
מבוא למערכות לומדות - ד"ר מתן גביש - סמסטר ב' 2021

~

הרצאות ותרגולים

מסכם: יחיאל מרכזבך

תוכן העניינים

7	I הקדמות מתמטיות
7	1 אלגברה ליניארית
7	1.1 העתקות ליניאריות
9	1.2 נורמות, מרחבי מכפלה פנימית והטלות
12	1.3 הצגה מטריציונית ג
14	2 אינפי
15	2.1 נגזרות, גרדיאנטים ויעקוביאנים
20	2.2 קירוב מסדר ראשון
22	3 הסתברות
22	3.1 עקרונות בסיסיים
27	3.2 קצט סטטיסטיקה - הערכת שונות ותוחלת
29	3.3 כמה משתנים
35	3.4 אי שוויונים
40	II רגרסיה ליניארית
41	1 מודלים של רגרסיה
42	1.1 רגרסיה ליניארית
43	1.2 עיצוב אלגוריתם למידה (Designing A Learning Algorithm)
49	1.3 שיקולים נומריים בעת מימוש אלגוריתם
49	1.4 הוספה רוש
52	1.5 משתנים קטגוריים
52	2 התאמת פולינומית (Polynomial fitting)
53	2.1 הטיה ושונות של אומדנים
56	III בעיות סיווג (Classification)
56	1 הקדמה לסיווג
56	1.1 פונקציית הפסד (Loss Functions)
56	1.2 שגיאות מסווג ראשון וסוג שני
57	1.3 מדדיות של ביצועים
58	1.4 גבולות החלטה (Decision Boundaries)
59	2 מסובג-חצאי-מרחיב (Half-Space Classifier)
59	2.1 הפסד אמפירי מינימלי (Learning Linearly Separable Data Via ERM)
60	2.2 פיתרון ERM בחצאי מרחבים

3	המewood SVM-Support Vector Machine	61
3.1	עקרון הלמידה של מקסום השול (Maximum Margin)	61
3.2	המקרה הריאלי - Hard-SVM	62
3.3	בלתי ניתן להפרדה ליניארית - Soft-SV	63
4	רגרסיה לוגיסטיבית	65
4.1	מודל הסתברותי עם רוש	65
4.2	מיימוש חישובי	67
4.3	פרשנות (Interpretability)	67
4.4	מחלקת ההיפතזה וכייצד חווים	67
5	השכנים הקרובים ביותר Nearest Neighbors	68
5.1	חיזוי באמצעות NN	68
5.2	בחירה k	69
5.3	חישובית	69
6	עצי החלטה	69
6.1	מחלקת ההיפתזה	70
6.2	עקרון הלמידה	70
6.3	שתייה עציים	70
IV	תיאוריות PAC של למידה סטטיסטית	70
1	הקדמה תיאורטית	71
1.1	למידה כמשחק - ניסיון ראשון	73
1.2	לומדים "בערך נכונים" ו"קרוב לוודאי נכונים" (Probably Correct & Approximately Correct Learners)	74
1.3	למידה כמשחק - ניסיון שני	77
2	אין ארכחות חינם ומחלקות היפותזה	77
2.1	משחק הלמידה (גרסה שלישית)	80
3	למידת PAC	82
3.1	למידת PAC של מחלקות היפתזה סופיות	83
3.2	מנגד ה-VC (VC-Dimension)	86
3.3	המשפט היסודי של למידה סטטיסטית (The Fundamental Theorem of Statistical Learning)	88
4	למידת Agnostic PAC	89
4.1	הקדמה לפונקציית התפלגות משותפת מעל $\mathcal{U} \times \mathcal{X}$	89
4.2	עדין הנחת הריאליות (Relaxing Realizability Assumption)	90
4.3	פונקציית הפסד כללית (General Loss Function)	90
4.4	למידת Agnostic-PAC	91
4.5	המשפט היסודי של הלמידה הסטטיסטית	93

99	V שיטות אנסמבל - Ensemble Methods
99	1 יחס הכוחות בין הטעיה והשונות (Bias-Variance Trade-off)
100	1.1 פירוק שגיאת ההכללה (Generalization Error Decomposition)
100	2 ועדות ואנסמבל (Ensemble/Committee Methods)
102	2.1 מנבאים בלתי מותאימים (Uncorrelated Predictors)
103	2.2 מנבאים מותאימים (Correlated Predictor)
104	2.3 שיטות ועדה במערכות למודדות (Committee Methods In Machine Learning)
105	3 צבירה אתחול (Bagging)
105	3.1 אתחול-אווזי-נעליים (The Bootstrap)
106	3.2 צבירה אתחול - Bagging
107	3.3 הפחיתת השונות על ידי bagging
107	3.4 יערות רנדומליים - Random Forests - ביטול הקורלציה בעצי החלטה ו-
108	4 הגברה - Boosting
109	4.1 אלגוריתם adaBoost
111	4.2 למידת PAC, למידה חלה (Weak Learnability) boosting-ו
112	4.3 הטיה ושונות ב-boosting
112	VI רגוליזציה ובחירה מודל
113	1 רגוליזציה
114	2 רגוליזציה עצי החלטה
115	3 רגוליזציה רגרסיבית
115	3.1 בחירת תתי קבוצות
116	3.2 רגוליזציה Ridge
118	3.3 רגוליזציה לאסו (Lasso) - נורמה ℓ_1
121	3.4 המקרה האורתוגונלי
124	4 בחירת מודל והערכתה (Model Selection and -Evaluation)
125	4.1 סכימת אימון-תיקוף- מבחן (Train-Validation-Test Scheme)
127	4.2 שיטת ה-Cross Validation
129	VII למידה בלתי מוחנית (Unsupervised Learning)
129	1 הקטנת ממדים
130	1.1 ניתוח גורמים ראשיים (Principal Component Analysis)
136	2 איגוד (Clustering)
136	2.1 k אמצעים (k-Means)
139	2.2 איחוד מלוכסן (Spectral Clustering)

140	VIII דרכי קרנלייזציה (kernel methods)
141	1 בעית למידה חלופית
142	2 אפיון פונקציות קרnel
143	2.1 פונקציונליות הkernel הפולינומיאלית והגאוסינית
144	2.2 תוכנות סגירות (Closure) עבור פונקציות ker nel PSD
145	2.3 ייצור ker nelים מker nelים קיימים
145	3 אלגוריתמים מker nelים (Kernelized Algorithm)
145	3.1 ker nel לרגרסיה ridge
145	3.2 ker nel לרגרסיה לוגיסטיבית לאחר רגוליזציה
146	3.3 ker nel ל-PCA
146	IX אופטימיזציה קמורה ולמידה عمוקה
147	1 קבוצות ופונקציות קמורות
147	1.1 הקדמה
147	1.2 קבוצות קמורות
149	1.3 פונקציות קמורות
152	2 תת-גרדיינט
152	2.1 תת-גרדיינט (sub - gradient)
153	2.2 תוכנות של תת-גרדיינט
154	2.3 חישבון של תת-גרדיינטים
154	3 תנאים מסדר גבורה לקמירות
156	4 בעיות אופטימיזציה
158	4.1 הקשר בין למידה ובין אופטימיזציה קמורה
158	5 מורד הגרדיינט (Gradient descent)
159	5.1 גרדיינט והקשר לשיפוע
159	5.2 האלגוריתם
160	5.3 איך נבחר את η ?
162	5.4 התוכניות של GD
163	6 מורד התת-גרדיינט (Sub-Gradient descent)
164	6.1 גודל הצעדים
165	7 שיטת Stochastic Gradient Descent
166	7.1 שימוש ב-SGD לפתור בעיות למידה קמורות
168	8 למידה عمוקה (deep learning)
168	8.1 רשתות ניורוניים
170	8.2 פונקציית אקטיבציה

170	רגרסיה לוגיסטיבית ל- <i>multiclass</i>	8.3
171	עширות מחלקת הhippozות	8.4
171	רשתות נוירונים عمוקות (deep neural net)	8.5
172	שימוש ב-GD ברשתות נוירונים	8.6

חלק I

הקדמות מתמטיות

1 אלגברה ליניארית

1.1 העתקות ליניאריות

הגדרה

יהיו $V \in \mathbb{R}^m$ -ו $W \in \mathbb{R}^d$ שני מרחבים וקטוריים. פונקציה $T : V \rightarrow W$ היא **העתקה ליניארית** מ- V ל- W אם לכל $v, u \in V$ ו- $c \in \mathbb{R}$ מתקיימות התכונות הבאות:

□ אדיטיביות: $T(u + v) = T(u) + T(v)$

□ מכפלה סקלרית: $T(cu) = cT(u)$

לכל W , בעלי מימד סופי, העתקה ליניארית יכולה להיות מיוצגת באמצעות מטריצה. לכן, مكان ולהלאה נטמקד בעיקר בiamiדים סופיים ובעיקר נטמקד בהציגת המטריציונית של העתקות ליניאריות.

הגדרה

העתקה אפינית היא העתקה מהצורה $w \in W$, כאשר $V \in \mathbb{R}^n$, $T(u) = Au + w$.

נבחן כי ההגדרה של העתקה אפינית היא לא העתקה ליניארית. כמו כן, נבחן כי העתקה ליניארית משמרת את איבר ה-0, כי $A \cdot 0_V = 0_W$, אבל במקרה של העתקה אפינית, מתקיים כי עבור $W \neq 0$ כאשר $.T(0_V) = A \cdot 0_V + w = w \neq 0_W$

cutet נגידיר מספר מרחבים וקטוריים הקשורים לכל העתקה ליניארית.

הגדרה

תהי A מטריצה המייצגת העתקה ליניארית $T : V \rightarrow W$. cutet נגידיר:

□ גרעין: נגידיר את הגרעין להיות $.N(A) := \{x \in V \mid Ax = 0\}$. מסומן גם בתור

□ התמונה - מוחב העמודות של A יוגדר בתור $.Col(A) := \{w \in W \mid w = Ax, x \in V\}$. מסומן גם בתור

□ מרחב השורות של A יסומן בתור $.Im(A) := \{x \in V \mid x = A^T w, w \in W\}$. בצורה דומה, הוא גם מוגדר בתור מוחב העמודות של A^T .

□ הגרעין של A^T יוגדר בתור $\ker(A^T) := \{x \in W \mid A^T x = 0\}$

נבחן כי לפי ההגדרה, $.Im(A) \subseteq W$ ו- $\ker(A) \subseteq V$, $.Row(A) \subseteq W$ ו- $\ker(A)$, $.Row(A) \subseteq V$. על ידי שימוש בהגדרות לעיל נוכל להגעה לכמה תובנות.

הגדרה

תהי $\text{rank}(A)$ (הדרגה) של $A \in \mathbb{R}^{m \times d}$. הוא המספר המקסימלי של שורות בת"ל של A , והוא מסומן בטור $\text{rank}(A)$.

מכאן ניתן להגשים למסקנה כי $\text{rank}(A)$ של A שווה גם למינימום השורות וגם למרחב העמודות. כתוצאה לכך, נאמר כי A מדרגה מלאה אם ורק אם $\text{rank}(A) = \min(m, d)$. אחרת, נאמר כי A מדרגה חסרה.

הגדרה

תהי $A \in \mathbb{R}^{d \times d}$ מטריצה ריבועית. A תיקרא הפיכה (או לא סינגולרית) אם ישנה מטריצה $B \in \mathbb{R}^{d \times d}$ כך $AB = I_d = BA$ ונסמן את המטריצה הההפוכה בתור A^{-1} .

טענה

תהי A מטריצה ריבועית. הטענות הבאות שקולות:

\square A היא הפיכה.

\square A היא מדרגה מלאה.

$\square \det(A) \neq 0$

$\square \text{Im}(A) = \mathbb{R}^m$

$\square \ker(A) = \{0\}$

דוגמה

נתבונן במקרה הבא. נניח ונთונה לנו קבוצה של n משוואות ליניאריות, כשל אחת מהצורה $y_i = \sum_{j=1}^d w_j \cdot x_{ij}$, כאשר x_{ij} ו- y_i נתונים לנו, אבל w_j לא נתון לנו. נרצה למצוא פתרון למערכת המשוואות. למעשה,ollo כל הוקטורים $w \in \mathbb{R}^d$ שמקיימים:

$$\forall i \in [d] \quad y_i = \sum_{j=1}^d w_j \cdot x_{ij} = w^\top x_i$$

כעת, נאוסף את המשוואות ונסדר אותן בתוך מטריצה, ולמעשה נרצה למצוא $w \in \mathbb{R}^d$ ש- $y = Xw$. קלומר $\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} - & x_1 & - \\ - & x_m & - \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$. עניין אותנו לדעת האם המטריצה היא הפיכה, כי אם היא הפיכה, בהכרח מושקלות מקודם יש לה מטריצה הופכית, שאז נקבל:

$$y = Xw \Rightarrow X^{-1}y = X^{-1}Xw \Rightarrow w = X^{-1}y$$

הערה

נתעניין בדברים כאלה, למשל אם נחשוב על כל וקטור $x \in \mathbb{R}^d$ כאיזושהי תצפית (למשל, של מקום בצד הארץ), וכל ערכי y הם ערכי הטמפרטורה, נרצה למצוא קשר בין המקום ובין הטמפרטורה. למשל, בהינתן מקום חדש נרצה לדעת מהו הטמפרטורה שלו, על סמך הקשר שמצאנו. (כנראה שבדרך כלל הקשר לא יהיה לiniاري, ונרצה לקרב אותו לiniاري - תהליך שנקרא גרגסיה לiniארית, עליו נדבר בהמשך).

2.2 נורמות, מרחבי מכפלה פנימית והטלות

הגדרה

פונקציה בקבוצה X המוגדרת על ידי $d : X \times X \rightarrow \mathbb{R}^+$ נקראת **מטריקה** אם מתקיימים שלושת התנאים הבאים:

□ **חויביות:** $d(v, u) = 0 \Leftrightarrow v = u$

□ **סימטריה:** $d(v, u) = d(u, v)$

□ **אי שוויון המשולש:** $d(v, u) \leq d(v, w) + d(w, u)$

מהתנאים האלו עולה כי מטריקה היא בהכרח אי שלילית. בעקבות כך, נקרא למטריקה (או פונקציית המרחק) חיובית בהחלה. דוגמה מוכרת למטריקה היא פונקציית הערך המוחלט או המרחק האוקלידי:

תרגיל

יהיו $u, v \in \mathbb{R}^k$. הראו שפונקציית המרחק המוחלט, שמוגדרת על ידי סכום הערך המוחלט של הפרשי הוקטורים, קלומר $d(v, u) := \sum_{i=1}^n |v_i - u_i|$ הוא פונקציית מרחק.

