.

בחדו"א למדמ"ח או אינפי 2 לממד"ח או בוקיפדייה או בשום מקום, ראיינו ש- ∇f מצביע בכיוון שבו f גדלהopi הכי הרבה מהנקודה הנוכחית ו- ∇f – בכיוון שבו f קטנהopi הכי הרבה.

הגדירה קו הגובה של f מוגדר ע"י $\{x' : f(x) = f(x')\}$.

טענה אם f גזירה ב- x ו- w הוא משיק ל- $\nabla f(x) | w$ כלומר $\nabla f(x)$ מאונך למסלנות שלא משנה את ערך הפ'.

טענה (אופטימליות מסדר ראשון) עבור בעיה קמורה ב- f עם אילוצים C , נקודת פיזיבילית x היא אופטימלית אם "ם"

$$\langle \nabla f(x) | y - x \rangle \geq 0$$

כלומר, לכל h צעד מ- x , אנחנו הולכים קצת בכיוון ∇f וرك מגדילים את הערך (או לא משנהים אותו אם המכ"פ 0).

דוגמה עבור $C = \mathbb{R}^d$ (כלומר אין אילוצים), x הוא אופטימלי אם $\nabla f(x)$ ניצב ל- y בכל נקודת, כלומר, כלומר 0 .

תרגיל תהי המטרה c $f(x) = \frac{1}{2}x^T Qx + b^T x + c$ כאשר Q היא PSD. הראו כי f גזירה ו- $b \notin \text{Im } Q^{-1}$ אז אין

פתרון אופטימלי (כי f לא חסומה מלמטה) ואם Q לא הפיכה אבל $b \in \text{Im } Q$ אז $x = Q^\dagger b + z$ הוא פתרון אופטימלי.

הערה אם x לא אופטימלי, אז קיים h כך $\langle \nabla f(x) | h \rangle < -\langle \nabla f(x) | h \rangle < 0$ כלומר אפשר לצוד למקומן יותר נמוך בכיוון (ולא בהכרח במקביל אבל בכיוון) $-\nabla f(x)$.

זה בדוק מה שהאלג' שלנו יעשה, באופן איטרטיבי, רק שצורך לדעת מה הצעדים האלה. אם הצעד גדול מדי, נצא מקבוצת הפיזיבילים, אם עשינו צעד קטן מדי, התקדםנו קצת מאוד והרבה מאוד זמן לא נגיע לפתרון.

אלג' GD

1. נתחל $x^{(1)} \in \mathbb{R}^d$ כלשהו.

2. באיטרציה ה- t , נעדכן $x^{(t+1)} = x^{(t)} - \eta_{t+1} \nabla f(x^{(t)})$.

3. נעצור ב- T כלשהו, לדוגמה כאשר $\| \nabla f(x^{(t)}) \| < \epsilon$.

4. נחזיר את הוקטור האחרון $x^{(T)}$, או את $\bar{x} = \frac{1}{T} \sum_t x^{(t)}$.

הערה η_t נקראים gradient step sizes.

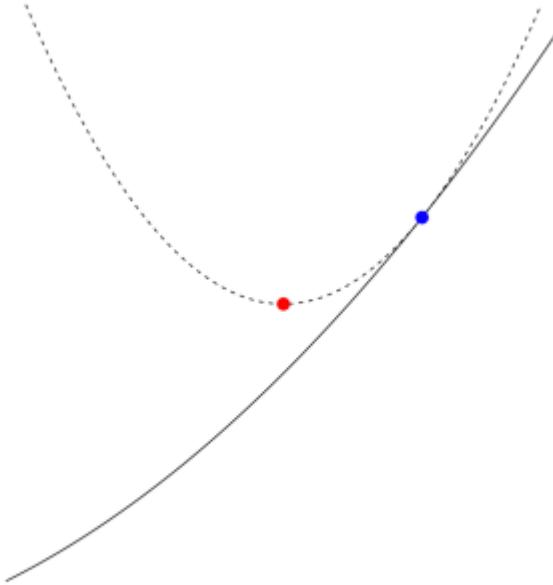
נבין אינטואטיבית איך GD עובד. נזכיר כי קירוב טילור של f מסדר 2 הוא

$$f(y) \approx f(x) + \langle \nabla f(x) | y - x \rangle + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x)$$

עכשו, GD הוא שיטה מסדר ראשון ולכן לא יודע האם f גזירה פעמיים ולא יודע מההסיאן ולכן הוא מקרב אותו נאיבית באמצעות

$$\nabla^2 f(x) = \frac{1}{\eta}$$

תרגיל הראו כי המזער y של הקירוב הריבועי הנאיבי ה"ל, או במפורש $f(y) \approx f(x) + \langle \nabla f(x) | y - x \rangle + \frac{1}{2\eta} \|y - x^{(t)}\|^2$, שמשמער את הקירוב מסדר שני (ראו איור) $x^{(t+1)} = x - \eta \nabla f(x)$.



Blue point is x , red point is

$$x^+ = \underset{y}{\operatorname{argmin}} \quad f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

בחירה η קבוע, לא נצליח לתפוס תופעות עדינות ואף פעם לא נכנס, אם η קטן מדי נידרש להרבה מאד איטרציות ואם הוא גדול מדי כל הזמן נקרב ונתרחק ממרכז הגולף באנלוגיה לגולף (שעכשו המצתתי).

Backtracking line search

BLS בוחר דינמית את η . נסמן $f(x)$ ו- $\Delta x = -\nabla f(x)$.

- אם $\alpha > 0$ והוא קטן ממספר או $f(x) + \alpha \eta \langle \nabla f(x) | \Delta x \rangle > f(x + \eta \Delta x)$

- אם הוא גדול ממספר או $f(x) + \alpha \eta \langle \nabla f(x) | \Delta x \rangle > f(x + \eta \Delta x)$

לכן ממשפט ערך הביניים קיימים $\eta_0 < \eta < \eta_1$ כך $f(x) + \alpha \eta_0 \langle \nabla f(x) | \Delta x \rangle = f(x + \Delta x)$, אותו אנחנו רוצים.

כיצד נמצא את η_0 זהה? נקבע עוד פרמטר $\beta \in (0, 1)$ וначילה מ- $t = 1$ וכל איטרציה נחשב $\beta t \mapsto t$ ונווחר על התהיליך הזה עד ש-

$$f(x) + \alpha \eta \langle \nabla f(x) | \Delta x \rangle < f(x + \eta \Delta x)$$

משמעות אם f היא קמורה וגזירה, ו- ∇f^* לפשיטית עם קבוע L ו- x^* אופטימלי ב- f אז GD עם צעדים קבועים η מקיים

$$f(x^{(t)}) - f^* \leq \frac{\|x^{(0)} - x^*\|_2^2}{2\eta t}$$

כלומר ב- $t \rightarrow \infty$ התכנסנו.

תרגילי תכנות

- כתבו פתרון GD לבעיה $f(x) = \frac{1}{2}x^T Qx + b^T x + c$ כאשר Q, b, c ארגומנטים.

- וודאו את תקינות הקלט.

- קובלו כפרמטר את η , וכברירת מחדל השתמשו ב-BLS לקביעת הצעד.

- השתמשו ב- ϵ כ- $\|\nabla f(x^{(t)})\| < \epsilon$ פרמטר.

- השוו את דיק הפלטים שראינו (אחרון, הכיתוב, ממוצע).

האם הפתרן שלכם סובל מבעיות כישיש ע"ז?

בתרגיל זה ובבאים, ציררו לכל איטרציה את η_t והשו על היפר-פרמטרים שונים. לפעמים קורה ש- $x^{(t)}$ מרחוק מ- x^* אבל עדין $f(x^{(t)})$ קרובה ל- $f(x^*)$.

להלן הקוד של CVX, פתרון גנרי ב Matlab לבעיות קמורות.

```
m = 20; n = 10; p = 4;
A = randn(m,n); b = randn(m,1);
C = randn(p,n); d = randn(p,1); e = rand;
cvx_begin
    variable x(n)
    minimize( norm( A * x - b, 2 ) )
    subject to
        C * x == d
        norm( x, Inf ) <= e
cvx_end
```

- כתבו פתרון יודי לריגרסיה ליניארית/רידג' באמצעות פתרון ה-QP הנ"ל שכתבתם, שמקבל בנוסף פרמטר רגולרייזציה λ .

- מדדו את הדיק שלכם מול הפתרון הסגור של רידג', עם כל הגרפים הנ"ל.

- האם הפתרון מושפע מ- λ ? לים בעמודות של X : האם הוא עובד אחרת עם $\lambda > 0$ ו- $\lambda = 0$? (כנראה שכן).

מוטל GD

עד כה עסקנו ב-GD בלי אילוצים, כיצד נכליל זאת לבעיה קמורה עם אילוצים?

הגדירהichi אופרטור ההטלה על האוסף הפיזיבילים C המוגדר ע"י

$$P_C(x) = \underset{w \in C}{\operatorname{argmin}} \|x - w\|_2$$

שהוא לא בהכרח הטלה אורתוגונלית.

הערה נשתמש ב- P_C לפתרו הכללי, אבל גם חישוב P_C הוא בעיה קמורה ולכן במקרה נפתרו עוד בעיה קמורה.

עתה באיטרציה ה- t -של GD מוטל, נחשב $x^{(t+1)} = P_C(x^{(t)} - \eta_{t+1} \nabla f(x^{(t)}))$ שבסך פעם אנחנו מティילים חוזרת אל תז' C .

הערה GD מוטל הוא יעיל אם אנחנו יכולים לחשב במהירות את $P_C(x)$.

Sub-Gradient Descent

מה נעשה ב-GD אם הגענו לנקודת גזירה? נזכיר כי f קמורה אם $\partial f(x) \neq \emptyset$ ולכן אם אנחנו יכולים למצוא איבר אחד בתת-דיפרנציאל (תת-גרדיינט כלשהו), נשתמש בו במקום הגרדיינט ויש לנו לפחות שעבוד טוב (לא כמו GD, אבל לפחות מואוד טוב).

טענה תהי f קמורה, אז $f(x^*) = \min f(x) \in \partial f(x^*)$ (נובע ישירות מהגדרת התת-גרדיינט).

תרגיל תהי f גזירה וקמורה ותהי הבעיה הקמורה $\min_{x \in C} f(x)$. הסיקו את תנאי האופטימליות מסדר ראשון מתנאי האופטימליות של תת-גרדיינטנים של הבעיה השcoleה $\min_{x \in C} f(x) + \mathbb{1}_C(x)$.

דוגמה בלאסנו מזערנו $\min_{w \in \mathbb{R}} \frac{1}{2} \|y - Xw\|_2^2 + \lambda \|w\|_1$ שבו y הוא אופטימלי אם $w \in \partial f(w^*)$

$$0 \in \partial \left(\frac{1}{2} \|y - Xw\|_2^2 + \lambda \|w\|_1 \right)$$

אם "

$$0 \in -X^T(y - Xw) + \lambda \partial \|w\|_1$$

$$\text{אם } v \in \partial \|w\|_1 \text{ ש-} v \in -X^T(y - Xw) + \lambda v$$

$$v_i \in \begin{cases} 1 & w_i > 0 \\ -1 & w_i < 0 \\ [-1, 1] & w_i = 0 \end{cases}$$

בסדר זהה כרגיל מוגן, אבל כאשר ציר אחד מתאפס, לא רק $v = 0$, פתאום התת-גרדיינט בנקודת חופשי לנوع ב- $[1, -1]$ רק

בקוורדינטת המאופסת. שכנו עצמכם שזה נכון (דמיינו את השפיצים ב- $x_1 = 0, x_2 = 1$ וכו').

יחד נקבל ש- w אופטימלי לאלסו אם "מ עבר C_i עמודות X

$$\langle C_i | y - Xw \rangle = \lambda \text{sign}(w_i)$$

$$\text{לכל } i \text{ שעבורו } 0 \leq \langle C_i | y - Xw \rangle \leq \lambda \text{ ו-} w_i \neq 0$$

עתה האיטרציה של Sub-Gradient Descent היא $x^{(t+1)} = x^{(t)} - \eta_{t+1} g^{(t)}$ כאשר $g^{(t)} \in \partial f(x^{(t)})$ הוא תת-גרדיינט כלשהו (לא משנה איזה). שאר השלבים נשארים אותו הדבר, חוץ מהעכירה שבאה סיבת להחזיר את $x^{(T)}$ כי לא בהכרח ש- $g^{(T)}$ הוא כיוון ירידה, אך ראיי להחזיר את הממוצע או ה-best-preforming.

דוגמה למעשה Sub-Gradient Descent הוא Perceptron על על-מיוריים.

הערה כרגע צריך לבחור בחוכמה את step sizes.

הערה גם כאן יש תורה שלמה מאחרי התכניות, ובתנאי לפישיות וכוכ' אכן נקבל התכניות לפתרון האופטימלי.

תרגיל השתמשו ב- $\eta_t = \frac{1}{t}$ וכתבו Sub-GD לאלסו, באמצעות פ' שמחזירה תת-גרדיינט כלשהו (נגזרת בכל נקודה גזירה ואחרות תת-גרדיינט כלשהו, עתה חישבנו). ציירו את הגרפים וראו שאכן מתקנסים.

תרגיל עשו את אותו הדבר לריגרסיה לוגיסטיית עם הגרדיינט של המטרה, והוסיפו רגולרייזציית ℓ_1 באמצעות פ' שמחזירה תת-גרדיינט בכל נקודה (ראינו עתה מהו התת-גרדיינט של ℓ_1).

תרגול

התהlik שבו בוחרים פיצ'רים למודל נקרא Feature Selection. אנחנו מוחפשים את תת-הקבוצה הכי טובה של פיצ'רים, רק שאנחנו לא יודעים מה גודל תת-הקבוצה שאנחנו רוצים ככלمر צריך לעبور על 2^d אפשרויות שווה לא פרקטן. לכן צריך היוריסטיקה.

אלג' Forward Stepwise Selection

1. נאותחל m_0 מודל ללא פיצ'רים.

2. לכל $k \in \{0, \dots, d-1\}$:

(א) ניצור $d-k$ מודלים מהצורה $\{j\} \cup m_k$.

(ב) נבחר את המודל הכי טוב למעלה ונגדירו m_{k+1} .

3. נחזיר את המודל הכי טוב מבין m_0, \dots, m_d .

FSS הוא אלג' חמוץ קלסי. ב-2 בבחירה המודל הכי טוב היה בעצם מקרה פרטי של Feature Selection ובחירה המודל הכי טוב ב-3 היה מקרה פרטי של Model Selection.

כיצד נמדד את איכות המודל? RSS הוא בין 0 ל- ∞ וסובל מבעיות של סקיליניג (מה זה סטייה של 1000? זה טוב? זה רע?) וגם לא מנורמל למספר הדוגמאות. MSE הוא בין 0 ל- ∞ וכן מנורמל לפי מספר הדוגמאות אבל עדיין חסר משמעות. לשם כך הומצא ה- R^2 .

הגדולה מקדם ה- R^2 מוגדר ע"י

$$R^2 = 1 - \frac{\text{Var Unexplained}}{\text{Variance Total}} = 1 - \frac{SSE}{SST} = 1 - \frac{\|y - \hat{y}\|_2^2}{\|y - \bar{y}\|_2^2}$$

$$\text{כasher } \bar{y} = \frac{1}{m} \sum y_i$$

עתה נוכל לאמן כל מודל ולמדוד את הביצועים שלו לפי ה- R^2 שלו - ככל שהוא יותר גבוה כך המודל יותר טוב. הביעיה היא שגם R^2 לא מעולה - הוא מונוטוני עולה ב-d, שכן ב-3 נוטה למודלים עם הרבה פיצ'רים (ב-2 זו לא בעיה כי יכולים באותו הגודל).

הגדולה מקדם ה- R^2 מוגדר ע"י Adjusted R^2

$$adjR^2 = 1 - \frac{\frac{RSS}{m-d-1}}{\frac{mvar(y)}{m-1}} = 1 - \frac{m-1}{m-d-1} (1 - R^2)$$

טענה למקסם את $adjR^2$ שקול למיקסום . $\frac{RSS}{m-d-1}$

בדומה ל-**FS Selection**, אפשר לעשות את אותו האלג' רק כל פעם להוריד את הפיצ'ר שמשפיע ככמה שפחות (לרעה) על $adjR^2$. עד כה השתמשנו בפורנרי בעיות קמורות כקופסה שחורה, ועתה נלמד את התאוריה שמאחוריהם..Columns משתמשים בוריאציה כזו או אחרת של העיקנון הבא: נתחילה מנוקודה כלשהי ונozo ימינה או שמאליה (בכמה מימדים) בהתאם לאיזו בחירה יותר טובה - כלומר אם הגדריאנט חיובי או שלילי. כשהבעיה קמורה, לכת תמיד נגד כיוון הגרדיאנט (להקטין את הערך) ייתן לנו את הפתרון.

דוגמא ברגression לוגיסטי ניסינו לモזער את

$$f(w) = -\ell(w) = -\sum \left(y_i \log \frac{e^{-y_i}}{1 - e^{y_i}} + (1 - y_i) \log \frac{e^{-y_i}}{1 + e^{y_i}} \right)$$

וניתן להוכחה כי זו פ' קמורה (לא מעניין).