הוכחה

ראשית, נבחן כי לכל $a, b \in \mathbb{R}$ מתקיים כי $|a - b| = 0$ אם ורק אם $a = b$.
בעקבות כך, d הוא הסכום של איברים אי שליליים שווים לאפס אם ורק אם כל האלמנטים שווים ל-0, אם ורק אם $u = v$.

סימטריה מגיעה ישירות מסימטריות של פונקציית הערך המוחלט.

כעת נוכחים א"ש המשולש. יהיו $w, u, v \in \mathbb{R}^k$ ולכן נקבל:

$$d(v, u) = \sum |v_i - u_i| = \sum |v_i - w_i + w_i - u_i| \stackrel{(*)}{\leq} \sum |v_i - w_i| + \sum |w_i - u_i| = d(v, w) + d(w, u)$$

כאשר (*) נובע מ"א"ש המשולש בפונקציית הערך המוחלט הרגילה.

כעת, נגדיר מהו גודל של וקטור.

הגדרה

נורמה היא פונקציה $: u, v \in \mathbb{R}^d \rightarrow \mathbb{R}_+$, שמקיימת את שלושת התכונות הבאות, לכל $a \in \mathbb{R}$ ולכל

□ **חויביות בהחלה:** $0 \leq \|v\| = 0$ אם ורק אם v הוא וקטור ה-0.

□ **הומוגניות:** $\|av\| = |a| \cdot \|v\|$

□ **אי שוויון המשולש:** $\|v + u\| \leq \|v\| + \|u\|$

ניתן לחשב על נורמה גם בתחום **המרחב מהרاسيית**, תחת פונקציית המרחק המוגדרת על ידי הנורמה.
מספר נורמות מוכרכות הינו:

□ נורמת הערך המוחלט $\|v\|_1 := \sum |v_i|$: (ℓ_1)

□ הנורמה האוקלידית $\|v\|_2 = \sqrt{\sum x_i^2}$: (ℓ_2)

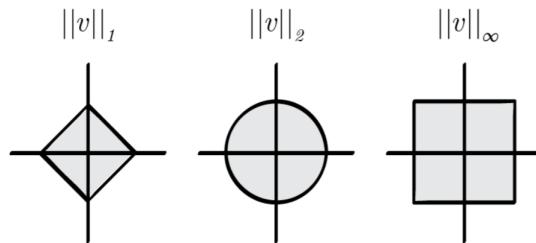
□ נורמת אינסוף: $\|x\|_\infty = \max |v_i|$.

נורמת הערך המוחלט והנורמה האוקלידית הן חלק משפחה של נורמות המכונה L_p , המוגדרת על ידי $\|v\|_p := (\sum |v_i|^p)^{\frac{1}{p}}$, כאשר $p \in \mathbb{N}$.

הגדרה

יהי V מרחב וקטורי ו- $\|\cdot\|$ נורמה על אותו המרחב.

. $B_{\|\cdot\|} = \{v \in V \mid \|v\| \leq 1\}$ קבוצה של וקטורים כך ש-



אחרי שדיברנו על 'גודל' של וקטורים, נרצה לדבר בפעם על 'מכפלה' של וקטורים.

הגדרה

מכפלה פנימית מעל מרחב וקטורי V ו- \mathbb{R}_+ עם העתקה $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}_+$ כך שלכל $v, u, w \in V$ מתקיים:

□ סימטריות: $\langle v, u \rangle = \langle u, v \rangle$

□ ליניאריות: $\langle \alpha v + w, u \rangle = \alpha \langle v, u \rangle + \langle w, u \rangle$

□ אי שליליות: $\langle v, v \rangle = 0 \Leftrightarrow v = 0$ ו- $\langle v, v \rangle \geq 0$

נניח כי ישנו דמיון בין מכפלה פנימית ובין נורמה. למעשה, בהינתן מרחב מכפלה פנימית כלשהו, אנו גם מקבלים נורמה על אותו המרחב.

טענה

תהי H מרחב מכפלה פנימית. אז הפונקציה $\| \cdot \| : H \rightarrow \mathbb{R}_+$ המוגדרת על ידי (לכל $v \in H$) $\|v\| = \langle v, v \rangle^{\frac{1}{2}}$ היא נורמה על H (הנורמה המושראית).

תרגיל

יהיו $v, u \in V$. הראו כי $\langle v, u \rangle = \|v\| \cdot \|u\| \cos \theta$, כאשר θ היא הזווית בין v ו- u .

הוכחה

תחילה, נזכיר בחוק הקוסינוסים. בהינתן משולש עם צלעות a, b, c נקבל:

$$c^2 = a^2 + b^2 - 2ab \cos \theta$$

על ידי הפעלת חוק הסינוסים על המשולש המוגדר על ידי v, u , ו- $v - u$, קיבל כי:

$$\|v - u\|^2 = \|v\|^2 + \|u\|^2 - 2\|v\| \cdot \|u\| \cos \theta$$

מצד שני, אנו יודעים כי:

$$\|v - u\|^2 = \langle v - u, v - u \rangle = \langle v, v \rangle - 2\langle v, u \rangle + \langle u, u \rangle = \|v\|^2 + \|u\|^2 - 2\langle v, u \rangle$$

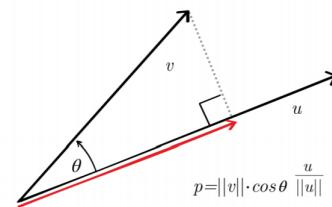
(כל המעברים נובעים מהגדירות מכפלות פנימיות שראינו קודם לכן) אם נשווה בין שתי המשוואות שקיבלנו, נוכל לראות כי $\cos \theta = \frac{\langle v, u \rangle}{\|v\| \cdot \|u\|}$, כנדרש.

מהאמור לעיל, קיבلנו ביטוי לזוית בין שני וקטורים, המשמש במכפלה הפנימית. נוכל גם להגיד מהי משמעות 'הטלה' בין וקטור אחד לשני. בשימוש בזיהות של $\cos \theta$ קיבל:

$$p \stackrel{(*)}{=} \|v\| \cos \theta \cdot \frac{u}{\|u\|} \stackrel{(**)}{=} \|v\| \frac{\langle v, u \rangle}{\|v\| \cdot \|u\|} \cdot \frac{u}{\|u\|} = \frac{\langle v, u \rangle}{\|u\|^2} \cdot u$$

המעבר הראשון (*) נובע אינטואטיבית מהגדירת קוסינוס, כאשר נרצה לקבל את 'הגודל' של הוktor, על ידי חלוקה בnormה.

המעבר השני (**) נובע מהזהות שקיבלנו קודם לכן. קל לראות זאת באמצעות הציור הבא:



הגדרה

טללה של וקטור v על וקטור u היא הוktor p באורך $\cos \theta \|v\|$ בכיוון של u .

נניח שבמקרה המיוחד המינוח של $\theta = 90^\circ$, קיבל כי $\langle v, u \rangle = 0$. במקרה זה, נאמר כי v והין **אורתוגונליים**, ונשתמש בסימון $v \perp u$.

אם v, u הם גם וקטורי יחידה, נאמר כי הם **אורתוגונרמיים** אחד לשני.

הגדרה

מטריצה אורתוגונלית היא מטריצה ריבועית כאשר העמודות הימן וקטורי יחידה שאורתוגונליים אחד לשני (כלומר, הימן וקטוריים אורתוגונרמיים), וגם השורה הימן וקטורי יחידה שאורתוגונליים אחד לשני.

למה

תהי $A \in \mathbb{R}^{d \times d}$ מטריצה אורתוגונלית, אזי $AA^T = I = A^T A$.

בצירוף שתי ההגדרות של הטלה ומטריצה אורתוגונלית, נוכל לקבל את ההגדרה של 'הטלה אורתוגונלית' של וקטור על מרחב ליניארי כלשהו.

הגדרה

תהי V תת מרחב בממד k של \mathbb{R}^d ויהיו v_1, \dots, v_k וקטורי בסיס אורתונורמלי של V . נגיד $P = \sum_{i=1}^k v_i v_i^T$. אזי המטריצה P היא מטריצת הטלת האורתוגונלית למוחב V .

למה

יהיו v_1, \dots, v_k קבוצת וקטורים אורתונורמלית ו- $P = \sum_{i=1}^k v_i v_i^T$ קיימות התכונות הבאות:

□ P היא סימטרית.

$$P^2 = P \quad \square$$

□ הערכים העצמיים של P הם 0 או 1. v_1, \dots, v_k הם וקטורים עצמיים של P המותאים לערך העצמי 1.

$$(I - P)P = 0 \quad \square$$

□ לכל $x \in \mathbb{R}^d$ ולכל $V \in u$ מתקיים כי $\|x - Px\| \geq \|x - u\|$.

$$x \in V \Rightarrow Px = x \quad \square$$

(אלו למעשה שילוב של תכונות של הטלה ומטריצה סימטרית, שראינו בליניארית עם המשפט הספקטרלי וכו').

הערה

כדי להבחן כי ההגדרה של מטריצת הטלת השתמשה בהגדרה של סכום מכפלה חיצונית.

1.3 הצגה מטרציאונית ג'

הגדרה

תהי A מטריצה ריבועית. נאמר כי A מטריצה לכסינה אם קיימת מטריצה P כך $AP^{-1}AP = P$ היא אלכסונית.

לפניהם נגדר את הקשר בין לכסינות וערכים עצמיים, נזכר בכלל בהגדרה שלערכים עצמיים.

הגדרה

תהי A מטריצה ריבועית. נאמר כי וקטור $v \in V \neq 0$ הוא וקטור עצמי של A המותאים לערך העצמי $\lambda \in \mathbb{R}$ אם מתקיים כי $Av = \lambda v$.

טענה

תהי A מטריצה ריבועית סימטרית. אזי קיימים בסיס אורתונורמלי $\mathbb{R}^d = u_1, \dots, u_n$ שלערכים עצמיים של A .

משפט (EVD)

תהי $A \in \mathbb{R}^{d \times d}$ מטריצה סימטרית מעל המשיים. אז קיימת מטריצה אורתונורמלית $U \in \mathbb{R}^{d \times d}$ ומטריצה אלכסונית D כך ש- $n \dots 1 = D_{i,i}$ הם הערכים העצמיים של A וגם $A = UDU^T$.

הפרק הזה של A נקרא הפירוק של A לערכים עצמיים. פירוק זה נמצא בשימוש נפוץ וייש לו תכונות חזקות. למשל, קל להשתמש בו על מנת לחשב חזקות של A :

$$A^k = UDU^T \cdot UDU^T \cdot UDU^T = UD^kU^T$$

כעת, נרצה להציג מושג שיאפשר לנו להכליל את ההגדלה הקודמת גם למטריצות שאין סימטריות ואולי שאין ריבועיות.

הגדרה

תהי $A \in \mathbb{R}^{m \times d}$ ו- $v \in \mathbb{R}^d$ ו- $u \in \mathbb{R}^m$ וקטורי יחידה. נאמר כי u , v הם וקטורים סינגולריים ימניים ושמאליים בהתחיימה, אם מתקיים ביחס לערך הסינגולרי $\sigma \in \mathbb{R}_+$ כי $u\sigma = v$.

טענה (SVD)

תהי $A \in \mathbb{R}^{m \times d}$ מטריצה מעל המשיים. אז היא יכולה להכתב בהצגה הסינגולרית הבאה- $A = U\Sigma V^T$, כאשר $U \in \mathbb{R}^{m \times m}$ ו- $V \in \mathbb{R}^{d \times d}$ הינם מטריצות אורתונורמליות ו- $\Sigma \in \mathbb{R}^{m \times d}$ היא מטריצה אלכסונית עם ערכים לא שליליים. אלו נקראים הערכים הסינגולריים של A .

(איינטואיציה - נזכיר כי ערכים אורתונורמליים הם למעשה שיקוף או סיבוב - מליניארית 2, הערכים העצמיים הם רק 1 או -1).

כמו כן, נזכיר כי הדרך שלנו 'להעביר' סקלרים למטריצה היא להשתמש במטריצה אלכסונית. במקרה שלנו, מדובר במקרה מלכנית, כך שהוא לא אלכסונית הרגילה לנו. במקרה כזה, המטריצה האלכסונית עשויה למעשה סוג של 'מתיחה'.

לכן, בפרט במקרה שלנו מתבצע 'סיבוב', מתיחה ו'סיבוב' נסף.)

טענה

תהי $A = U\Sigma V^T$ יציג הערך הסינגולרי של A . מכאן נובע כי העמודות של U והשורות של V^T הם וקטוריים סינגולריים שמאליים וימניים של A , בהתאם לערכים הסינגולריים המיוצגים באקסון של Σ .

(איינטואיציה - אם נכפול את שני האגפים ב- V , נקבל $U = \sum_i u_i \sigma_i v_i^T$, ובבדיקה נקבל לכל עמודה i כי $Av_i = u_i \sigma_i$ (בדיקת הוקטורים הסינגולריים הדרושים)

נניח שהדרגה של A שווה ל- r . מכאן נובע כי המספר של ערכים סינגולריים שאינם אפס שווה ל- r . כמו כן, נבחן כי $\min\{d, m\} \leq r$, כאשר בה"כ $m \leq d$ או $d \leq m$ (שניהם מטריצות רחבות יותר (יש להן יותר שורות ועמודות):

$$A = U\Sigma V^\top = \left[\begin{array}{c|ccccc} & & & & & \\ \hline u_1 & \cdots & u_r & \cdots & u_m & \\ \hline & & & & & \end{array} \right] \left[\begin{array}{ccc|c} \sigma_1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & 0 \\ 0 & \cdots & \sigma_r & \\ \hline & & & 0 & \cdots & 0 \\ 0 & & & \vdots & \ddots & \vdots \\ & & & 0 & \cdots & 0 \end{array} \right] \left[\begin{array}{ccc} - & v_1^\top & - \\ \vdots & & \vdots \\ - & v_r^\top & - \\ \vdots & & \vdots \\ - & v_d^\top & - \end{array} \right]$$

כיוון שככל הערכים הסינגולריים $\sigma_m, \dots, \sigma_{r+1}, \dots, \sigma_r$ הם אפסים, כל הערכים הסינגולריים הימניים והשמאליים שגדולים מ- r נכפלים ב-0 ואין להם משמעות.