לפתרו כשייש נזרת זה קל, אבל דברים כמו לא גזירים ואז צריך טרייך אחר. אינטואיטיבית, נרצה לו שתמיד יהיה מתחת לגרף הפ', במקרה של נקודת גזירה זה כMOVEDן הגרדיאנט, אבל אם היא לא גזירה נרצה נרצה עדיין לו כזו.

הגדולה התת-דיפרנציאלי של f ב- x היא אוסף כל הקווים שעוברם מתחת לגרף הפ' (התת-גרדיינטים שלו).

שבוע VII | למידה عمוקה

הרצאה

SGD

.Sub-GD הוא הרחבה של GD, או יותר נכון של Stochastic Gradient Descent

הגדרה יהיו G וקטור מקרי ב- \mathbb{R}^d המקיים $E [G^{(t)}] \in \partial f(x^{(t)})$. האיטרציה של SGD היא

$$x^{(t+1)} = x^{(t)} - \eta_{t+1} g^{(t)}$$

כאשר $g^{(t)} \sim \mathcal{B}^n$

הערה מה שקרה כאן זה שאנו מגרילים כיוון, שבוחלת הוא נכון (כיון הירידה הכי גדול).

הערה אם f היא גזירה אז $E [G^{(t)}] = \nabla f(x^{(t)})$

כרגע נctrיך לקבע תנאי עצירה, ונבחר להחזיר את ממוצע הוקטוריים.

Sub-GD $\sum f_i(w) = \sum \partial f_i(w)$ כאמור (ולכן האיטרציה ב-Sub-GD) מתקיים $f(w)$ כאשר $f_i(w)$ קבועות. דוגמה נניח שאנו רוצים למצער $\sum_i f_i(w)$ הוא

$$x^{(t+1)} = x^{(t)} - \eta_{t+1} \sum_i g_i^{(t)}$$

כאשר $g_i^{(t)} \in \partial f_i(x^{(t)})$

ב-SGD נוכל לעשות איטרציה $x^{(t+1)} = x^{(t)} - \eta_{t+1} g_{k(t)}^{(t)}$

הערה יתרו אחד שאנו כבר רואים הוא יתרו חישובי משמעותי בכל איטרציה (רק תת-גדריאנט אחד בדוגמה הקודמת ולא m).

הגדרה היא הכללה של בחירת פ' אחת בכל פעם, שגוררת תת-קובוצה של תת-גדריאנטים בכל פעם וביצעת איטרציה

$$x^{(t+1)} = x^{(t)} - \frac{\eta_{t+1}}{|K(t)|} \sum_{j \in K(t)} g_j^{(t)}$$

כאשר $K(t) \subseteq [m]$ הוא אוסף נדgm אחד וב"ת, או מבודסראט, או בכל קבוע.

דוגמה תהי בעיתת למידה קמורה עם מחלוקת היפותזות \mathcal{H} ו-loss $\ell(w, (x, y))$. נסמן $L_S(w)$ הסיכון האמפירי על מדגם אימון. אנחנו מנסים לפתור את הבעיה הקמורה $\min_{w \in \mathcal{H}} L_S(w)$.

$$w^{(t+1)} = w^{(t)} - \eta_{t+1} g^{(t)}$$

כאשר $g^{(t)} \in \partial \ell(w^{(t)}, z_t)$ זוהי כפי שתיארנו לפני איטרציית SGD שמצוירת את הפ' w .

עובדת, קיבלנו אלג' חזק מאוד (רק loss אחד בכל איטרציה).

השינוי המركזי הוא ב-GD מסתכלים כל פעם על כל S וב-SGD רק על מספר קטן של דוגמאות. למעשה SGD עשויה איטרציה עם $g^{(t)}$ שהוא אומד בלתי מותה לגרדיאנט, כי מתקיים

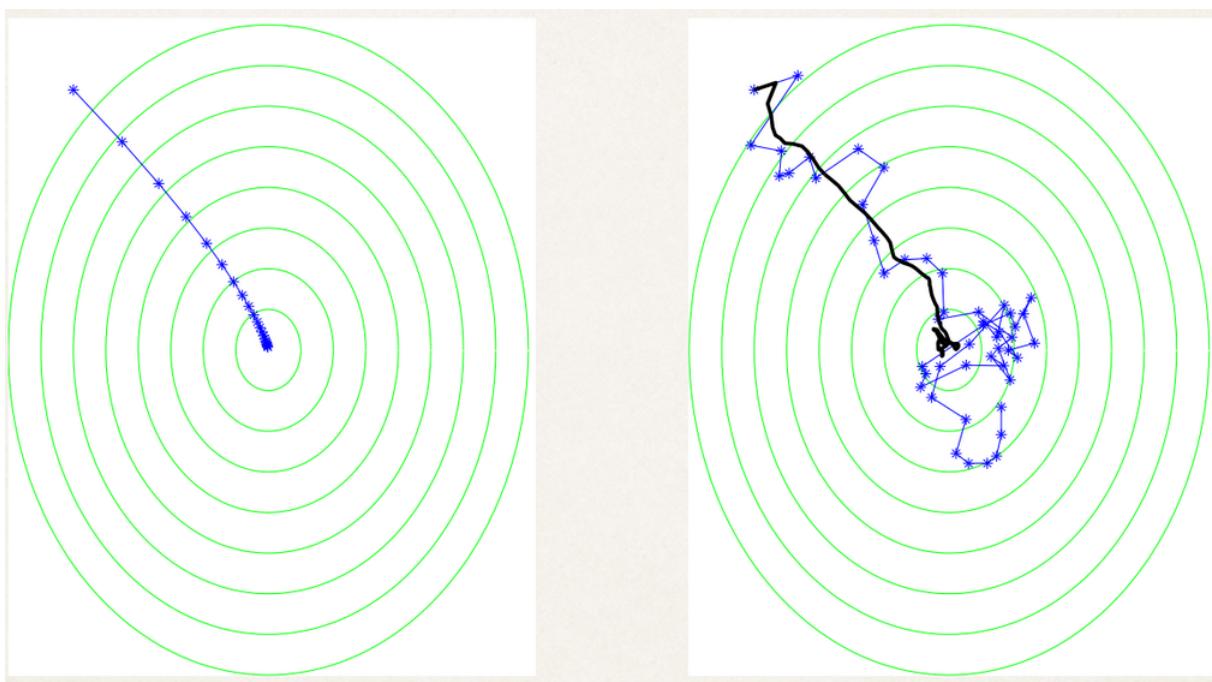
$$E \left[\nabla \ell \left(w^{(t)}, (x_i, y_i) \right) \right] = \nabla E \left[\ell \left(w^{(t)}, (x_i, y_i) \right) \right] = \nabla L_S \left(w^{(t)} \right)$$

הערה שימוש ב-mini-batch מאפשר להוריד את השונות לעומת הת-גרדיאנט של דוגמה אחת (מציעים על דוגמאות אקריאיות) וכך גם נקבל הוכנסות יותר מהירה של האומד לתוצאה. כמובן שהוא בא על חשבונו קושי חישובי יותר גדול.

הגדרה לפי שיטת cyclic rules, במקרה באקראי-mini-batch, עבור אופן סדרתי על הדוגמאות כך שנעבור על כל מוגם האימון. epoch הוא מספר הצעדים של SGD שצריך לעשות כדי לעבור על כל מוגם האימון פעם אחת, אם אנחנו עושים SGD ציקלי עם b דוגמאות כל פעם, האפקט הוא $\frac{m}{b}$ mini-batch.

SGD עם mini-batch ציקלי הוא כבר לא מיידת Batch, כי אנחנו בונים את המודל בהדרגות ודורשים רק את b האיברים הנוכחיים, בלי לדעת שהאחרים קיימים בכלל. זו בעצם... למידת אונליין! אפשר לאמן על S_1 , להשתמש במודל לניבוא, לקבל עוד נתונים וכאילו זה עוד איטרציות של mini-batch ציקלי לאמן את המודל דינאמית עם S_2 ולהשתמש במודל החדש.

הערה השונות של האלגוריתם גובאה משל GD כי אנחנו מושפעים בכל פעם מדוגמאות ספציפיות, נזון עוד בטרידוף בהמשך. **דוגמה** באירור הבא ניתן לראות משماה את GD ומימין את SGD, במבנה ש-GD עובד יותר טוב מבחינה "פדןית" אבל SGD עדיין מבצע לא רע בכלל.