לכן המידע החשוב שנקבע ב- SVD מיוצג במטריצה $r \times r$ קטנה, שלעיתים מכונה SVD קומפקטיבית של A , שנייה נתה היכתב גם באמצעות:

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top = \overbrace{\left[\begin{array}{c|cc} & & \\ \hline u_1 & \cdots & u_r \\ & & \end{array} \right]}^{m \times r} \left[\begin{array}{ccc} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{array} \right] \overbrace{\left[\begin{array}{ccc} - & v_1^\top & - \\ \vdots & & \vdots \\ - & v_r^\top & - \end{array} \right]}^{r \times d}$$

בשביל להימנע מבלבול, השתמש בסימון המקורי ונתיחס מראש ל- V , $\sum U$ בהצגה הקומפקטיבית.

שתי הצורות שראינו קשורות זו לזו, כפי שניתן לראות מהלמה הבאה (שגם מראה כי ניתן לחשב את ה- SVD בזמן פולינייאלי ביחס ל- m ו- d).

למה
 $A \in \mathbb{R}^{m \times d}$ SVD - $A = U\Sigma V^\top$ של $A^T A = V\Sigma^T \Sigma V^\top$ ו- $A A^T = U\Sigma \Sigma^T U^\top$ הוא ה- EVD של A (או EVD של A^T)
(אינטוואיציה - פשוט כפל בטרנספורם שמתהprec)

מחלמה זאת עולה כי הערכים העצמיים של $A^T A$ ו- $A A^T$ שוים לריבוע של הערכים הסינגולריים של A .
בנוסף, כיוון שהמטריצות האורתוגונליות של ה- EVD מכילות את הוקטורים העצמיים של המטריצה, הוקטורים העצמיים של $A A^T$ הם הוקטורים הסינגולריים השמאליים של A , והוקטורים העצמיים של $A^T A$ הם הוקטורים הסינגולריים הימניים של A .

2 אינפי

לעתים קרובות במהלך שימוש בטכניקות שונות של 'מערכות לומדות', השתמש בפונקציות מכמה ממדים $f : \mathbb{R}^d \rightarrow \mathbb{R}$.

הפונקציות הללו יסמלו את המרחק בין הערכים האמתיים ובין ערכי החיזוי שלנו. בעקבות כך, נרצה לסייע את המרחק על מנת לקבל תוצאה אופטימלית.
בקיצור, נרצה לפטור את הבעה $\arg \min_{w \in \mathbb{R}^d} f(w)$.

2.1 נגזרות, גרדיאנטים ויעקביאנים

2.1.1 נגזרות

הדרך הנפוצה למצוא דבר מינימלי היא לחשב את הנגזרת, להשוות לאפס ולפתרו את המשוואה. כאשר השתמש בפונקציות ככמה משתנים, השתמש לרוב בגרדיינט ולא בנגזרת סקלרית (כלומר, כאשר נציב x מסוים).

הגדרה

תהי $f : \mathbb{R} \rightarrow \mathbb{R}$. הנגזרת של f בנקודה $x \in \mathbb{R}$ מוגדרת על ידי:

$$\frac{d}{dx} f(x) = \lim_{a \rightarrow 0} \frac{f(x+a) - f(x)}{a}$$

דוגמה

ניקח את הפונקציה ReLU שמוגדרת על ידי $\max(0, x)$ כחלק החיובי של הערך. ככלומר, הנגזרת של הפונקציה הינה:

$$\frac{df(x)}{dx} = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

נבחן כי בנקודה $x = 0$ הנגזרת איננה מוגדרת. בהמשך נדוע כיצד אפשר להתמודד עם מקרים כאלה.

כעת, נעבור לדען בהגדרות המתאימות לפונקציות ככמה משתנים.

הגדרה

תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$. הנגזרת החלקית של f בנקודה $x \in \mathbb{R}^d$ ביחס ל- x_i , מוגדרת על ידי:

$$\frac{\partial}{\partial x_i} f(x) = \lim_{a \rightarrow 0} \frac{f(x + ae_i) - f(x)}{a} = \lim_{a \rightarrow 0} \frac{f(x_1, \dots, x_i + a, \dots, x_d) - f(x_1, \dots, x_i, \dots, x_d)}{a}$$

כאשר e_i הוא וקטור הבסיס הסטנדרטי ה- i .

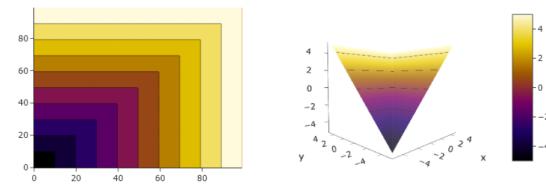
למעשה, הנגזרת החלקית היא הנגזרת ביחס לאחד המשתנים, כאשר שאר המשתנים נשארים קבועים.

דוגמה

ניקח את הפונקציה $f(x) = \max(x_1, \dots, x_d)$. הנגזרות החלקיות של הפונקציה הינה:

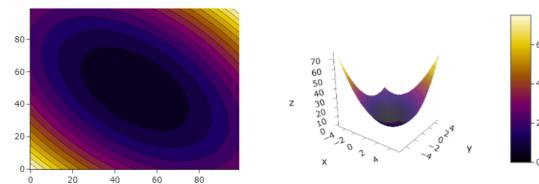
$$\frac{\partial}{\partial x_i} f(x) = \begin{cases} 1 & i = \operatorname{argmax}(x_1, \dots, x_d) \\ 0 & i \neq \operatorname{argmax}(x_1, \dots, x_d) \end{cases}$$

בתמונה, זה נראה כך:

**דוגמה נוספת**

תהי $f(x, y) = x^2 + xy + y^2$. הנגזרות החלקיות של f בנקודה (x_0, y_0) הינו:

$$\frac{\partial}{\partial x} f(x_0, y_0) = 2x_0 + y_0, \quad \frac{\partial}{\partial y} f(x_0, y_0) = x_0 + 2y_0$$



(אפשר לראות אינטואיטיבית את קווי הגובה, דבר שיתחבר לנו בהמשך)

2.1.2 גראדינטים**הגדרה**

הגרדיינט של f בנקודה x הוא וקטור הנגזרות החלקיות:

$$\nabla f(x) := \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_d} \right)$$

(הגרדיינט מציג את כיוון העלייה הגדול ביותר ובעקבות כך, הגרדיינט תמיד ניצב לקווי הגובה, כיוון שהקוויים מייצגים את המקרה בו 'איןנו גדלים')

דוגמה

על ידי שימוש בנגזרות החלקיות שהשכחנו קודם לכן, של הפונקציה $y^2 + xy + x^2$, נקבל:

$$\nabla f(t_0) = (2x_0 + y_0, 2y_0 + x_0)^T$$

תרגיל

תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ המוגדרת על ידי $f(x) = w^T x$. חשבו את הגרדיינט של f בנקודה x .

הוכחה

מליניאריות הנגזרת¹ נקבל:

¹נגזרת של סכום היא סכום הנגזרות, ונבחין שכיוון שמדובר במכפלה פנימית מדובר בסכום

$$\frac{\partial}{\partial x_j} f(x) = \sum_i \frac{\partial}{\partial x_j} f(x)_i = \sum_i \frac{\partial}{\partial x_j} w_i x_i = w_j$$

כלומר, הגרדיינט הינו $w = \nabla f(x)$.

תרגיל

תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ המוגדרת על ידי $f(x) = \|x\|^2$. חשבו את הגרדיינט של f בנקודה x .

הוכחה

בדומה לתרגיל הקודם, משתמש בליינאריות הנגזרת:

$$\frac{\partial}{\partial x_j} f(x) = \sum_i \frac{\partial}{\partial x_j} x_i^2 = 2x_j$$

איןטאיטיבית, אפשר לומר כי אנחנו מתחבננים בכל אחד מהרכיבים וגוזרים ביחס אליו. הרि הנורמה הסטנדרטית מוגדרת באמצעות $x_1^2 + x_2^2 + \dots + x_d^2$. כאשר נזכיר ביחס לכל אחד מהמשתנים, קיבל בגירה הרלוונטי $2x_j$ ובכל השאר אפס.

לכן, סך הכל ניתן לכתוב את הגרדיינט בתור $\nabla f(x) = 2x$.

2.1.3 יעקוביאנים

לעתים רבות, הפונקציה שנרצה למשוך תהיה פונקציה שמורכבת ממשפּר פונקציות, דהיינו $f : \mathbb{R}^m \rightarrow \mathbb{R}^d$. למשל, נניח שצברנו מידע שיכיל תוכן על עונות השנה, זמני היום והמקומות (קווי אורך ורוחב) ונניח שיש פונקציה שמקשרת בין המיקומים והזמן לחץ האוויר והטמפרטורה. במלילים אחרים, מדובר על פונקציה $f : \mathbb{R}^4 \rightarrow \mathbb{R}^2$. על מנת לפטור את הבעיה, נגידיר את המושג הבא.

הגדרה

תהי $f : \mathbb{R}^m \rightarrow \mathbb{R}^d$ כasher $f(x) = (f_1(x), \dots, f_d(x))^T$ הוא מטריצת הנגזרות החלקיים:

$$J_x(f) := \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_d(x)}{\partial x_1} & \dots & \frac{\partial f_d(x)}{\partial x_m} \end{bmatrix}$$

דוגמה

נזכיר בדוגמה מקודם שראינו, כאשר $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ שמוגדרת על ידי $f(x) = x_1^2 + x_2^2$. היעקוביאן הינו $J_x(f) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_2 & 2x_1 \end{bmatrix}$. לעומת זאת, היעקוביאן הוא הטורנספו של הגרדיינט (כאשר $d=1$, אין הגדלה של הגרדיינט כלל).

² שימו לב ששמדוור בגדירה שונה ממנה שהגדרכנו בקורסי האינפי למיניהם. בימנו הגדרכנו את היעקוביאן בטור הדטרמיננטה של המטריצה המדוברת - היעקוביאן כאן הוא פישוט הדיפרגיאל המוכר לנו. בקיצור, תאמינו למתמטיקאים. סתם, יש בלבול בין זה ובין מטריצת יעקובי, שמוגדרת גם כאשר הפונקציה לא דיפרגיאלית בהכרח.

תרגיל

תהי $A \in \mathbb{R}^{m \times d}$ המוגדרת על ידי $f(x) = Ax$. מצאו את היעקוביאן של f .

הוכחה

תחילה, נבחן כי לכל $i \in [d]$ מתקיים כי $f_i(x) = A_i^T X$, וכך מוגדרת f כ:

$$J_x(f) = \begin{bmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_d(x)^\top \end{bmatrix} = \begin{bmatrix} -A_1 - \\ \vdots \\ -A_{dm} - \end{bmatrix} = A$$

2.1.4 כלל השרשראת**טענה - כלל השרשראת למשתנה אחד**

תהי $f : \mathbb{R} \rightarrow \mathbb{R}$ ו- $g : \mathbb{R} \rightarrow \mathbb{R}$ שתי פונקציות גזירות, אז הנגזרת של ההרכבה $f \circ g$ הינה:

$$(f \circ g)' := (f' \circ g) \cdot g'$$

במילים אחרות, אם $h(x) = f(g(x)) \cdot g'(x)$ אז לכל $x \in \mathbb{R}$ מתקיים כי $h(x) = f(g(x))$

טענה - כלל השרשראת לפונקציות בכמה משתנים

תהי $f : \mathbb{R}^k \rightarrow \mathbb{R}^d$ ו- $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ היעקוביאן של ההרכבה $(f \circ g) : \mathbb{R}^k \rightarrow \mathbb{R}^m$ מוגדר על ידי:

$$J_x(f \circ g) = J_{g(x)}(f) J_x(g) := \begin{bmatrix} \frac{\partial f_1(g(x))}{\partial g_1(x)} & \dots & \frac{\partial f_1(g(x))}{\partial g_d(x)} \\ \vdots & & \vdots \\ \frac{\partial f_m(g(x))}{\partial g_1(x)} & \dots & \frac{\partial f_m(g(x))}{\partial g_d(x)} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \dots & \frac{\partial g_1(x)}{\partial x_k} \\ \vdots & & \vdots \\ \frac{\partial g_d(x)}{\partial x_1} & \dots & \frac{\partial g_d(x)}{\partial x_k} \end{bmatrix}$$

תרגיל

יהי $f(x) = ||x||^2$ עבור $x \in \mathbb{R}^m$. חשבו את היעקוביאן של $f \circ g$.

הוכחה

נשתמש בכלל השרשראת. ראשית, נבחן כי $J_x(g) = A$ ממה שפרטנו קודם לכן. נבחן כי $\text{Im}(f) \subseteq \mathbb{R}$, היעקוביאן של f שווה לטרנספורם של הגראונט.

כמו כן, ראיינו כי $J_{g(x)}(f) = (2g(x))^T$. בukt, מתקיים מכל השרשראת:

$$J_x(f \circ g) = J_{g(x)}(f) \cdot g'(x) = (2Ax)^T \cdot A = 2x^T A^T \cdot A$$

כנדרש.

תרגום

חשבו את הגרדיאנט של הפונקציה הבאה $f(x) = \frac{1}{2} \|Ax - y\|^2$. בהמשך נראה כי מדובר בפונקציה לחישוב טעות ריבועית ממוצעת³.

הוכחה

לשם הנוחות, נסמן $h(x) = Ax$. $g(x) = \|Ax\|^2$, $h(x) = \|x\|^2$. ראיינו כי מתקיים:

$$\begin{aligned} J_x(g \circ h) &= 2x^T A^T A \Rightarrow (J_x(g \circ h))^T = \\ &(x^T A^T A)^T = 2A^T Ax \end{aligned}$$

וגם A היא מטריצה لكن קיבל כי

$$\begin{aligned} J_x(y^T Ax) &= 2y^T A \Rightarrow (J_x(2y^T Ax))^T = \\ &(2y^T A)^T = 2A^T y \end{aligned}$$

לכן קיבל בסך הכל כי :

$$\nabla f(x) = (J_x(g \circ h))^T - (J_x(2y^T Ax))^T = 2A^T Ax - 2A^T y$$

כנדרש.