PAC במסגרת SGD

נסתכל על SGD בהקשר של PAC. אנחנו מניחים שדגימות מוגעות מ- \mathcal{D} $\sim \text{i.i.d}$. המטרה שלנו היא לפתור

$$\min_{w \in \mathcal{H}} L_{\mathcal{D}}(w)$$

כאשר $L_{\mathcal{D}}(w) = E_{z \sim \mathcal{D}}[\ell(w, z)]$ בתקופה \mathcal{D} (או $L_{\mathcal{D}}(w) = E_{z \sim \mathcal{D}}[\ell(w, z)]$ ב-ERM). אין לנו דרך לモזער את $L_{\mathcal{D}}(w)$ ולכן אנחנו משתמשים ב- $\nabla L_{\mathcal{D}}(w)$ עם GD. אבל אין לנו אותו. מתקיים

$$\nabla L_{\mathcal{D}}(w) = \nabla E_{\mathcal{D}}[\ell(w, z)] = E_{\mathcal{D}}[\nabla \ell(w, z)]$$

כלומר שהגדריאנט של loss על דוגימה אחת הוא אומד בלתי-מדויק של הגדריאנט $(\nabla L_{\mathcal{D}}(w))$.

משפט תהי בעיתת למידה קמורה חסומה-ליפשיצית עם פרמטרים B, ρ, ϵ , אם נרים את SGD על ERM כדי $L_{\mathcal{D}}(w)$ אז לכל $0 < \epsilon$, אם נקבע $T \geq \frac{B^2 \rho^2}{\epsilon^2}$ נקבל

$$\text{אין לנו } (\mathcal{D}) \text{ עם מספר איטרציות } \eta = \sqrt{\frac{B^2}{\rho^2 T}}, \text{ נקבל}$$

$$E[L_{\mathcal{D}}(\bar{w})] \leq \min_{w \in \mathcal{H}} L_{\mathcal{D}}(w) + \epsilon$$

כאשר \bar{w} ממוצע הוקטורים באיטרציות או במלים, עם מספיק איטרציות, אנחנו מצליחים להגיע בתוחלת לשגיאת $L_{\mathcal{D}}$ קטנה ככל שנרצה.

דוגמא להלן אלג' Soft-SVM לפתרון

```

initialize:  $\theta^{(1)} = \mathbf{0}$ 
for  $t = 1, \dots, T$ 
    Let  $\mathbf{w}^{(t)} = \frac{1}{\lambda t} \theta^{(t)}$ 
    Choose  $i$  uniformly at random from  $[m]$ 
    If  $(y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle < 1)$ 
        Set  $\theta^{(t+1)} = \theta^{(t)} + y_i \mathbf{x}_i$ 
    Else
        Set  $\theta^{(t+1)} = \theta^{(t)}$ 
output:  $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}^{(t)}$ 

```

בכל איטרציה אנחנו בוחרים דוגימה, אם היא לא בצד הנכון של העל-מישור, נסיט את המישור (העל-מישור הסופי הוא ממוצע האיטרציות,

שבו בغالל ש- w הוא $\frac{1}{t} \theta$, כלומר w הוא ממוצע איטרציות כז השינוי זניח).

יש לנו גרסה שקולה ל-Soft-SVM מקורן.

תרגיל ממשו ריגריסה לוגיסטיבית עם SGD.

ממשו ריגריסה לוגיסטיבית עם רגוליזציית ℓ_1 עם SGD, כאשרם מקבלים כארגומנט את גודל ה-mini-batch. המשטו בכל הגרפים מהתרגילים הקודמים כדי להשוות את GD עם Sub-GD ו-Sub-GD עם SGD ו-mini-batch עם SGD. הניחו ש- m ו- d כל כך גדולים שאפשר להכניס רק b דוגמאות לזכרון במכונה אחת. תכנו אלג' מבוזר ל-SGD, עם מחשב מסטר שמחזיק את $w^{(t)}$.

רשתות נירונים

יש הרבה מאוד שימושים ל-Deep Learning, ראו האינטרנט לדוגמאות.

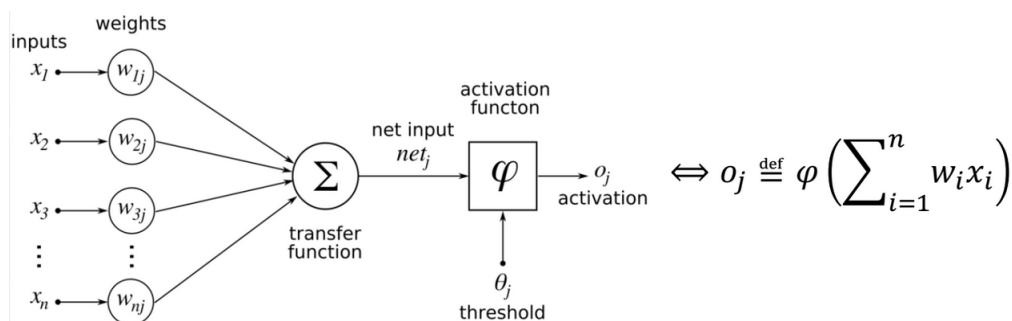
זהו פשוט שימוש ב-Deep Net, שהוא מודל ענק שמבצע הרבה ריגריסות לוגיסטיות קטנות (או אבני יסוד אחרות).

תרגיל הסטודנטית המשקיפה לתכוב ת"ז לומד Supervised Forwardfeeding Neural Networks בסוף הרצאה.

זהו פשוט רשת נירונית עם הרבה מושגים עם רשת נירונית, עכשו נלמד על רשתות נירונים.

זיכרון ריגריסה לוגיסטיבית משתמשת בעיקרון MLE למציאת הפוטזה מתוך המחלקה $\mathcal{H} = \{x \mapsto \phi(\langle w | x \rangle) : w \in \mathbb{R}^{d+1}\}$ כאשר ϕ היא פ' ה-logit, $\phi(x) = \frac{e^x}{1+e^x}$, שמקבלת ערכים ב- \mathbb{R} ומוחירה ערך ב- $[0, 1]$ שהוא הסבירות לקבל 0 או 1 בסיווג (ערך קטן יהיה מחלוקת 0 וערך גדול יהיה מחלוקת 1).

נציג את ריגריסה לוגיסטיבית כfcn יסוד עם קלטים ופלטים,

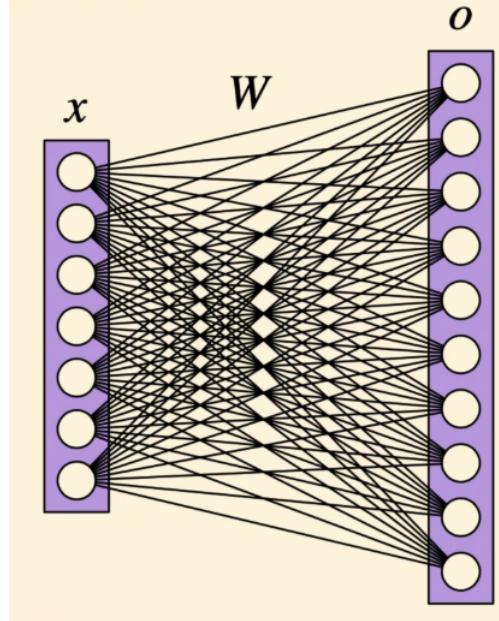


כאשר למעשה מה שקרה כאן זה פשוט הפעלה של הפ' $\phi = \varphi(f(x)) = \phi(\langle w | x \rangle)$.

הערה הקשר לנירונים הוא שנירונים מקבלים קלטים, עושים משחו ושולחים הלאה, בדומה למודל הזה.

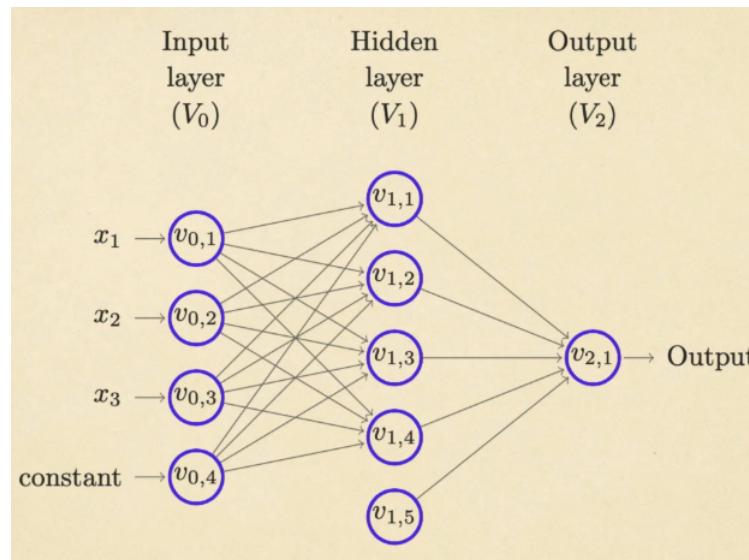
רשת נירונית לוקחת את הידיות האלה, שמה אותם ביחיד למודל מוצלח. נניח שיש לנו מודל ריגריסה לוגיסטיבית $x \mapsto \phi\left(\langle x | w_1^{(1)} \rangle\right)$ ומודל נוסף $x \mapsto \phi\left(\langle x | w_2^{(1)} \rangle\right)$.