דוגמה - פונקציית Soft-Max

במידת מוכנה, משתמש לעיתים בפונקציה $S : \mathbb{R}^d \rightarrow [0, 1]^d$ מחזירה וקטור שהקווארדינטות שלו נסכמוות ל-1. דבר זה מוגדר על ידי :

$$S(\mathbf{a})_j = \frac{e^{a_j}}{\sum_{k=1}^N e^{a_k}}$$

כיון שהמקדמים תלויים בפונקציית האקספוננט, כל הערכים שמתזקבים מהפונקציה הינם חיוביים בהחלה. מעבר לכך, עובדה או נוררת כי כל הערכים הם בתחום (0, 1) (שהרי הווקטור נסכם ל-1). כך למשל, אם נפעיל את הפונקציה על (1, 2, 5) קיבל את (0.02, 0.05, 0.93). ניתן לראות כי היחס סדר בין האיברים נשמר, וכי הם נסכימים ל-1, כפי שרצינו ואפשר גם לשים לב כי ההפעלה על 5 רוחקה מאוד מההפעלה על 2.

aintoaitivit, פונקציית הסופטמакс היא גרסה מעודנת של argmax. במקומות לבחור את הערך המקסימלי ביותר, פונקציית הסופטמакс מפרקת את הווקטור לחלקים, כשהערך המקסימלי מקבל 'נתה' משמעותית, אך גם שאר חלקיו הווקטור מקבלים חלק מסוים. משתמשים בפונקציה זו ברשותות נירוניים (לחילק פלט לא מנורמל להתפלגות מתאימה).

כעת, נרצה למצוא את הנגזרת של פונקציה זו. נסמן $h(a) = \sum_{k=1}^d g_i(a)$ ו- $g_i(a) = e^{a_i}$. נקבל:

Mean Square Error-MSE³

$$\frac{\partial S_i}{\partial a_j} = \frac{\partial}{\partial a_j} \frac{e^{a_i}}{\sum_{k=1}^N e^{a_k}} = \frac{\partial}{\partial a_j} \frac{g_i}{h}$$

(בתכל"ס, אפשר להסתכל על כל קוארדינטה מכפלה של $\frac{1}{\sum_{k=1}^N e^{a_k}} \cdot e^{a_j}$ במקומות $-j$ ואז נתבונן בנגזרת מכפלה).
נבחין כי הנגזרת של המונח, כולם של h היא e^{a_j} כאשר $j = i$ ו- $i \neq j$ כאשר $j = i$:
לכן נקבל, במקומות $-j$, כאשר נגזר לפי $j = i$:

$$\frac{\partial}{\partial a_j} \frac{e^{a_i}}{\sum_{k=1}^N e^{a_k}} = \frac{e^{a_i} \left(\sum_{k=1}^N e^{a_k} \right) - e^{a_i} e^{a_j}}{\left(\sum_{k=1}^N e^{a_k} \right)^2} = \frac{e^{a_i}}{\left(\sum_{k=1}^N e^{a_k} \right)} \cdot \frac{\left(\sum_{k=1}^N e^{a_k} \right) - e^{a_j}}{\left(\sum_{k=1}^N e^{a_k} \right)} = S_i (1 - S_j)$$

כאשר $j \neq i$, נקבל בנגזרת המכפלה של $0 - \frac{1}{\sum_{k=1}^N e^{a_k}} \cdot e^{a_j}$ במחובר הראשון, $-j$ במחובר השני.
כנדרש.

2.2 קירוב מסדר ראשון

כפי שראינו כבר קודם לכן, לעיתים נתעניין במצבית הממצאים של פונקציות בכמה משתנים. לעיתים, חלק מהfonקציות האלה מודוס מושכות או אף בלתי אפשריות לחישוב בצורה אנליטית. לשם כך, משתמש בקירוב של הפונקציה באמצעות פונקציה פשוטה יותר שנוכל לפתור.

תהיה $f : \mathbb{R} \rightarrow \mathbb{R}$. נזכיר⁴ בהגדירה של טור טיילור:

$$T(x_0 + x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} x^n = f(x_0) + f'(x_0)x + \frac{1}{2}f''(x_0)x^2 + \dots$$

הקירוב הליינרי (או קירובה מסדר ראשון) היא קירוב של פונקציה כלשהיא בשימוש בפונקציה ליינרית. עבור פונקציה $f : \mathbb{R} \rightarrow \mathbb{R}$ שגזרה פעמיים ברציפות, עולה משפט טיילור כי:

$$f(x_0 + x) \approx f(x_0) + f'(x_0)x$$

נוכל כעת להרחיב את ההגדירה הזאת על מנת להגיד קירובים ליינריים של פונקציות בכמה משתנים.

הגדרה - קירוב ליינרי

תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ו- $p_0 \in \mathbb{R}^d$. הקירוב הליינרי של f בכל p קרוב לפ-0 הינו:

⁴עבור חלקנו זה יהיה זכרון לשימושם לא למדנו.

$$f(p_0) + \langle \nabla f(p_0), p - p_0 \rangle$$

בצורה דומה, אם נתיחס ל- p כ'סטייה' מ- p_0 , נקבל:

$$f(p_0 + p) \approx f(p_0) + \langle \nabla f(p_0), p \rangle$$

לדוגמא, אם f היא פונקציה ליניארית עצמה, נוכל להבין אינטואיטיבית כי הקירוב הליניארי יהיה הפונקציה עצמה. תהי $f : \mathbb{R}^d \rightarrow \mathbb{R}$ -ו $b \in \mathbb{R}^d$ המוגדרת על ידי $f(x) = b^T x$. אזי הקירוב הליניארי הינו:

$$\begin{aligned} f(p_0) + \langle \nabla f(p_0), p - p_0 \rangle &= \\ b^T p_0 + \langle b, p - p_0 \rangle &= \\ b^T (p_0 + p - p_0) &= \\ b^T p \end{aligned}$$

כנדרש.

דוגמה נוספת

תהי $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ המוגדרת על ידי $f(x, y) = \sqrt{x^2 + y^2}$. חשבו את הקירוב הליניארי של f קרוב ל-(3, 4) הוכחה תחילה, נחשב את הגרדיאנט של f ⁵. הנגזרות החלקיות לפי x ולפי y בנקודות x_0 ו- y_0 הינו בהתאם $\frac{2x_0}{\sqrt{x_0^2 + y_0^2}}$ ולכן הגרדיאנט הינו $\left(\frac{2x_0}{\sqrt{x_0^2 + y_0^2}}, \frac{2y_0}{\sqrt{x_0^2 + y_0^2}} \right)^T$ אם כך, עבור הנקודה (3, 4) כאשר נציב נקבל:

$$f(3 + x, 4 + y) \approx 5 + \frac{3}{5}x + \frac{4}{5}y$$

אם $x = 0.1$ ו- $y = 0.2$, אז:

$$f(3 + 0.1, 4 + 0.2) = 5.2201 \approx 5.22 = 5 + \frac{3}{5} \cdot 0.1 + \frac{4}{5} \cdot 0.2$$

נוכל להשתמש בקירובים גם לטובת הדבר הבא. נניח שחקרנו פונקציה $f : \mathbb{R}^d \rightarrow \mathbb{R}$ בנקודה $p_0 \in \mathbb{R}^d$. נרצה 'לקחת צעד' ב- \mathbb{R}^d בכיוון בו f גדלה בקצב המהיר ביותר. כיצד נוכל למצוא את הכיוון זהה?

⁵במהלך אינפי 3 ראיינו לסבר נוח יותר לחישוב הפולינום טילור הדורש. צריך לחשב את הנגזרות החלקיות לפי x ולפי y ולהציב במקום המתאים בפולינום. למעשה זה מה שעשינו כאן, אלא שהחישוב באמצעות גראדיאנט עלול לבלבול.

נזכיר כי $p \cdot f(p_0 + p) \approx f(p_0) + \nabla f(p_0) \cdot p$. בנוסח, הزاوية בין $\nabla f(p_0)$ ו- p היא:

$$\nabla f(p_0) \cdot p = \|\nabla f(p_0)\| \cdot \|p\| \cos \theta$$

כיוון שאנו מתעניינים בכיוון בלבד ולא בגודל הוקטור, נניח כי $|p| = 1$, לשם הפשטות. כמו כן, נשים לב כי $f'(x) \in [-1, 1]$ ולכן הכוון ש'מקסם' את f הינו $p_{\max} := \frac{f(p_0)}{\|\nabla f(p_0)\|}$. כך, במקרה דומה הכיוון ש'מינימע' את f הוא $p_{\min} := -\frac{f(p_0)}{\|\nabla f(p_0)\|}$. כמובן, על מנת ל赞美ר או למקסם את f באמצעות 'יעדים קטנים', לכלת בכיוון של הגרדיאנט או בכיוון המנוגד לגרדיאנט זה רעיון טוב.⁶

3 הסתברות

חלק משמעותי מעקרונות מערכות למדות מבוססים על עקרונות של סטטיסטיקה והסתברות.

3.1 עקרונות בסיסיים

מורחבי הסתברות

מורחבי הסתברות מורכבים משני מרכיבים מרכזיים. מרחב מדגם ופונקציית הסתברות.

הגדרה

מרחב מדגם Ω הוא קבוצה שמכילה את כל התוצאות האפשרות. $\Omega \subseteq w$ מסמלת תוצאה יחידה.

הגדרה

מאורע A הוא תת קבוצה של התוצאות האפשרות, כלומר $A \subseteq \Omega$.

הגדרה

מרחב הסתברות הוא זוג (Ω, \mathcal{D}) כאשר Ω זה מרחב המדגם ו- $\mathcal{D} : 2^\Omega \rightarrow \mathbb{R}$ היא פונקציית הסתברות כך ש:

$$\mathcal{D}(\Omega) = 1 \quad \square$$

$$\square \text{ לכל } \Omega \in \omega \text{ מתקיים כי } \mathcal{D}(\omega) \in [0, 1]$$

$$\mathcal{D}(A \cup B) = \mathcal{D}(A) + \mathcal{D}(B) \quad \square \text{ אם } A, B \subseteq \Omega \text{ ש-}\emptyset = A \cap B = \emptyset \text{ מתקיים כי}$$

בהתבסס על ההגדרה הזאת, נוכיח את השימוש הכלכלי וההדרה.

טענה

לכל $\Omega \subseteq A, B \subseteq \Omega$ מתקיים כי $\mathcal{D}(A \cup B) = \mathcal{D}(A) + \mathcal{D}(B) - \mathcal{D}(A \cap B)$

הוכחה

נוכיח כי מתקיים:

$$A = (A \setminus B) \cup (A \cap B)$$

$$B = (B \setminus A) \cup (A \cap B)$$

⁶ אזכיר במאמר שבדנו לגבי הבדיקה שהגרדיאנט מאונך לקווי הגובה והוא מבטא את קצב השינוי המקסימלי. ישacha סרטון ביוטיוב שמודים את זה.

כמו כן, מתקיים כי $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$
ולכן, נקבל סך הכל:

$$\begin{aligned}\mathcal{D}(A \cup B) &= \mathcal{D}(A \setminus B) + \mathcal{D}(B \setminus A) + \mathcal{D}(A \cap B) \\ &= \mathcal{D}(A) - \mathcal{D}(A \cap B) + \mathcal{D}(B) - \mathcal{D}(A \cap B) + \mathcal{D}(A \cap B) = \\ &= \mathcal{D}(A) + \mathcal{D}(B) - \mathcal{D}(A \cap B)\end{aligned}$$

כנדרש.

טענה - חסם האיחוד
יהי (Ω, \mathcal{D}) מרחב הסתברות. לכל קבוצה (A_k) של מאורעות, מתקיים כי:

$$\mathcal{D}(\cup_{k=1}^{\infty} A_k) \leq \sum_{k=1}^{\infty} \mathcal{D}(A_k)$$

הוכחה

נגידר $B_3 = A_3 \setminus (A_2 \cup A_1)$, ומכאן והלאה נגדיר $B_2 = A_2 \setminus A_1$ וגם $B_1 = A_1$ ולכן נקבל:

$$\mathcal{D}(\cup_{k=1}^{\infty} A_k) = \mathcal{D}(\cup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} \mathcal{D}(B_k) \leq \sum_{k=1}^{\infty} \mathcal{D}(A_k)$$

הגדירה

נאמר כי מאורעות $A, B \subseteq \Omega$ הם בלתי תלויים אם מתקיים:

$$\mathcal{D}(A \cap B) = \mathcal{D}(A) \cdot \mathcal{D}(B)$$

תרגיל

הראו כי אם A ו- B בלתי תלויים, אז גם $A \cup B^c$ בלתי תלויים.

הוכחה

מתקיים:

$$\mathcal{D}(A) = \mathcal{D}(A \cap B) + \mathcal{D}(A \cap B^c)$$

ולכן בפרט:

$$\mathcal{D}(A \cap B^c) = \mathcal{D}(A)(1 - \mathcal{D}(B)) = \mathcal{D}(A) \cdot \mathcal{D}(B^c)$$

משתנים מקרים

באופן כללי, דיברנו עד כה על השאלה האם מאורע מסוים התקיים או לא ושאלנו מה ההסתברות לכך. נרצה לשאול שאלות אחרות. למשל, להסיק מידע נרחב יותר על סמך המאורעות. למשל, נוכל לשאול אם אנחנו מקבלים H בהטלה, קיבל דולר, כמה דולרים נקבל לאחר שלוש הטלות. על מנת לענות על השאלה הזאת, נוכל להגיד את מרחב המדגם:

$$\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{THH}, \text{HTT}, \text{THT}, \text{TTH}, \text{TTT}\}$$

לאחר מכן, נוכל להגיד פונקציה שתספור כמה פעמים התקבל H .
למשל:

$$(\omega \in \Omega) : X(HHH) = 3, X(HHT) = 2$$

הפונקציה X מוגדרת כמשתנה מקרי.

הגדרה

בහינתן מרחב הסתברות (Ω, \mathcal{D}) , המשתנה המקרי הוא פונקציה $X : \Omega \rightarrow \mathbb{R}$.

הגדרה

יהי Ω מרחב הסתברות בדיד ו- X משתנה מקרי מעל Ω . פונקציית התפלגות PMF של X מוגדרת על ידי :

$$\mathcal{D}(\{X = x\}) = \sum_{\omega: X(\omega) = x} \mathcal{D}(\omega)$$

במקום לכתוב $\{X = x\}$ נוכל פשוט להשתמש בסימון $\mathcal{D}_X(x)$ או $\mathcal{D}(x)$ אך זה נוח רק כאשר הפונקציה מוגדרת וברורה היטב).

דוגמה

טיפילים מטבע הוגן פעמיים. מרחב המדגם הינו $\Omega = \{T, H\}^2$.
כיון שמדובר במטבע הוגן, אזי לכל $\omega \in \Omega$ מתקיים כי $\mathcal{D}(\omega) = \frac{1}{4}$. נגיד את X להיות מספר הראשים שמתוקבים.