קיבלונו מודל שמקבל דוגמאות ב- \mathbb{R}^d ומוחזר פלטים ב- \mathbb{R}^k . כך נראה רשת נירונית בלי שכבות נסתרות (מה שתיארנו עכשו).



למעשה אפשר לכתוב בכתב מטריציוני את הניבוא שלנו כ- $f(x) = \phi(W^T x)$. שום דבר לא מונע מאייתנו להוסיף עוד שכבות כאלה.

דוגמה נוסיף שכבה מוחבאת אחד. נקבל



למעשה קיבלנו עדיין מודל שmbחוץ נראה כמו ריגריסה לוגיסטיית קצר - מקבל וקטור קלטים ופולט מספר בין 0 ל-1. בשכבה המוחבאת, עושים מכ"פ ואוז מפעילים פונקציית אקטיבציה.

הגדרה מחלוקת ההיפותזות של feedforward neural network עם שכבה מוחבאת אחת שמכילה k נוירונים עם אקטיבציה σ לשימוש בינארי, היא

$$\mathcal{H} = \{\phi(\sigma(\langle W_1^T x | w_2 \rangle))\}$$

כasher $\phi : \mathbb{R} \rightarrow \mathbb{R}$, σ עובדות על וקטוריים איבר-איבר, $w_2 \in \mathbb{R}^k$, $W_1 \in \mathbb{R}^{d \times k}$.

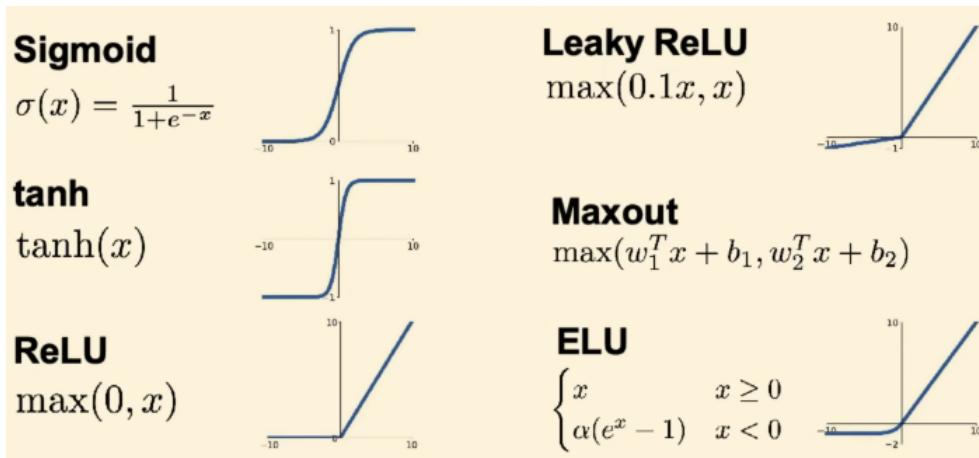
רשת שבה כל המשקولات הן לא אפס. fully connected.

הערה ההגדירה הנ"ל פשוט מגדירה מודל עם שתי פ' אקטיבציה שונות (אחד לכל שכבה).

הערה אם אנחנו בסיווג נשתמש בפ' שלוקחת את \mathbb{R} ומעבירה אותו ל- $[0, 1]$ (מוניוניות) ובריגריסיה נשתמש פ' מוניוניות לא לינארית כלשהי.

אם נשתמש בריגריסיה בפ' הזהות לדוגמה, נקבל ריגריסיה לינארית קלאסית.

דוגמה ראו כמה אפשרויות לפ' אקטיבציה, כולם לא לינאריות ומוניוניות.



בשביל multi-regression (ריגריסיה עם k תגבות), נשתמש בשכבה פלט עם k נוירונים.

בריגריסיה לינארית עם מספר תגבות פשוט נשתמש במודל $x \mapsto (\langle x | w_1 \rangle, \dots, \langle x | w_k \rangle)$ של מחלקה $j \in [k-1]$. ה-likelihood $x \mapsto (\langle x | w_1 \rangle, \dots, \langle x | w_k \rangle)$ של מחלקה j היא

$$\frac{e^{\langle x | w_i \rangle}}{1 + \sum_{\alpha=1}^{k-1} e^{\langle x | w_\alpha \rangle}}$$

(ישל המחלקה האחרונה זה אחד פחות- סכום האחרים). נאמנו מודל עם MLE כרגיל וזה יוצאה די דומה, הסטודנטנית המשקיעה תעשה זאת פורמלית.

נחוור לרשתות נוירונים - כדי לעשות מולטי-סיווג על k מחלקות, נשתמש בשכבה הפלט (שתכילה k נוירונים) באקטיבציה softmax הומוגדרת ע"י

$$\phi : (z_1, \dots, z_k) \mapsto \left(\frac{e^{z_1}}{\sum_\alpha e^{z_\alpha}}, \dots, \frac{e^{z_k}}{\sum_\alpha e^{z_\alpha}} \right)$$

כasher z_i הוא הפלט של הנירון ה- j בשכבה האחורונה. פ' זו נקראת softmax.

כלומר, יש לנו מודל, שכשוברים מהשכבה המוחבאת האחורונה לפלט, לא מבצעים מכ"פ אלא רק מפעלים softmax על כל הנוירונים

ביחד לקבלת k ערכים בין 0 ל-1 שהם הסבירות לקבל כל מחלוקת, ומציגים זאת למשתמש (softmax) אמונ על נרמול לערכים שמסתכמים ב-1).

תרגיל ה@studentית המשקיפה תוויה כי $\log(\text{logit})$ מקרבת את sign , וכן שעבור

$$-\log(\phi) : z \mapsto \left(\dots, \log \left(\sum_{\alpha} e^{z_{\alpha}} \right) - z_i, \dots \right)$$

$$\dots \log \left(\sum_{\alpha} e^{z_{\alpha}} \right) = z^* + \log \left(\sum_{\alpha} e^{z_{\alpha} - z^*} \right) \text{ ש } z^* = \max \{z_i\}$$

נסיק כי אם z^* גדול משמעותית מכל שאר z_i אז $z^* - z_i \approx 0$, ואז $\log \left(\sum_{\alpha} e^{z_{\alpha} - z^*} \right) \approx z^*$.
כלומר, אם רשת הנוירונים מבאת נסוך (y, x) , לאמור ϕ על שכבות הפלט נותן את התגובה היכי גדולה על הליבול u , הרי שהדגימה תתרום באופן זניח לפ' שאנחנו מנסים למצער ($-\log -\text{likelihood}$).

דוגמה כמשמעותם ספרות, בנו מודל מוצלח שימוש בשכבה קלט (784 נוירונים, אחד לכל פיקסל), שכבה מוחבאת עם 15 נוירונים ושבבת פלט עם 10 אפשרויות. יהיו לנו 2 פ' אקטיבציה נוספת על ה-softmax (מהקלט למוחבאת, המוחבאת לפלט, ופלט למולטי-סיווג סופי).

הגדרה רשת נוירונים feedforward כללית עם קלט $x \in \mathbb{R}^d$, שכבות, פ' אקטיבציה σ (זוהה לכל השכבות) ופלט ב- \mathbb{R}^k מוגדרת ע"י:

$$G = (V, E) \cdot \bigcup_{t=0}^T V_t \text{ גראף מכון חסר מעגלים, כאשר } V_0 \text{ חן השכבות.}$$

$$\cdot \text{ פ' משקל } \mathbb{R} \rightarrow E : w \text{ על הצלעות (מיוצגת ע"י מטריות } .W_0, \dots, W_{T-1} \text{).}$$

$$\cdot \text{ פ' אקטיבציה } \sigma \text{ ופ' פלט } \mathbb{R} \rightarrow \mathbb{R} : \sigma \text{ ופ' פלט }$$

נדיר ריקורסיבית $x \mapsto \phi(o_T), o_0 = \sigma(W_{t-1}^T o_{t-1})$.

הערה תחת ההגדרה זו, רשת נוירונים עם שכבה מוחבאת אחת היא רשת כללית עם 2 שכבות (לא 3 כפי שאמרנו בדוגמה בהתחלה).

רשתות עמוקות

במקום להציג כל נוירון בנפרד, נציג כל שכבה כבלוק וכך נוכל לתאר גרפית העברת שכבה אחת לאחרת, איחוד של שתיים, פלט וכו'. רשתות עמוקות יכולות להכיל גם 10 שכבות מוחבאות (GoogleNet, ResNet).