הערכים האפשריים של X הינם 0, 1, 2, ולכן נקבל:

$$\mathcal{D}(x) = \begin{cases} \frac{1}{4} & x = 0, 2 \\ \frac{1}{2} & x = 1 \\ 0 & \text{else} \end{cases}$$

הגדרה

יהי Ω מרחב הסתברות רציף ו- X משתנה מקרי מעל Ω .
נאמר ש- X הוא משתנה מקרי רציף, אם קיימת פונקציה $0 \leq f(x) \leq \infty$ כך שנוכל לכתוב, לכל \mathbb{R}

$$\mathcal{D}(X \in S) = \int_S f(x) dx$$

פונקציה זו מכונה בתור פונקציית הצפיפות (PDF) של X .

במילים אחרות, דבר זה אומר כי לכל \mathbb{R} $a, b \in \mathbb{R}$ מתקיים כי $\int_a^b f(x) dx$ על מנת להראות כי מדובר על פונקציה שקשורה ל- X , אמם גם כן כאן כמו קודם, נוכל לסמן את f בתור $f_X(x)$ על מנת להראות כי מדובר על פונקציה שקשורה ל- X .
בקורס שלנו, נניח כי f היא פונקציה רגילה ולכן פונקציית הצפיפות בכל נקודה שווה לאפס.
כלומר:

$$\mathcal{D}(X = a) = \int_a^a f(x) dx = 0$$

לכל $a \in \mathbb{R}$
נבהיר כי פונקציית הצפיפות מקיימת:

$$\begin{aligned} f(x) &\geq 0 \\ \int_{-\infty}^{\infty} f(x) dx &= 1 \end{aligned}$$

$f(x) > 1$ הינה פונקציית צפיפות, לא התפלגות. כלומר, "יתכן כי"

3.1.1 שונות ותוחלת

כאשר נתעסק במשתנים מקרים, לעיתים קרובות נדבר על שונות ותוחלת של משתנים מקרים.

הגדרה
התוחלת של המשתנה המקרי X היא:

$$\mathbb{E}[X] := \sum_{x \in \mathcal{X}} x \mathcal{D}(x) \quad \text{or} \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx$$

טענה
יהיו X, Y משתנים מקרים בעלי תוחלת סופית $\mathbb{E}[X], \mathbb{E}[Y]$
אזי מתקיימות התכונות הבאות:

$$\square \text{ ליניאריות התוחלת: } \mathbb{E}[aX + Y] = a\mathbb{E}[X] + \mathbb{E}[Y]$$

$$\square \text{ חוק הסטטיסטיائي חסר ההכרה: } \mathbb{E}[g(x)] = \sum g(x) \mathcal{D}(x)$$

□ אם X ו- Y בלתי תלויים אז $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

□ מונטוניות התוחלת: אם $X \leq Y$ אז $\mathbb{E}[X] \leq \mathbb{E}[Y]$

הגדלה

יהי X משתנה מקרי בעל תוחלת סופית $\mathbb{E}[X]$ השונות של X תהיה:

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

статistica התקן של X מוגדרת על ידי $\sigma := \sqrt{\text{Var}(X)}$

טענה

יהיו X ו- Y משתנים מקרים בעלי שונות ושותפות סופיות. אז מתקאים:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

□ לכל $a, b \in \mathbb{R}$ מתקאים כי $\text{Var}(aX + b) = a^2\text{Var}(X)$

□ חיוביות השונות: מתקאים כי $\text{Var}(X) \geq 0$. כמו כן $\text{Var}(X) = 0$ אם ויחד X קבוע.

הגדלה

השותפות המשותפת של X ו- Y מוגדרת על ידי:

$$\text{Cov}(X, Y) := [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

כמו כן, היא מסומנת גם בטור $\sigma_{XY} = \text{Cov}(X, Y)$

טענה

$$\text{Var}(X + Y) = \text{Var}(Y) + 2\text{Cov}(X, Y) + \text{Var}(X)$$

טענה

יהיו X, Y משתנים מקרים בעלי שונות ושותפות סופיות. אז מתקימות התכונות הבאות:

$$\text{□ סימטריות: } \text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{□ לכל } a, b, c, d \text{ מתקאים כי } \text{Cov}(aX + b, cY + d) = a \cdot c \cdot \text{Cov}(X, Y)$$

$$\text{□ } \text{Cov}(X, X) = \text{Var}(X)$$

למעשה, המשמעות של השונות היא המרחק מהתוחלת (או הגובה של בנאדם ביחס לאוכלוסייה) והשונות המשותפת היא כיצד המשתנים המקרים מותנהגים ביחד. כמו כן, נזכיר כי השונות המשותפת היא תבנית ביליניארית סימטרית, על כל המשתמע לכך.

דוגמה

יהי X משתנה מקרי ברנולי. אז השונות של X היא $\text{Var}[X] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = p - p^2 = p(1-p)$.

תרגיל

יהי $X \sim \text{Unif}([a, b])$ לכל $a, b \in \mathbb{R}$. חשבו את השונות והתוחלת של X .

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{b-a} \int_a^b x dx = \frac{b+a}{2} \\ \text{מайдן: } \mathbb{E}[X^2] &= \frac{1}{b-a} \int_a^b x^2 dx = \frac{b^2+ab+a^2}{3} \end{aligned}$$

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{b+a}{2} - \frac{b^2+ab+a^2}{3} = \frac{(b-a)^2}{12}$$

3.2 קצט סטטיסטיקה - הערכת שונות ותוחלת

עד כה, על סמך פונקציית התפלגות נתונה, רצינו לבדוק הסtabירות מסוימת. בסטטיסטיקה נעבד הפוך, נקבל דוגמאות ונרצה לחשב או להעריך את ההתפלגות اللا ידועה (למשל, רוצים לחשב כמה פיציות אורחים יאכלו, ואנו מוצאים דוגמאות מסוימות על האכילה של האורחים).

נרצה לחשב את השונות והתוחלת של ההתפלגות اللا ידועה.

נגדיר $(X_1, \dots, X_m) \sim \mathcal{D}(X)$ משתנים מקרים שווים ההתפלגות ובתי תלוים.

נתונות לנו x_1, \dots, x_m דוגמאות, ככל דוגמה x_i מתאימה למשתנה המקרי X_i .

נעיר כי בסטטיסטיקה נשתמש במונח "אוכולוסיית המדגם" ונרצה לחשב את התוחלת או השונות של האוכולוסייה (באה לתאר כי מדובר על תוחלת של התפלגות לא ידועה).

ישנו מספר דרכים לחשב זאת, וכל אחת מהן מוגדרות בתור **אומדן**. בהמשך נשתמש באומדנים נוספים.

על מנת לחשב את האומדן **لتוחלת** נשתמש בהגדרה $\hat{\mu}_X = \frac{1}{m} \sum_{i=1}^m x_i$ - ממוצע המדגים ועל מנת למצוא את

$$\text{האומדן לשונות} \text{ נשימוש ב-} \hat{\sigma}_X^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \hat{\mu}_X)^2 \text{ (בהמשך נסביר מדוע מחלקים בשונות ב-} m-1 \text{ ולא } m\text{).}$$

תרגיל

יהי X משתנה מקרי. הראו כי התוחלת של ממוצע המדגם ושל שונות המדגם, שווה לתוחלת ולשונות בהתאם.

$$\mathbb{E}(\hat{\mu}_X) = \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m x_i\right) = \frac{1}{m} \sum_{i=1}^m \mathbb{E}(x_i) = \mathbb{E}(X) \frac{1}{m} \sum_{i=1}^m 1 = \mathbb{E}(X) = \mu_X$$

נסמן $V = \text{Var}(X)$ ונקבל בנוסח:

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2 &= \frac{1}{n-1} \cdot \sum_{i=1}^n \mathbb{E} [((X_i - E) + (E - \mu))^2] = \\ &= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E} [(X_i - E)^2 + 2(X_i - E)(E - \mu) + (E - \mu)^2] = \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(V + \mathbb{E}(2(X_i - E)(E - \mu)) + \frac{V}{n} \right) \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(V + \mathbb{E}(2(X_i E - X_i \mu - E^2 + E\mu)) + \frac{V}{n} \right) \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(V + (2(E^2 - \mathbb{E}(X_i \mu) - E^2 + E^2)) + \frac{V}{n} \right) = \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(V + (2(E^2 - \mathbb{E}(X_i \mu))) + \frac{V}{n} \right) = \\ &= \frac{1}{n-1} \left(nV + 2 \left(nE^2 - \sum_{i=1}^n \mathbb{E}(X_i \mu) \right) + V \right) = \end{aligned}$$

נחשב את $\mathbb{E}(X_i \mu) = \mathbb{E} \left(\sum_{j=1}^n X_j \frac{1}{n} \sum_{i=1}^n X_i \right)$ ונקבל:

$$\frac{1}{n} \mathbb{E} \left(\sum_{j=1}^n \sum_{i=1}^n X_i X_j \right) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(X_i X_j)$$

אם איז נקבע פשוט $E(X_i^2) = V + E^2$ ויש n כאלה. אם איז התוחלת הינה (בגלל האי תלות):

$$E(X_i X_j) = \mathbb{E}(X_i) \mathbb{E}(X_j) = E^2$$

יש $(n(n-1))$ כאלה. סך הכל נקבע כי:

$$\begin{aligned}
& \frac{1}{n-1} \left(nV + 2 \left(nE^2 - \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i \mu) \right) + V \right) = \\
& \frac{1}{n-1} (nV + 2(nE^2 - (V + E^2) - (n-1)E^2) + V) = \\
& \frac{1}{n-1} \left(nV + \frac{2}{n} (-nV) + V \right) = \\
& \frac{1}{n-1} (V(n-2V+V)) = \\
& \frac{1}{n-1} V(n-1) = V
\end{aligned}$$

נחלף את V ב- σ_X^2 וסיימנו.

הגדלה
 ה- $\hat{\theta}$ אומדן של θ .
 ההטיה של $\hat{\theta}$ מוגדרת להיות:

$$B(\hat{\theta}) := \mathbb{E}[\hat{\theta}] - \theta$$

$$\text{נאמר כי } \hat{\theta} \text{ הוא איינו מוטה, אם } 0.$$

אם כך, ממה שראינו קודם לכן, ממוצע המדגמים הוא אומדן בלתי מוטה של ממוצע האוכלוסייה וכך גם השונות.
 (אם היינו משתמשים ב- m ולא $m-1$ – m היינו מקבלים אומדן מוטה).

3.3 כמה משתנים

עד כה התעסקנו בעיקר במשתנים מקריםים שבבאים תוצאה ייחודית, אך לעיתים רבות נרצה מספר רב יותר של משתנים, ולחשב את ההתפלגות המרובה שלהם.

הגדלה
 וקטור של משתנה מקרי.
 $X := (X_1, \dots, X_d)^T$ הוא אוסף סופי של משתנים מקרים המסומנים X_d, \dots, X_1 , ומשמעותו על ידי מרחיב הסתברות משותף (Ω, \mathcal{D}) .

הגדלה
 בהינתן משתנים מקרים X_d, \dots, X_1 , ההתפלגות המשותפת הינה ההתפלגות שככל אחד מהמשתנים המקרים י'פול' בתחום הרצוי עבור משתנה מקרי רציף ועבור ערכים ספציפיים במשתנה מקרי בדיד.

הגדלה

יהו X_1, X_2 שני משתנים מקרים. נאמר כי X_1, X_2 מתפלגים באופן משותף, אם קיימת פונקציה אי שלילית $f_{X_1, X_2} : \mathbb{R}^2 \rightarrow \mathbb{R}$ כך שלכל $A \in \mathbb{R}^2$:

$$\mathcal{D}((X, Y) \in A) = \int_A f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

מוגדרת בתור פונקציה הצפיפות המשותפת של X_1, X_2 .

משמעותו לב כי אם המשתנים המקרים אינם בלתי תלויים, יתכן כי יהיה הבדל משמעותי בין פונקציית ההתפלגות המשותפת והפונקציה הבודדת (למשל, בדקו עבור $([-a, a])$ $X_1, X_2 \sim -X_1$ ובדקו את ההסתברות למצוא את (X_1, X_2) בראיבוע $[0, 1] \times [0, 1]$ לעומת (X_1, X_1) באותו ריבוע).

3.3.1 ההתפלגות נורמלית

הגדרה

משתנה מקרי X הוא בעל ההתפלגות נורמלית עם תוחלת μ ושונות σ^2 אם יש לו פונקציית צפיפות מהצורה

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

 במקרה זה נסמן כי $x \sim \mathcal{N}(\mu, \sigma^2)$

הגדרה

וקטור $x \in \mathbb{R}^d$ הוא בעל ההתפלגות נורמלית מרובת משתנים עם תוחלת μ ומטריצת שונות משותפת (הגדרה בהמשך)
 \sum אם יש פונקציית צפיפות משותפת מהצורה:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}$$

במקרה זה נסמן $x \sim \mathcal{N}(\mu, \Sigma)$

נוכל להוכיח כי החקלה האחידונה מכיליה את ההגדרה הקודמת, כאשר $d = 1$, ומכאן נוכל להוכיח מהיכן מגיע
 הביטוי \sum_{σ}^{-1} (ככיבול כאנלוגיה ל- $\frac{1}{\sigma^2}$) - השונות המשותפת הינה אנלוגיה לשונות במקרה של משתנה אחד

לפעמים, נרצה לראות רק כיצד אחד המשתנים פועל - קלומר 'לבודד' את אחד המשתנים ולא לראות את ההתפלגות המשותפת.

הגדרה

התפלגות השולית של אוסף של משתנים מקרים עם ההתפלגות משותפת, הוא ההתפלגות של כל המשתנים בקבוצה:

$$f(x) = \int_y f(x, y) dy$$

כאשר y היא אינטגרציה על כלל המשתנים שלא נמצאים בקבוצה.

(הכללה של המקרה שראינו בהסתברות עבור התפלגות שולית של 2 משתנים, כאשר 'מקבילים' אחד וmbצעים אינטגרציה על השני).

דוגמה

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \quad x \in \mathbb{R}^2, \mu = (\mu_1, \mu_2)^T$$

יביאו את פונקציית הצפיפות של ההתפלגות השולית של x_1 .