איך לאמן רשתות

מחלקת ההיפותזות עבור $G = (V, E)$ היא \mathcal{H}_G , שזה וקטורים בגודל $\mathbb{R}^{|E|}$.

עקרון הלמידה שלנו הוא MLE, כלומר בהינתן $\mathcal{L}(w; S) = \sum_{(x,y) \in S} \ell_w(x, y)$, כאשר w הוא משקלות הרשת ו- ℓ הוא לוגיליליהו.

בניגוד להיות הלוג-לייקליהוד של ריגריסה לוגיסטי פ' קמורה (קורה בעיירון), הלוג-לייקליהוד של רשותות נוירונים מאוד לא קמורה. עם זאת, שימוש ב-GD עדין מצליח למקסם לוג-לייקליהוד וכך הוא נהי אלג' מאד חזק ללמידה, אפילו שפ' המטרה לא קמורה. אף אחד לא יודע למה זה עובד, ואנשים מנסים להוכיח למה מקרים פרטיים צריכים לעבוד וכו'.

הסיבה שזה כן עובד (לא יודעים למה סיבת זו קורת אבל), היא שהמנימומיים מקומיים בפ' המטרה הם מקומיות טובים להיות בהם. back propagation המשמש ב-SGD כדי לאמן את הרשות ובכל פעם משתמש בדוגמה אחת, ככלומר נצעד עם $\nabla \ell_w(x, y)$. במקרה, נשתמש ב-mini-batch כדי לאמן את הרשות ע"י

$$w^{(t)} = w^{(t)} - \eta_t \sum_{(x,y) \in B_t} \nabla L(w^{(t)}; B_t)$$

תרגול

פסאודו-קוד של שיטות Descent

תהי f קמורה וdifrenzialabile.

- נבחר $x \in \text{dom}(f)$ נקודת התחלה.

• בכל איטרציה נבצע:

1. נבחר כיוון ירידה Δx .

2. נבחר גודל צעד η .

3. נעדכן $x = x + \eta \Delta x$.

- כל עוד לא מתקיים תנאי עצירה.

בתרגול עוסוק בשלב 1 ו-2 באיטרציה, ובתנאי העצירה.

בחירה כיוון ירידה

אם f היא $\mathbb{R} \rightarrow \mathbb{R}$ פשוט הולכים בכיוון הנגורת, כך בעצם גם במקרה הרבה מימדי רק שאז זה עם הגדריאנט. הקירוב טילטור של f ב- x (הנקודה האופטימלית לצורך העניין) ביחס ל- $x^{(t)}$ הוא

$$f(x) \approx f(x^{(t)}) + \frac{1}{1!} \nabla f(x^{(t)})^T (x - x^{(t)}) + \frac{1}{2!} (x - x^{(t)})^T H_f(x^{(t)}) (x - x^{(t)})^T + \dots$$

שיטות מסדר ראשון יבחרו כיוון רק עם הגורם הראשון (הגדריאנט). כש- f קמורה, הקירוב הוא תמיד מתחת לגרף הפ' ולכן אנחנו עושים הערכת-חווסף לערך של הפ' (אנחנו חושבים שהפ' יותר קטן מאשר ממשית באמות).

לכן נרצה לצעוד $x^{(t+1)} = x^{(t)} - \eta_{t+1}^T \nabla f(x^{(t)})$ כאשר $\eta_{t+1} = x^{(t+1)} - x^{(t)}$ (אם נשתכל על המכ"פ כהטלה של η_{t+1} על הגדריאנט, בעצם מה שקורח זה שאנו מטילים את הצעד שאנו רוצים לו על כיוון הירידה הכי תלולה). אנחנו לא יודעים מה זה $x^{(t+1)} - x^{(t)}$ במציאות ולכן צריך לבחור צעד אחרות.

$$\text{הערה אם נסמן } \Delta x = -\nabla f(x^{(t)}) \text{ אז אנחנו מעדכנים } x^{(t+1)} = x^{(t)} + \eta_{t+1} \Delta x \text{ כאשר } 0 > \eta_{t+1}.$$

הערה גם אם η קבוע, הגדריאנט מכיל לא רק כיוון אלא גם גודל - ככל שהזיה יותר תלול כך הגדריאנט יותר גדול ואז הצעד יותר גדול, ואם אין שינוימשמעותי נצא קצר מאוד. יש שיטות Descent שמנרמלים את הגדריאנט שלא נלמד.

בחירה גודל הצעד

גודל הצעד הוא חשוב לאלג' איקוטי, כי אם הוא גדול מדי נתבדר/נפספס את הנקודה האופטימלית ואם הוא קטן מדי, האלג' יגיע לנקודה האופטימלי לאט מדי.

בעיקרון נרצה גודל צעד שדווקע עם הזמן אבל איך הוא דועך וזה בחירה על בסיס מקרה ספציפי. אפשר לדעך לינארית, פולינומיאלית, אקספוננציאלית, מדרגות וגו' וכו'.

יש גם אלג' יודדים שבוחרים דינאמית את הצעד, ביניהם BLS.

נקבע Δx . נגיד $\langle \nabla f(x) | \Delta x \rangle = f(x) + \alpha \eta$ כאשר $\eta > 0, \alpha \in (0, 1)$. אם $\alpha = 1$ או $\alpha < 1$ הוא בכיוון הקירוב לינארית של מסדר ראשון, כאשר η קבוע כמו אנחנו נעים על המשיק בנקודה.

אם $\alpha = 0$ אנחנו מקבלים פ' קבועה שערכה הקירוב. נסיק (אפשר לראות במחשב) ש- α קובע את הזווית שלנו ביחס למשיק (ככל ש- α יותר גדול, כך אנחנו יותר קרובים למשיק).

עכשו נרצה למצוא η טובה כך שנכוון היטב לנקודה האופטימלית. אם η גדולה מדי קיבל קירוב מעיל הפ' ואם η קטנה מדי קיבל ערך מתחת לפ', ולכן ממשפט ערך הביניים קיימת η שעובדת יש שוויון (הקירוב ה"מושלם", ננסה לכוון אליו).

BLS-קוד של BLS

יהיו היפר-פרמטרים $\alpha \in (0, 1)$ ($\alpha = \frac{1}{2}$ נניח) ו- $\beta \in (0, 1)$ ($\beta = \alpha$ במקרה ש- α יתיר עדינה).

1. נתחל $\eta = 1$.

2. כל עוד $\eta = \beta \eta$, $f(x) + \alpha \eta \langle \nabla f(x) | \Delta x \rangle < f(x + \eta \Delta x)$

3. נחזיר η .

במקרה שהנגזרת לא גזירה בכל נקודה, משתמש במקום בגדריאנט בתת-גדריאנט, כל קו שעובר מתחת לגרף הפ', בلومר שקיימים

$$f(u) \geq f(x) + \langle v | u - x \rangle$$

לכל $u \in \text{dom}(f)$

שבוע VII | רשותות נוירוניים וסיכום

הרצאה

לא כולל תכנים שאינם רלוונטיים לחומר הקורס, אפ"ע, שהם חשובים ומעניינים.

נסתכל על מוטיבים מרכזיים ב-ML ועל חלק מהנושאים מזוויות אחרות.

ב-ML שהוא supervised אנחנו מקבלים דוגמאות ומנסים לבנות מודל לא מושלם אבל יהיה שימושי בחיקוי של המנגנון האמתי. כדי לפתור בעיות אמתיות אנחנו צריכים עולם מתמטי שייצג את העולם האמתי, וינה על שאלות כמו متى אפשר למדוד, כמה דוגמאות צריך, באיזו איקוס הنبוא שלנו עוד. PAC עונה על חלק מהשאלות האלה אבל לא על כלן.

דוגמה לפি PAC, דוגמאות מגיעות מהתפלגות לא ידועה באופן ב'ת'.

לפי PAC Agnostic אנחנו יכולים למדוד גם עם רוש והאקריאות האינהרנטית של הטבע.

בנוסף, ראיינו שאפשר למדוד אם \mathcal{H} סופית, או אם \mathcal{H} היא בעלת $\infty < \text{VCdim}(\mathcal{H})$ כאשר מושג של "מורכבות" של מחלוקת היפותזות.

ראיינו בעיות סיוג ביןארי (שתי מחלקות), בעיות סיוג מרובה (כמה חלוקות דיסקרטיות), בעיות עם הרבה לייבלים (התשובה הנכונה היא אוסף של לייבלים מתוך קבוצה).

ראיינו גם ריגריסה - תגובה רציפה. יש סוג שלישי של ניבוא, שהוא גנרטציה - לדוגמה בהינתן קבוע ליצור קבוע אחר שהוא התרגום של קבוע הקלט, ושם מרחב הפלט דומה למרחב הקלט.