הוכחה

קודם כל, נבחן כי אנחנו יכולים לכתוב את פונקציית הצפיפות בתור:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{(2\pi)^2 |\Sigma|}} \exp \left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \\ &= \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2}} \exp \left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \\ &= \frac{1}{\sqrt{(2\pi)\sigma_1^2}} \exp \left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right) \cdot \frac{1}{\sqrt{(2\pi)\sigma_2^2}} \exp \left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \end{aligned}$$

כעת, נשימוש בהגדרת הצפיפות המשותפת:

$$\begin{aligned} f(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2}} \exp \left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right) \cdot \frac{1}{\sqrt{(2\pi)^2 \sigma_2^2}} \exp \left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) dx_2 \\ &= \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2}} \exp \left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right) \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^2 \sigma_2^2}} \exp \left(-\frac{1}{2} \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) dx_2 \end{aligned}$$

נבחן כי אכן ימין סוכם את כל ההסתברויות' של צפיפות של המשתנה הנורמלי, ולכן שווה ל-1. לכן נקבל:

$$f(x_1) = \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2}} \exp \left(-\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 \right)$$

כלומר, קיבלנו את הצפיפות של המשתנה הנורמלי מהצורה $.x_1 \sim \mathcal{N}(\mu, \sigma_1^2)$

הגדרה

יהי $X = (X_1, X_2, \dots, X_d)^T$ משתנה מקרי (מרובה משתנים).

מטריצת השינויות המשותפות Σ הינה $d \times d$ כאשר עבור (i, j) כלשהו מתקיים כי $\Sigma_{i,j} = \sigma(X_i, X_j)$ (כל כניסה מסמלת שונות משותפת בין שני משתנים מקרים):

$$\Sigma := \begin{pmatrix} \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_1 - \mathbb{E}[X_1])] & \dots & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_d - \mathbb{E}[X_d])] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_d - \mathbb{E}[X_d])(X_1 - \mathbb{E}[X_1])] & \dots & \mathbb{E}[(X_d - \mathbb{E}[X_d])(X_d - \mathbb{E}[X_d])] \end{pmatrix}$$

האלכסונים על המטריצה הינם $\sigma_{X_i}^2 = \text{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2]$.

אם נזכור לסטטיסטיקה, נרצה לחשב את 'מטריצת השוניות המשותפת', בהינתן דוגמאות מסוימות.

במקרה שלנו, Σ היא מטריצת השוניות המשותפת של האוכלוסייה.

נתחיל בלחישוב עבור 2 משתנים (גובה ומשקל).

אם ניקח m אנשים ונמצא אותם בשני המשתנים האלו, נקבל שתי עמודות של דוגמאות, עבור הגובה והרוחב בהתאם:

$$X = \begin{bmatrix} x_{1,1} & x_{2,1} \\ \vdots & \vdots \\ x_{m,1} & x_{m,2} \end{bmatrix} = (x_1, \dots, x_m)^\top$$

נניח כי העמודה הראשונה מסמלת את הגובה של האנשים, והעמודה השנייה מסמנת את המשקל.

הגדרה

האומדן הבלתי מוטה של השוניות המשותפת של המדגם של המשתנים המקרים i ו- j הינו:

$$\hat{\sigma}(X_i, X_j) = \frac{1}{m-1} \sum_{k=1}^m (x_{k,i} - \hat{\mu}_i)(x_{k,j} - \hat{\mu}_j)$$

כאשר $\hat{\mu}_i$ הוא ממוצע המדגם של משתנה מקרי X_i .

3.3.2 מטריצת השוניות המשותפת

הגדרה

יהי $X = (X_1, \dots, X_d)^\top$ משתנה מקרי d מימי. וניקח $x_1, \dots, x_m \in \mathbb{R}^d$ דוגמאות של כל אחד מהמשתנים המקרים בהתאם.

מטריצת השוניות המשותפת של המדגם היא מטריצה d -ריבועית כך שבכינסה i, j מתקיים כי $\hat{\Sigma}_{i,j} = \hat{\sigma}(X_i, X_j)$. מבחן כתיבה מטרצונית, ניתן לתאר זאת כך:

$$\hat{\Sigma} := \frac{1}{m-1} \sum_{i=1}^m (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^\top = \frac{1}{m-1} \tilde{X}^\top \tilde{X}$$

כאשר \tilde{X} הוא המרכז של X (הפחיתה התוחלת).

(כמו שראינו על מטריצת שוניות משותפת, רק שהפעם החלפנו בשינוי משותפת של המדגם, כפי שתיארנו לפני רגע).

דוגמה

$$X \text{ דגימות של משקל וגובה של } 3 \text{ אנשים, נחשב את מטריצת השוניות המשותפת של המדגם.}$$

$$\text{ניקח} \quad \left(\begin{array}{c|c} 150 & 45 \\ 170 & 74 \\ 184 & 79 \end{array} \right)$$

תחילה, נפחית את הממוצע של הדגימות - נמוך את המידע.
נבחין כי ממוצע המדגם הינו $\hat{\mu} = (168, 66)^T$. וכך נקבל:

$$X_{\text{centered}} = X - \left(\begin{array}{cc} 168 & 66 \\ 168 & 66 \\ 168 & 66 \end{array} \right) = \left(\begin{array}{cc} 150 & 45 \\ 170 & 74 \\ 184 & 79 \end{array} \right) - \left(\begin{array}{cc} 168 & 66 \\ 168 & 66 \\ 168 & 66 \end{array} \right) = \left(\begin{array}{cc} -18 & -21 \\ 2 & 8 \\ 16 & 13 \end{array} \right)$$

ובצע כפל מטריצות ונקבל:

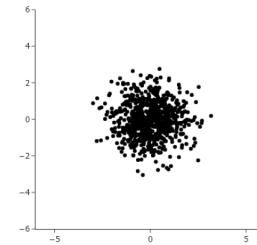
$$\hat{\Sigma} = \frac{1}{3-1} \tilde{X}^T \tilde{X} = \frac{1}{2} \cdot \left(\begin{array}{ccc} -18 & 2 & 16 \\ -21 & 8 & 13 \end{array} \right) \cdot \left(\begin{array}{cc} -18 & -21 \\ 2 & 8 \\ 16 & 13 \end{array} \right) = \left(\begin{array}{cc} 292 & 301 \\ 301 & 337 \end{array} \right)$$

3.3.3 העתקות ליניאריות על קבועות מידע

ונסה לראות כיצד העתקות ליניאריות פועלות על מאגרי מידע.
נזכיר כי העתקה ליניארית יכולה להיות מתיחה או סיבוב (גם שיקוף הוא סוג של סיבוב). נתבונן במקרה של $d = 2$ אבל התוצאה דומה גם במספר משתנים.
מטריצת השוניות המשותפת במקרה הדו ממדי הינה:

$$\Sigma = \begin{pmatrix} \sigma(X_1, X_1) & \sigma(X_1, X_2) \\ \sigma(X_2, X_1) & \sigma(X_2, X_2) \end{pmatrix}$$

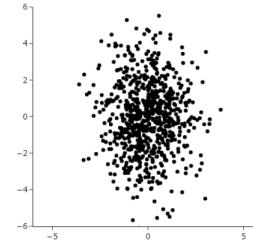
לשם הדוגמה נבחר דגימות שנלקחות בצורה בלתי תליה ושווות התפלגות עם ממוצע 0 ושוניות שווה.
דבר זה גורר באופן ישיר כי המשתנים המקרים X_1 ו- X_2 הם **בלתי מתואמים** ולכן נסמן את השוניות של כל אחד מהמשתנים המקרים ב- σ^2 נקבל מטריצה מהצורה $\sigma^2 I_2$.
כיוון שההתוללת אפס, במקרה זה **המוכז** איןנו עושה דבר ולכן $\hat{\Sigma} = \Sigma = \sigma^2 I_2$.
 מבחינה גרפית, הדגימה של המשתנים המקרים נראית כך:



כעת נראה כיצד העתקות ליניאריות משפיעות על המדגם שלנו. קודם כל, נכפול במטריצה המותחת $S = \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix}$. כיוון שהמטריצה אלכסונית, המתיחה של ציר ה- x' לא משפיעה על המתיחה של ציר ה- y' . וכך נקבל (נתיחס ל- $X = \tilde{X}$):

$$\hat{\Sigma}_{\text{scaled}} = \frac{1}{m-1} S X (S X)^T = S \left(\frac{1}{m-1} X X^T \right) S^T = \begin{pmatrix} (s_1 \hat{\sigma})^2 & 0 \\ 0 & (s_2 \hat{\sigma})^2 \end{pmatrix}$$

בציור, המתיחה נראה כך:



לבסוף, נפעיל 'סיבוב' על המידע. נזכיר כי מטריצה אורתוגונלית היא למעשה מטריצת סיבוב, ובפרט סיבוב של זווית θ ב- \mathbb{R}^2 נתון על ידי:

$$R = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

נחשב את מטריצת השינויות המשותפת, לאחר מתיחה וסיבוב⁷:

$$\begin{aligned} \hat{\Sigma}_{\text{rotated}} &= \frac{1}{m-1} (R S X) (R S X)^T \stackrel{\text{הגדרת טרנספוי}}{=} R S \left(\frac{1}{m-1} X X^T \right) (R S)^T R (S \hat{\Sigma} S^T) R^T \\ &= R \begin{bmatrix} (s_1 \sigma)^2 & 0 \\ 0 & (s_2 \sigma)^2 \end{bmatrix} R^T = \sigma^2 \begin{bmatrix} s_1^2 \cos^2 \theta + s_2^2 \sin^2 \theta & \sin \theta \cos \theta (s_1^2 - s_2^2) \\ \sin \theta \cos \theta (s_1^2 - s_2^2) & s_1^2 \sin^2 \theta + s_2^2 \cos^2 \theta \end{bmatrix} \end{aligned}$$

⁷ המעבר השני לא ברור. צריך לחזק

כיוון שמחוץ לאלכסון איברי המטריצה אינם הפיכים, בהכרח שני המשתנים מותואמים (כלומר, בצורה הופכית למה שרינו קודם, שם המשתנים המקרים היו בלתי מותואמים).
כמו כן, אם הינו בוחרים $s_1 = s_2$ אז הערכים מחוץ לאלכסון יהיו שווים לאפס. משמעו, סיבוב בלבד לא ייעיל⁸ מספק כדי לקבוע את התאימות\אי התאימות בין משתנים מקרים.

3.4 או שוויוניות

במהלך 'למידת מכונה', ישנו שימושים רבים לא"ש של הסתברויות, למשל, לחישוב חסמים על 'שגיאות' במהלך למידה של אלגוריתם, או לחישוב הביצועים של אלגוריתם בהינתן מרחב מדגם כלשהו. (למשל, החוק החלש של המספרים הגדולים אומר לנו כי אם נעשה הרבה ניסויים, יהיה קרוביים מאוד לתוחלת).

דוגמאות

נניח ויש לנו שק המכיל m כדורים צהובים ואדומים.
נרצה לחשב כמה כדורים (בשבירם) אדומים נשארו בסל, כאשר בכל פעם נוציא כדור אחד רנדומלית ונחזיר אותו. כיצד נוכל לעשות זאת? נגדיר x משתנה מקרי אינדיקטור, שווה 1 אם הכדור שהזינו אדום, ו 0 אחרת.
נדגום m משתנים מקרים כאלו ונוכל באמצעות ממוצע המדגם $\hat{p} = \frac{1}{m} \sum_{i=1}^m x_i$ לקבל את הערכה הבאה:

$$\hat{p} = \frac{1}{m} \sum_{i=1}^m x_i = \frac{\text{balls red of Number}}{m}$$

אמנם, נבחן כי עשינו רק m ניסויים ולכן יש לנו מידע מוגבל - יתכן שבמהלך m הדגימות הינו חסרי מצל וקיבלו תוצאה שאינה מייצגת.
דבר זה יכול לקרות גם עבור m מאד גדול, ולכן 'מייצר' משתנה $0 < \epsilon$ שמייצג את 'מרחב השגיאה' מהערך האמתי.
בכל אופן, גם עבור m מאד גדולים, לא נוכל להבטיח כי המדגם יוצר לנו את התוצאה המדויקת.
ברור כי m הדגימות הן סופיות, ולכן התוצאה עלולות לא תהיה מדויקת לחולוין. אך אכן נוכל להשיג **תוצאה מדויקת מספק**. ככלומר, בהינתן $0 > \delta$ קטן מאד, נרצה בסיכוי של δ – $1 - \hat{p}$ יהיה מספק מדויק (כלומר, במרחב ϵ הרצוי מהיעד δ) – ונרצה לבדוק איזה m ישפוק לנו את הסיכוי הזה.
בניסוח פורמלי נוכל לומר: בהינתן 'מקדם שגיאה' $\epsilon > 0$ ומקדם ביטחון $(0, 1) \in \delta$, כמה דגימות נctrarך לחתול על מנת להבטיח כי בסיכוי של δ – $1 - \epsilon$ אנחנו במרחב מקסימום ϵ מהערך הנכון.
בהמשך, נדון בכך כshedbar על PAC – "למידת קרובה לוודאי בערך נכון".

3.4.1 א"ש מركוב וצ'יבישב

נזכיר תחילת באי השוויונות שלמדנו בעבר, וניחסם אותם בהקשר של הניסוי שהזכרנו לעיל.

טענה - א"ש מרכוב

יהי X משתנה מקרי אי שלילי. נסמן את התוחלת ב- $\mathbb{E}[X]$. אזי מתקיים, לכל $a > 0$:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

⁸מהו האינטואיציה לכך? למה סיבוב לא מספק כדי לקבוע תאימות? מה שונה ממשר מתיחה?

הוכחה

תהי $f(x)$ פונקציית הצפיפות של X . כיוון ש- X אי שלילי אז $f(x) = 0$ לכל $x < 0$. ואז נקבל:

$$\frac{\mathbb{E}[X]}{a} = \frac{1}{a} \int_0^\infty f(x)xdx \stackrel{\text{בחירה חלק מהאיברים}}{\geq} \frac{1}{a} \int_{x=a}^\infty f(x)xdx \geq \frac{1}{a} \int_{x=a}^\infty f(x)adx = \mathbb{P}(X \geq a)$$

מסקנה

יהיו X_1, \dots, X_n משתנים מקרים אי שליליים שווים התפלגותם ובת"ל. נסמן את התוחלות $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i = \mathbb{E}[X_i]$ לכל $1 \leq i \leq n$. נסמן בנוסח a מתקיים, לכל $a > 0$:

$$\mathbb{P}[\bar{X} \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

הוכחה

ונכל להפעיל את מרכיב על \bar{X} כיוון שמדובר במשתנה מקרי אי שלילי. אז נקבל:

$$\mathbb{P}[\bar{X} \geq a] \leq \frac{\mathbb{E}[\bar{X}]}{a}$$

אבל כמו כן, מתקיים:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[X_i] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[X] = \mathbb{E}[X]$$

ולכן קיבלנו בסך הכל כי:

$$\mathbb{P}(\bar{X} \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

כנדרש.

ונכל גם להשתמש במרכיב על מנת להגיד חסם התלו依 בשונות ולא בתוחלת.