כדי למדוד, נרצה שהמודל ימזרע את שנייה הכללה ביחס לפ' עלות כלשהי. בגלל שאנחנו לא יודעים מה שגיאת הכללה באמת, אנחנו משערכים אותה עם מודג הטסט. העלות צריכה לתואם את הדרישות שלנו, אם אנחנו מסווים סיכון של תרומות, שגיאות מסווג 1 ייכבו לנו הרבה יותר משגיאות מסווג 2.

ראיינו הרבה מודלים שונים לריגריסה, ביניהם ריגריסה לינארית, לוגיסטיבית, עצי ריגריסה, רשותות נוירוניים. לסיוג ראיינו SVM, Half-space,

לכל לומד ראיינו איך לבנות לו ת"ז - איך נראה גבול ההחלטה שלו, איך שומרים אותו בזיכרון, האם הוא ניתן לפרשו, כיצד נממש אותו וכו'.

דוגמה עצי סיוג ממומשים באמצעות אלג' CART, נשמר כע' עם ערכי סף בקודוקדים, מנביאים באמצעות traverse, אפשר לפרש בקלות בклות את המודל אבל אין לנו הסת' למחלקות - רק החלטה סופית.

אם אין לנו לייבלים (די שכח כי זה יקר), אפשר לעשות קלאסטרינג כדי לוזחות תופעות (k-means או ספקטרלי), להוריד מימד באמצעות PCA (או t-SNE שלא למדנו).

מוטיבים מרכזיים

• טירידוף ה-Bias-Variance : כל הזמן עלה, והוא מבטא את הרצון שלנו להקליל היטב כשאנו חונכו לא רוצחים לעשות אובר-פייט. הבעיה של ML (אחרת הכל היה קל) זה שמודלים לא יכולים להבדיל בין תבניות של דאטא ספציפי (אובר-פייט) לבין תבניות של כל הדאטא (הכללה טובה).

כל שמודל הוא יותר אקספרסיבי, הוא יכול לתאר תבניות באופן יותר מדויק ולפנן הוא יביא שגיאה יותר נמוכה אבל גם יכול להוביל לאובר פיט בגלל השינויים הגדולה.

כדי להילחם בבעיה זו, משתמשים רגולרייזציה. באמצעות רגולרייזציה, נבחר מודל מורכב רק אם אפשר להוכיח את המורכבות באמצעות שגיאת הכללה מספיקת טובה.

באופן יותר כללי, הפרדה טובה בין נתונים שראיינו (דאטא מזוהם), לבין שלא ראיינו (טסט ובחן כזו או אחרית ולידציה).

• פרמטרים נלמדים מול היפר-פרמטרים : כל מודל מתאים באמצעות קביעת פרמטרים נלמדים כלשהם (משכולות בריגרישה לינארית) אבל באותו הזמן רבים מהם דורשים פרמטרים שיהיו נתונים להם לפני האימון, למשל היפר-פרמטר הרגולרייזציה, צעד ב-DG, ארכיטקטורה של רשת נוירונים ועוד ועוד.

לכוארה יכולים לנו גם מודל כדי לקבל אוטומטיות היפר-פרמטרים אבל זה מגביל לנו את יכולת לקבוע כמה מרכיב אנחנו רוצחים שהמודל יהיה וכו'.

• מודלים על סטרואידים : אחרי שראיינו מודלים בסיסיים, הוספנו עוד שיטות מטה כדי לשפר אותם בהינתן מודל נתון. בניהם אנסמבלים, קרנליים, באגינג וכו'.

• פתרון אנלטי מול נומי : זו הפעם הראשונה שראיינו שקsha לממש תאוריה באופן נומי - אפילו בריגרישה לינארית אנחנו מתחשים לחשב ערכים סינגולריים באופן יציב נומי, שלא לדבר על SGD וכו'.

• תאוריה מול אמפיריות : אפע"פ שיש הרבה תאוריה מתמטית מאחוריה למידת מכונה, יש הרבה מאוד טריקים חדשים שלא קיימים בתאוריה ונוטנים לנו ביצועים טובים ואשף ML בקיא לא רק בתאוריה אלא גם בשיקולים מציאותיים.

• שיקולים מציאותיים : כל רעיון צריך להריץ על מחשב ומחשב פיזי מוגבל בזמן החישוב, זיכרון, דיקוק וכו'. זה משנה על ערכים סינגולריים ב-SVD, בגודל ה-batch-SGD, ו-vanishing gradients (לא למדנו, אבל זה המקירה שבו הגדריאנטים כל כך קטנים ב-GD שהמחשב לא מצליח ליצג אותם וחושב שהם 0 ופושט לא זו).

השעה השנייה והשלישית של ההרצאה זו היו אתיקה (מאוד חשוב אבל לא מהותי לתוכן הקורס) ומבוא ל-Pretraining ומבט אל העתיד מבחינה AI כקונספט, ML בתעשייה ובאקדמיה וקורסי המשך, שכאמור לא נכללים בהיקף הסיכום הזה.

תרגול

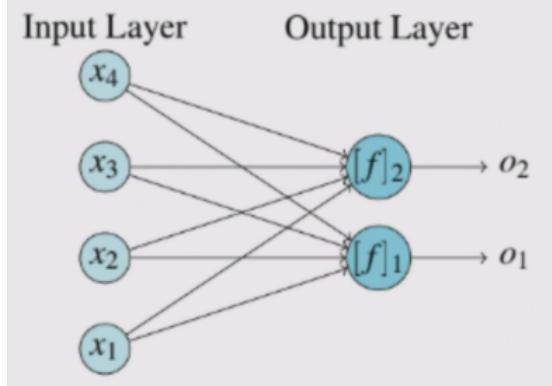
נוירון מורכב ממשקולות ותקטיבציה. סה"כ קיבל שנוירון מקבל קלט $\mathbb{R}^d \in x$ ופולט ($w | x$) $\varphi = o$. חשוב ש- φ תהיה לא לינארית כי אחרת פשוט מקבל מודל לינארית מאוד מוסף אבל עדין לינארית.

הגדולה הארכיטקטורה של רשת נוירונים היא תיאור מספר הנוירונים בכל שכבה ברשות.

.element-wise ו- $W \in \mathbb{R}^{k \times d}$ כאשר $f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_k(x) \end{pmatrix} = \begin{pmatrix} \varphi(\langle w_1 | x \rangle) \\ \vdots \\ \varphi(\langle w_k | x \rangle) \end{pmatrix} = \varphi(Wx)$ שכבה שלמה מчисבת

הערה במקרה הכלל, כל הנוירונים באותו שכבה ישתמשו באותה האקטיבציה אבל בין שכבות יכולות להיות פ' שונות.

דוגמא באירוע הבא, נקבל $(\dots_{w_1}^{\dots} \dots_{w_2}^{\dots}) = W_0$ עבור השכבה הראשונה (כאשר $w_i \in \mathbb{R}^4$) ואז פ' האקטיבציה מקבלת קלט את $\langle w_1 | x \rangle$ (עבור הנוירון התיכון בשכבה הפלט) או $\langle w_2 | x \rangle$ (עבור הנוירון העליון).



הגדירה נסמן את הפ' שמחשבת השכבה ה- j (זו שמחשבת $o_j = f_j(W_j x)$ הפלט שלה ו- a_j את הערך רגע לפני האקטיבציה, כלומר $W_j x$ ולכן מתקיים

$$f_t(o_{t-1}) = \varphi_t(W_{t-1}o_{t-1}) = \varphi_t(a_t) = o_t$$

הערה תחת הסימונים הקודמים, רשות פשוט מחשבת את הפ' f כדי לנבأ על דוגמה.

מה שנוטר לרשות ללמידה הם המשקولات (האקטיבציות אנחנו בחרנו וכן את הארכיטקטורה). נלמד באמצעות SGD כאשר אנחנו מנסים למזער את ה-Loss, כלומר אנחנו מעדכנים לפי

$$w^{(t)} = w^{(t-1)} - \eta_t \nabla L(w^{(t-1)}; B)$$

כאשר B הוא mini-batch או דוגמה אחת. עם זאת, L היא מאוד מורכבת כי היא מכילה בה את f , שהיא הרכבה של הרבה פ' עם הרבה משקولات.