טענה - א"ש צ'בישב

לכל משתנה מקרי סופי X עם ממוצע (או תוחלת) סופיים $\mathbb{E}[X]$ ושותות סופית $\text{Var}(X)$ ולכל $a > 0$ יתקיים:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

הוכחה

נגידיר את המשתנה המקרי, $Y = (X - \mathbb{E}[X])^2$. זהו משתנה מקרי אי שלילי ולכן נוכל להפעיל את מרכוב ונקבל:

$$\mathbb{P}[(X - \mathbb{E}[X])^2 \geq a^2] \leq \frac{\text{Var}(X)}{a^2}$$

על מנת לסייע את ההוכחה, נבחן כי $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] = \mathbb{P}[(X - \mathbb{E}[X])^2 \geq a^2]$

בשונה ממוצע המדגם, נבחן כי השונות של סכום m המשתנים המקרים שווה:

$$V\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m^2} V\left[\sum_{i=1}^m X_i\right] = \frac{1}{m^2} \sum_{i=1}^m \text{Var}(X_i) = \frac{1}{m^2} \sum_{i=1}^m \text{Var}(X) = \frac{1}{m} \text{Var}(X)$$

מסקנה

בහינתן X_1, \dots, X_n משתנים מקרים שווים התפלגות ובת"ל עם שונות סופית $\text{Var}(X)$.

נגידיר $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$. לכל $a > 0$ מתקיים:

$$\mathbb{P}[|\bar{X} - \mathbb{E}[\bar{X}]| \geq a] = \mathbb{P}[\bar{X} - \mathbb{E}[\bar{X}] \geq a] \leq \frac{\text{Var}(\bar{X})}{m \cdot a^2}$$

(מציר מאד את החוק החלש של המספרים הגדולים)

הגדרה

לכל מספר טבעי k , המומנט ה- k של המשתנה מקרי X מוגדר על ידי $\mathbb{E}[X^k]$.

כפי שראינו, א"ש מركוב מספק מידע על המומנט הראשון בלבד, ואילו צ'יבש משתמש הן במומנט הראשון והן בשני. בהמשך, נושא בין החסמים ונראה כי השימוש בצ'יבש ומילא בשני המומנטים מביא לנו חסם מדויק יותר מאשר שימוש במומנט הראשון בלבד.

בקרוב נאמר כי ככל שנשתמש ביותר מומנטים (אם קיימים), אז נקבל חסמים מדויקים יותר.

3.4.2 דוגמת חיזוי מטבע

נשותמש בדוגמה להערכת הטיפה של מטבע, או בקצרה 'חיזוי מטבע'. באמצעות דוגמה זו נוכל להבין את המושג החשוב **'סיבוכיות המציג'**, להדגים את השימושות וההגבלות של א"ש שהארכנו לעיל, ולהכיר א"ש חדש בשם א"ש הופדינג.

פורמלית, הטלת מטבע היא משתנה מקרי ברנולי Z שמקבל ערך 1 להטלת H ו-0 להטלת T .
 נסמן את ההסתגלות של Z באמצעות \mathcal{D}_p כאשר $\mathcal{D}_p(0) = 1 - p$ ו- $\mathcal{D}_p(1) = p$.
 תחילה, ניקח מטבע הוגן, כלומר $Z_1, \dots, Z_m \stackrel{\text{iid}}{\sim} \text{Ber}(p)$.
 נסמן את התוצאות של הטלות בהתאם, באמצעות $S = (z_1, \dots, z_m)$, ואת הסיכוי לקבלת התוצאה הספציפית באמצעות $\mathcal{D}_p^m(S)$.

שיטת לימוד האלגוריתם, 'חיזוי מטבע' היא שיטה שמקבלת בתור קלט, סדרה S , בהתאם ל- \mathcal{D}_p^m , ומיצאת בתור פלט הערכה של p .

שיטת זו, שמכונה גם בתור חיזוי, מסומנת בתור $\hat{p}(S) = \mathcal{A}(S)$ או \hat{p} .
 כיוון ש- S הינו סופי, אנחנו לא מצפים כי הערכה \hat{p} תהיה מדויקת לחולוין.
 לעומת זאת, נגידר אלגוריתם שנקלע עבורה - $\hat{p} = p$ עבור $\epsilon < 0$, שייקרא בתור **מקדם הדיק**.
 לעומת זאת, כפי שראינו קודם, תמיד יש סיכוי ('א'כ' p שווה ל-1 או 0) שהסדרה שהתקבלת אינה מייצגת. בעקבות כך, נגידר **מקדם ביטחון** ($\delta \in (0, 1)$), ונרצה כי המאורע $\{|\hat{p} - p| > \epsilon\} \subset \{S : \hat{p} \in (\hat{p} - \delta, \hat{p} + \delta)\}$ יתקיים בסיכוי מקסימום של δ אם הינו מדברים במונחים של שאיפה לאינסוף, נרצה כי הסיכוי של הדבר הזה שואף ל-0). בambilים אחרים,
 נדרש כי הסיכוי שהאלגוריתם שלנו הציג תוצאה לא מדויקת בסדרה של m הטלות (כלומר $\epsilon > |\hat{p} - p|$), תהיה קטנה או שווה ל- δ .
 אינטואיטיבית, ככל שנעשה יותר ניסויים, נקבל יותר מידע ונוכל להבטיח את הדיק והבטיחון הנדרשים. כיוון שמקדמי הביטחון והדיק הם **קבועים**, קיים בהכרח מספר סופי של הטלות $m_{\mathcal{A}}$ (שתלי ב- δ, ϵ ו- \mathcal{A}), כך שלכל מוגדל $m \geq m_{\mathcal{A}}$, האלגוריתם שלנו יתאים לדרישות מקדמי הביטחון והדיק.
 אם קיים $m_{\mathcal{A}}$ כזה, נאמר שהאלגוריתם שלנו הוא 'אלגוריתם לומד'⁹, למשימת חיזוי המטבע.
 מתמטית, אנחנו מוחפשים אלגוריתם שיתאים להגדלה הבאה.

הגדרה

יהי \mathcal{A} אלגוריתם, שיסומן נס בתור $\hat{p}(S) = \mathcal{A}(S)$, שמיירה $\hat{p} \in [0, 1]$, שמקבל סדרה של דוגמאות הטלת מטבע $(S \in (0, 1))$, ומקיימת את התנאים הבאים:

□ לכל $\delta \in (0, 1)$ קיים מספר אי שלילי $m_{\mathcal{A}}(\epsilon, \delta)$ כך שאם סדרה S של m מספרים, כאשר \mathcal{A} נוצרת בהתאם ל- \mathcal{D}_p^m , אז לכל $0 \leq p \leq 1$ מתקיים:

$$\mathcal{D}_p^m[|\hat{p}(S) - p| > \epsilon] \leq \delta$$

כלומר, הסיכוי לקבל מוגם S כך שהפלט של האלגוריתם $\hat{p}(S)$ לא יהיה בקטע $[p \pm \epsilon]$ קטן או שווה ל- δ .

□ אם ניצור סדרה S של m מספרים, כאשר $m < m_{\mathcal{A}}$, קיים $\hat{p} \in [0, 1]$ כך ש:

⁹'learning algorithm'

$$\mathcal{D}_p^m[|\hat{p}(S) - p| > \varepsilon] > \delta$$

הפונקציה $\mathbb{N} : (\varepsilon, \delta) \rightarrow [0, 1] \times [0, 1]$ מכונה **סיבוכיות המדגם** של האלגוריתם.

מהתנאי הראשון עולה כי בלי קשר לערך האמתי של p , מספיק ליצר $m_{\mathcal{A}}$ דוגמאות, בשביל לדעת מהו p עם וודאות של $\delta = 1 - \text{ודיק ש } \pm \varepsilon$.

התנאי השני אומר כי לפחות חלק מערכיו p , יצירת $1 - (\varepsilon, \delta)$ ¹⁰ לא תהיה מספקה. נבחין כי על מנת שדבר זה יפעיל, כמוות הדגימות איננה יכולה לצריכה להיות תליה בהסתברות האמיתית p .

בחירה אלגוריתם

נבחר אלגוריתם שיתן לנו את ההגדרה למעלה. בהינתן מוגם $S = (z_1, \dots, z_m)$, הערכה הכי ישירה של p היא $\hat{p}(S) = \frac{1}{m} \sum_{i=1}^m z_i$ - תספר את ה-1-ים ותחלק במספר הרוחות. נראה כיצד האלגוריתם מקיים את ההגדרה. תחילה נבחן כי האומד הזה הוא אומד לא מותה (כפי שראינו כבר). כזאת, הוא יכול לקבל ¹¹ עבור S מותאים כי $0 = |\hat{p} - p|$.

לאחר מכן נרצה לחשב כמה הטלות מטבע נדרש על מנת לוודא כי \hat{p} הוא קרוב ל- p הדורש. על מנת לענות על שאלת זו, נשתמש בא"ש שראינו.

הערכת סיבוכיות המדגם באמצעות א"ש מركוב

ניקח $|p - \hat{p}|$ בתור משתנה מקרי. נרצה קודם כל לחשב את התוחלת של $|\hat{p} - p|$. נקבל ¹²:

$$\mathcal{D}_p^m[|\hat{p} - p| \geq \varepsilon] \leq \frac{1}{\sqrt{4m\varepsilon^2}}$$

כלומר, אם נבחר $\left\lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta^2} \right\rceil$, אז אגן ימין שווה ל- δ . ולכן, לכל $(\varepsilon, \delta) \in (0, 1)$, אם נדגים $m_{\mathcal{A}}(\varepsilon, \delta) \geq \left\lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta^2} \right\rceil$, נקבל כי האלגוריתם הלומד מקבל כי:

$$\mathcal{D}_p^m[|\hat{p}(S) - p| > \varepsilon] > \delta$$

הערכת סיבוכיות המדגם באמצעות א"ש צ'בישב

על מנת לחשב את החסם העליון שראינו למעלה (כלומר, למצוא פונקציית סיבוכיות שדורשת פחות דוגמאות), נשתמש בא"ש צ'בישב.

נבחן כי השונות של משתנה ברנולי הינה $\leq \frac{1}{4}(p - 1)^2$, ולכן כאשר נפעיל את א"ש צ'בישב נקבל:

$$\mathcal{D}_p^m[|\hat{p} - p| \geq \varepsilon] = \mathcal{D}_p^m[|\hat{p} - \mathbb{E}[\hat{p}]| \geq \varepsilon] \leq \frac{p(1-p)}{m\varepsilon^2} \leq \frac{1}{4m\varepsilon^2}$$

¹⁰ מהי משמעות ה-1?

¹¹ למה?

¹² אל תהאלו שאלות, לא תשמעו שקרים. סתם, העונה על פיתרון החידה יזכה בפרס מוגנת טרקלין חשמל

כאשר המעבר האחרון נובע מכך ש- $\frac{1}{4}$.
כלומר, במקומות $\frac{1}{\sqrt{m}}$, קיבלנו חסם של $\frac{1}{m}$ - שהינו חסם טוב יותר.

מסקנה

$^{13}.m_{\mathcal{A}}(\varepsilon, \delta) \leq \left\lceil \frac{1}{4\varepsilon^2} \cdot \frac{1}{\delta^2} \right\rceil$ סיבוכיות המדגם של חיזוי מטבעות חסומה מלמעלה על ידי

הערכת סיבוכיות המדגם באמצעות א"ש הופding

אפשר לשאול האם החסם שקיבלנו מלמעלה הוא האופטימי. למעשה, נוכל לשפר את החסם, באמצעות הובדה שהמשתנה המקרי שלנו חסום בין 0 ו-1. לשם כך נוכל להשתמש בא"ש הופding.

טענה - א"ש הופding

יהיו X_1, \dots, X_m משתנים מקרים בלתי תלויים וחסומים על ידי
נסמן $\overline{X} = \frac{1}{m} \sum_{i=1}^m X_i$ וזו מותקית:

$$\mathbb{P}[|\overline{X} - \mathbb{E}[\overline{X}]| \geq \varepsilon] \leq 2 \exp \left(\frac{-2m^2\varepsilon^2}{\sum_{i=1}^m (b_i - a_i)^2} \right)$$

מסקנה

יהיו X_1, \dots, X_m סדרה של משתנים מקרים שווי התפלגות ובת"ל, שככל אחד תוחלת $[X]$ וכשכולם חסומים
.a $\leq X_i \leq b$
נסמן $\overline{X} = \frac{1}{m} \sum_{i=1}^m X_i$ ונקבל:

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \geq \varepsilon) \leq 2 \exp \left(\frac{-2m\varepsilon^2}{(b-a)^2} \right)$$

אם נפעיל את א"ש הופding על המקרה של חיזוי המטבעות, נקבל כי $(\widehat{p} - p) \geq \varepsilon$ $\leq 2 \exp(-2m\varepsilon^2)$, וכן
לראות כי קיבלנו חסם שמתכנס באופן אקספוננציאלי כתלות ב- m .
כלומר, אם ניקח $\left\lceil \frac{1}{2\varepsilon^2} \log \left(\frac{2}{\delta} \right) \right\rceil \geq m$ דגימות, נקבל כי ההסתברות חסומה על ידי δ , וכן נוכל להגיעה למסקנה
הבא.

מסקנה

אלגוריתם חיזוי המטבעות, (S, \widehat{p}) , שמעיריך את p לפי מספר ה- H -חלקי מסpur הטלות המטבע, מקיים את ההגדרה
 $^{14}.m_{\mathcal{A}}(\varepsilon, \delta) \leq \left\lceil \frac{1}{2\varepsilon^2} \log \left(\frac{2}{\delta} \right) \right\rceil$ דלעיל עם סיבוכיות מדגם שחסומה מלמעלה על ידי

¹³ההבדל בין זה ובין מרקוב - במרקוב יש δ^2 .
¹⁴אם שרדתם עד לפה, בראבו - אבל זו סך הכל חזרה.