טענה יהיו $J_x(f \circ g) = J_{g(x)}(f) J_g(x)$ אזי היעקוביאן של $f \circ g$ הוא $f : \mathbb{R}^d \rightarrow \mathbb{R}^m, g : \mathbb{R}^k \rightarrow \mathbb{R}^d$ וברמת האלמנט מתקיים

$$[J_x(f \circ g)]_{ij} = \sum_l \frac{\partial f_i(g(x))}{\partial g_l(x)} \cdot \frac{\partial g_l(x)}{\partial x_j}$$

דוגמה נבנה רשת שתקבל מקלט $\mathbb{R} \in x$ ויש לה שתי שכבות וначשב עבורה את ה-loss,

$$x \rightarrow \frac{w_0 x}{a_1} \rightarrow \frac{\varphi(w_0 x)}{o_1} \rightarrow \frac{w_1 o_1}{a_2} \rightarrow \frac{\varphi(a_2)}{o_2} \rightarrow C(o_2) = (o_2 - y)^2$$

וכשיו נчисב את הגדריאנט של ה-loss,

$$\begin{aligned}\frac{\partial c}{\partial w_1} &= \frac{\partial c}{\partial o_2} \cdot \frac{\partial o_2}{\partial a_2} \cdot \frac{\partial a_2}{\partial w_1} = 2(o_2 - y) \cdot \varphi'(a_2) \cdot o_1 \\ \frac{\partial c}{\partial w_0} &= \frac{\partial c}{\partial o_2} \cdot \frac{\partial o_2}{\partial a_2} \cdot \frac{\partial a_2}{\partial o_1} \cdot \frac{\partial o_1}{\partial a_1} \cdot \frac{\partial a_1}{\partial w_0} = 2(o_2 - y) \cdot \varphi'(a_2) \cdot w_1 \cdot \varphi'(a_1) \cdot x\end{aligned}$$

כלומר הצלחנו לחשב נזורת מרכיבת באמצעות הרכבה חישובים פשוטים.

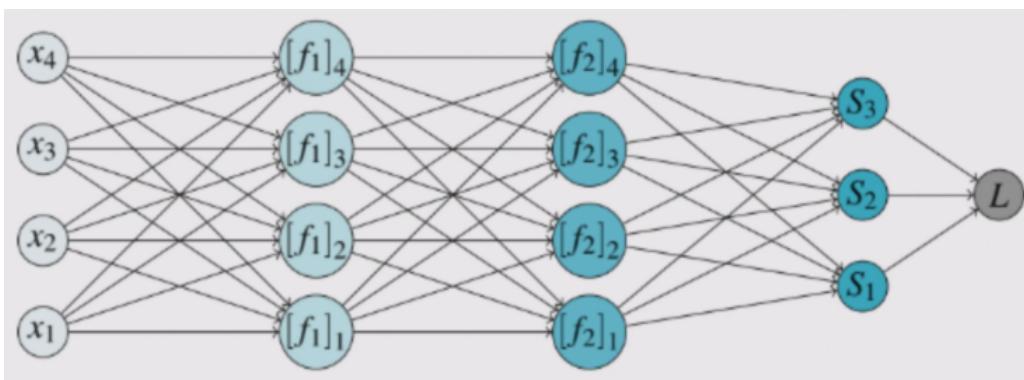
בנוסף, שני האיברים הראשוניים של שתי הנזורות החלקיות הם שווים ולכן אם נשמר את החישובים הקודמים, נחשוך חישובים חוזרים.

לכן, כדי לחשב ביעילות את הגדריאנט של ה-Loss נחשב מההתחלה לסוף כנסמoring את החישובים הקודמים.

הערה לרוב בוחרים אקטיבציות שקל לנזור (לחשב את הגדריאנט בנקודה כלשהי) כדי שחישוב גדריאנט ה-Loss יהיה פשוט (עד כדי תחת-גדריאנטים אם היא לא גזירה בנקודה כלשהי).

הערה נשים לב כי בחישובים הנ"ל יש שלוש סוגי נזורות, o לפני a , a לפני o ו- a לפני w (עד כדי אינדקסים).

דוגמה נבנה רשת כבאיר,



נסמן את החישוב כתוצאה מהשכבה הראשתית L_1 , הרשת השנייה L_2 והשלישית L_3 . מתקיים

$$\frac{\partial (L_T \circ L_{T-1} \circ \dots \circ L_1)}{\partial w_{t-1}} = \prod_{i=T}^t J_{o_{t-1}}(L_t)$$

מכלול השרשרת, כלומר הגזירה לפי משקלות הייעקוביאנים של כל האיברים החל מהשכבה שלה ועד הסוף. נפתח את

$$\begin{aligned}
\frac{\partial (L_T \circ L_{T-1} \circ \dots \circ L_1)}{\partial W_{t-1}} &= \prod_{i=T}^t J_{o_{i-1}}(L_i) \\
&= \left(\prod_{i=T}^{t+1} J_{o_{i-1}}(L_i) \right) J_{o_{t-1}}(L_t) \\
(*) &= \left(\prod_{i=T}^{t+1} J_{a_i}(o_i) J_{o_{i-1}}(a_i) \right) J_{o_{t-1}}(L_t) \\
(*) &= \left(\prod_{i=T}^{t+1} J_{a_i}(o_i) J_{o_{i-1}}(a_i) \right) \frac{J_{a_t}(o_t)}{\varphi'(a_t)} \frac{J_{W_{t-1}}(a_t)}{o_{t-1}}
\end{aligned}$$

(*) שוב הפענו את כל השרשראת, הפעם מהחישוב (

$$J_{o_{i-1}}(L_i) = J_{o_{i-1}}(o_i \circ a_i) = J_{a_i}(o_i) J_{o_{i-1}}(a_i)$$

עד כל השרשראת, רק שהפעם הגזירה של a_t לא מעניינת אותנו ביחס ל- o_{t-1} כי אנחנו מנסים להבין איך W_{t-1} משפייע על $Loss$.

עתה האקטיבציה השלישית היא Softmax שנوتנת וקטור הסט' (נסכם לאחד) ו- L -ייהה cross-entropy, שזו פ' שמודדת את ה"מරחך" של וקטורי הסט' (כמו שייתר קרוב כמה שיותר טוב).
לכן בעזרת הפיתוח הנ"ל,

$$\begin{aligned}
\frac{\partial}{W_1} (L \circ S \circ L_2) &= J_S(L) J_{o_2}(S) J_{a_2}(o_2) J_{W_1}(a_2) \\
\frac{\partial}{W_0} (L \circ S \circ L_2 \circ L_1) &= J_S(L) J_{o_2}(S) J_{a_2}(o_2) J_{o_1}(a_2) J_{a_1}(o_1) J_{W_0}(a_1)
\end{aligned}$$

מה חשוב לוזהות בדוגמה זו הוא העובדה שבהינתן שיש לנו את הנגורות החלקיות לפי W_{t-2} , גזירה לפי W_{t-1} מתקבלת ע"י מכפלה בעוד שני איברים בלבד (כי המכפלה מ-1 עד T כבר בוצעה) - האיבר מחוץ ל-[] והאיבר ה-t-[]]. אלג' שזוכר מה חישב עד כה באופן זהה יכול לחשב ביעילות את הגדריאנט, וזה מה שעושה Back-Propagation.

Back-Progpagation

האלג' מרכיב שני חלקים - חישוב קדימה כדי לחשב את o_t , a_t . במהלך החישוב אחרת, אנחנו מחשבים את הגדריאנט באמצעות שמירה של ה-[] בפיתוח שלנו) וכן של הנגורות החלקיות לפי כל משקלות, שיחד יהוו וקטור שהוא הגדריאנט (ראו פסאודו קוד).

Algorithm 1 Back-propagation

```
procedure BACK-PROPAGATION( $\langle G, \{\mathbf{W}_t\}, \{\sigma_t\}, \phi \rangle, (\mathbf{x}, y)$ )
    Denote  $N := L_T \circ L_{T-1} \circ \dots \circ L_1$ 
    FORWARD PASS
        Denote  $\mathbf{o}_0 \leftarrow \mathbf{x}$ 
        for  $t = 1, 2, \dots, T$  do
            Compute pre-activations  $\mathbf{a}_t \leftarrow \langle \mathbf{W}_{t-1}, \mathbf{o}_{t-1} \rangle$ 
            Compute activations  $\mathbf{o}_t \leftarrow \sigma_t(\mathbf{a}_t)$ 
        end for

    BACKWARD PASS
        Set  $\Delta_T = \phi'(\mathbf{o}_T)$ 
        for  $t = T-1, T-2, \dots, 1$  do
            Set derivation chain  $\Delta_t \leftarrow \Delta_{t+1} \cdot J_{\mathbf{a}_t}(\mathbf{o}_t) \cdot \mathbf{W}_{t-1}$ 
            Set partial derivatives  $\nabla_{\mathbf{W}_{t-1}} N \leftarrow \Delta_{t+1} \cdot J_{\mathbf{a}_t}(\mathbf{o}_t) \cdot \mathbf{o}_{t-1}$ 
        end for

    return  $\nabla N$ 
end procedure
```

כאן הסימונים הם $\{W_t\}$ המשקלות לכל שכבה, $\{\sigma_t\}$ האקטיבציה בכל שכבה, ϕ בסוף (לדוגמא Softmax).

. סוף.