חלק II

רגרסיה ליניארית

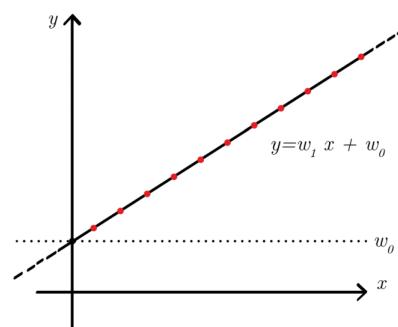
1 מודלים של רגרסיה

הבה וניקח דוגמה שתבהיר לנו היבט את הצורך והשימוש במודלי רגרסיה. נניח שיש לנו חנות מקוונת, ונרצה לחזות את ערך חי הלוקוח, ככלור ככמה רכישות יבצע הלוקוח באתר. על מנת להעריך זאת, נדרש לקחת כל מיני תוכנות רלוונטיות של הלוקוח (גיל, הכנסה, זמן כניסה לאתר, זמן ממוצע באתר וכו'), שיספקו לנו מידע על העתיד. המרחב הוקטוריו שמוגדר על סמך זאת, נקרא **תחום המדגם** (sample domain), שנסמנו ב- \mathcal{X} , ובמקרה שלנו הינו \mathbb{R}^d כאשר d הוא מספר התוכנות. את **קבוצת התגובה** (response set) נסמן ב- \mathcal{Y} , כשהיא מסמלת את ערך חי הלוקוח - במקרה שלנו \mathbb{R} .

כעת, נוכל לקחת דבר זה עבור m ל��חות, כשכל אחד הינו למעשה דגימה - זוג סדר (y, x) - **א הינו וקטור העמודה של תוכנות המדגם (נקודות נתונים)** ו- $x \in \mathcal{X}$ הוא התגובה המתאימה של הלוקוח. דבר זה נקרא **סט האימון** (training dataset) - נרצה להשתמש בו על מנת להעריך את ערך חי הלוקוח של לקוחות חדשים. דרך זו נקראת **Learning** ועל מנת להשתמש בה, יש לבנות **מודל רגרסיה** (regression model). **מודל רגרסיה** הוא הדרך לסלול קשר פונקציוני בין קבוצת התוכנות ב- \mathcal{X} וסקלר תגובה ב- \mathcal{Y} . נניח שקיים פונקציה $y \rightarrow \mathcal{X} : f$ כזו, שמקשרת בין הדגימות ב- \mathcal{X} והתשובות ב- \mathcal{Y} . פונקציה זו איננה ידועה לנו ונרצה לגנות אותה - היא יכולה להיות **דטרמיניסטית** או בעלת רכיב רנדומי.

תחליה, נניח כי הקשר בין $\mathcal{X} \in x$ ו- $\mathcal{Y} \in y$ הוא דטרמיניסטי. כמובן, נניח כי קיימת פונקציה $y \rightarrow \mathcal{X} : f$ כך שכל מודם שנסקרו, כתע או בעתיד, הוא מהצורה (y, x) כאשר $(x, y) = f(x)$, או בפרט $(x_i, y_i) = f(x_i)$, לכל $i \leq m$. דוגמאות האימון שלנו.

המטרה שלנו היא **ללמוד** (learn) את f מתוך מודם האימון $S = \{(x_i, y_i)\}_{i=1}^m$, על מנת שנוכל להעריך או לחזות את (x, y) עבור x חדש - מודם כאה נקרא **לעיטים מבחן** (test sample). באמצעות מודם האימון (training sample) נבנה פונקציה שנקווה שקרובה לפונקציה f והיא נקראת **כלל החיזוי** (prediction rule) ונסמנה ב- \hat{f} או h_S (כתלות במרחב המדגם S). דוגמה לפונקציה כזו (האדוות הן הדגימות), ניתן לראות כאן:



בשל סיבות שנרחיב עליה בהמשך, נצטמץ לקבוצת פונקציות שנקראת **מחלקה היפותזות**. נקבע אותה לפני שנتابון במידע, והפונקציה \hat{f} תהיה חייבת להיות באותה מחלוקת מדוברת. הדרך פשוטה ביותר לבנות מודל רגרסיה מעלה \mathcal{X} בלבד, היא מעלה \mathbb{R}^d , וכך יהיה בכל הפרק ובהמשך הקורס.

1.1 רגרסיה ליניארית

תחליה, נניח שהקשר $\mathcal{X} \rightarrow \mathcal{Y}$ הוא ליניארי.

הגדרה

המודל הליניארי, או מחלוקת הhipotезה הליניארית היא קבוצת הפונקציות הליניאריות¹⁵ מתחום למרחב התגובה:

$$\mathcal{H}_{\text{reg}} := \left\{ h(x_1, \dots, x_d) = w_0 + \sum_{i=1}^d x_i w_i \mid w_0, w_1, \dots, w_d \in \mathbb{R} \right\}$$

על מנת להפוך את ההגדרה לנוחה יותר, נרצה להסתכל עלייה כמו מכפלה פנימית (מכפלת וקטורים). 'בעה' היא ש- $x \in \mathbb{R}^d$ ו- $w \in \mathbb{R}^{d+1}$ הפונקציה הנו ב- $w^\top x$ (כי יש לנו את ה-intercept).
לכן נוסיף 1 ל'קוארדינטת ה-0' ונקבל למעשה $x = (1, x_1, \dots, x_d)^\top \in \mathbb{R}^{d+1}$ ואז נקבל $w^\top x$ ובקיצור:

$$\mathcal{H}_{\text{reg}} := \{h_w(x) = x^\top w \mid w \in \mathbb{R}^{d+1}\}$$

בקיצור, נחפש נספח w כך ש- $w^\top x_i = y_i$ לכל $i \in [m]$. לא תמיד זה יהיה פשוט כל כך.

נרצה לסדר את $S = \{(x_i, y_i)\}_{i=1}^m$ בתור מטריצה. נגדיר את וקטור התגובה $y \in \mathbb{R}^m$ ובתור:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

ואת מטריצת הרגרסיה $X \in \mathbb{R}^{m \times (d+1)}$ בתור:

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ \vdots & & \\ - & x_m & - \end{bmatrix}$$

השורות של X מסמלות את מספר דוגמאות האימון שלנו, ו- $d + 1$ מסמלות את התכונות (features) ואת ההזזה. אם כך, אנחנו מחפשים וקטור $w \in \mathbb{R}^{d+1}$ שקיים את הביטוי $y = Xw$.

¹⁵תכל'ס סתם קוראים לזה ליניארית, זה בכלל אפנית (משמעותם בקבוע).

חשיבות מספר הדגימות

בשלב זה נניח כי $d + 1 \geq m$, כלומר כי יש מספיק דוגמאות ולפחות ככמויות התוכנות. אחרת, לא ניתן לפתור את מערכת המשוואות.

2.1.2 עיצוב אלגוריתם למידה (Designing A Learning Algorithm)**2.1.2.1 רילזיביליות (Realizability)**

תחליה, נתבונן במקרה הרילזיבלי¹⁶. נזכיר כי אנחנו מחפשים פונקציה $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ כך ש- $f \in \mathcal{H}_{\text{reg}}$. במקרה בו יש פיתרון למשווה דלעיל נקרא המקרה הרילזיבלי. יהיו \hat{f} פיתרון עבור המשווה שראינו קודם לכן, אם כך, הפונקציה שנבחר תהיה $\hat{f}(x) = \hat{f}^T(x)$.

המקרה בו אין פיתרון למשווה זו נקרא במקרה הלא רילזיבלי - נדרש למצוא $\hat{f} \in \mathcal{H}_{\text{reg}}$ שהינה המתאימה ביותר לצרכינו.

אלגוריתם הלמידה שלנו צריך להתמודד עם שני המקרים הללו. במקרה הרילזיבלי פשוט, מה קורה במקרה בו אין פיתרון למשווה?

2.1.2.2 פונקציית הפסד (Loss Function)

את הדרכים לבחור את $\hat{f} \in \mathcal{H}_{\text{reg}}$ במקרה הלא רילזיבלי הוא להתאים לכל f מד איות, ולבחר את 'הטובה' ביותר. הפונקציה שמוגדרת על מנת למדוד את האיות מוגדרת כפונקציית הפסד (Loss Function), והיא מוגדרת כך:

$$\sum_{i=1}^m L(f(x_i), \hat{f}(x_i)), \quad i = 1, \dots, m,$$

נבחן את כלל החזוי על פי הפונקציה 'מתאימה ביותר'. אפשר לבדוק זאת באמצעות ערך מוחלט (Absolute Value). $L(y, \hat{f}(x)) := |y - \hat{f}(x)|$ או באמצעות 'ריבועי' (Squared Loss). $L(y, \hat{f}(x)) := (y - \hat{f}(x))^2$. אנו נתמקד ברגression הלייניארית שימושת בפונקציית הפסד הריבועי, כי זה יותר יouter.

2.1.2.3 מיזור הסיכון האמפירי (Empirical Risk Minimization)

נתרכז כעת ב- \hat{f} חדשה שבבסיסה על פונקציית הפסד הריבועי. נסמנה בתווך $\hat{f}(x) - y$ - ברור כי נרצה לבחור \hat{f} שמנזרת את פונקציית הפסד הריבועי. האסטרטגיה של בחירת \hat{f} מוגדרת בתווך 'מיזור הסיכון האמפירי' (Empirical Risk Minimization). בהינתן כלל חזוי $\hat{f} \in \mathcal{H}$, הנקודות $(y_i, \hat{f}(x_i))$ מוגדרת בתווך הסיכון האמפירי $\sum_{i=1}^m L(y_i, \hat{f}(x_i))$. במקרה של פונקציית הפסד הריבועי, הסיכון האמפירי של פונקציה w נתון בתווך:

$$\begin{aligned} & \underbrace{\sum_{i=1}^m (y_i - \hat{f}(x_i))^2}_{\text{מ''פ}} \\ & \downarrow \\ & \underbrace{\|y - \hat{f}(x)\|^2}_{\text{מ''פ}} \\ & \downarrow \\ & (y - \hat{f}(x))^\top (y - \hat{f}(x)) \end{aligned}$$

¹⁶אליעזר בן יהודה עדיין לא המציא מילה זאת.

1.2.4 שיטת הריבועים הפחותים (least squares)

נרצה למצוא את הפונקציה 'הכי קרובה' במנחים של 'מרחב שגיאה ריבועי'.
 הביטוי $\mathbf{w}^\top \mathbf{x}_i - y_i$ נקרא 'השארית' (residual) והסיכון האמפירי הכלול מוגדר בתור 'שארית סכום הריבועים' (Sum of Squares)

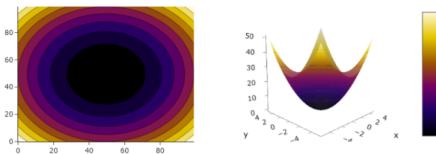
$$RSS_{\mathbf{X}, \mathbf{y}}(\mathbf{w}) := \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

לפעמים נתעלם מ- \mathbf{y} , \mathbf{X} בעת כתיבת הביטוי.
 אם כך, נרצה למצוא את הביטוי, כולם למצוא את:

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} RSS(\mathbf{w}) = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

חשוב להבחין כי **בעיות אופטימיזציה** מתייחסות הן ל מקרה הרלייזבלי והן ל מקרה הלא רלייזבלי:

- במקרה הרלייזבלי, כלומר כאשר $\mathbf{y} \in \text{Im}(\mathbf{X})$, אנו יודעים כי יש לפחות פתרון אחד $\hat{\mathbf{w}}$ כך $\mathbf{y} = \mathbf{X}\hat{\mathbf{w}}$. במקרה זה, הפונקציה מקבלת 0 עבור הפתרון, והוא אכן המזער של הפונקציה.
- במקרה הלא רלייזבלי, כאשר $\mathbf{y} \notin \text{Im}(\mathbf{X})$, אין פתרון כלל ולכן נצורך למצוא וקטור 'מספיק טוב'.



תנאי הכרחי ש- \mathbf{w} יהיה מזער הוא כי כל הנגזרות החלקיים 'געלות' ב- \mathbf{w} . כלומר:

$$\begin{aligned} \frac{\partial}{\partial w_j} RSS(\mathbf{w}) &= -2 \sum_{i=1}^m (\mathbf{x}_i)_j \cdot (y_i - \mathbf{x}_i \mathbf{w}) = 0 \Rightarrow \\ \nabla RSS(\mathbf{w}) &= -2 \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \end{aligned}$$

לכל $1 \leq j \leq n$.
 (עשינו נגזרות חלקיות של המכפלה הפנימית לפי כל אחד מ- w_j , כלומר אלו הם רכיבי ה- \mathbf{w}).

1.2.5 שוויונות נורמלים (The Normal Equations)

מהabitio לעיל עולה כי

$$\mathbf{x}^\top (\mathbf{y} - \mathbf{x}\mathbf{w}) = 0 \iff \mathbf{x}^\top \mathbf{y} = \mathbf{x}^\top \mathbf{x}\mathbf{w}$$

משמעות גיאומטרית

ננסה להבין דבר זה בצורה גיאומטרית שתמחיש לו היטב את הסיטואציה. אנו רגילים לחשב על $\mathbf{X} \in \mathbb{R}^{m \times (d+1)}$ בטור מטריצה שמכילה m שורות, לכל מרחב אימון. במקומות זאת, נוכל לחשב על \mathbf{X} בטור מטריצה עם $d + 1$ عمودות, כשלכל אחת פיצ'ר אחר:

$$\mathbf{X} := \begin{bmatrix} | & & | \\ \varphi_0 & \cdots & \varphi_d \\ | & & | \end{bmatrix}$$

התמונה למעשה נפרשת על פי העמודות של \mathbf{X} , ככלומר מתקיים: $\text{span}(\varphi_0, \dots, \varphi_d) = \text{Im}(\mathbf{X}) \subset \mathbb{R}^m$. נקבל כי התמונה של \mathbf{X} היא תת מרחב של \mathbb{R}^m . אם $m = d + 1$ אז התמונה ממש שווה ל- \mathbb{R}^m . אם $m = d + 1$ אז מדורבת רק בתת מרחב של \mathbb{R}^m .

כעת, ניקח את וקטור התגובה $\mathbb{R}^m \in \mathbf{y}$. נחלק למקרים:

□ אם $\mathbf{y} \in \text{Im}(\mathbf{X})$ אז \mathbf{y} הוא צירוף לינארי של $\varphi_d, \dots, \varphi_0$ וקיים $\mathbf{w} \in \mathbb{R}^m$ כך $\mathbf{y} = \mathbf{w}$. גם כאן יתכנו שתי אפשרויות:

- אם $\varphi_d, \dots, \varphi_0$ בלתי תלויים, אז \mathbf{y} יכול להיכתב בטור צירוף לינארי יחיד של \mathbf{X} . במקרה זה, למערכת שראינו מקודם יש פיתרון יחיד.

- אם $\varphi_d, \dots, \varphi_0$ תלויים, ישנו אינסוף דרכים להציג את \mathbf{y} בטור קומבינציה של עמודות \mathbf{X} .

□ $\mathbf{y} \notin \text{Im}(\mathbf{X})$ - במקרה זה \mathbf{y} לא מהו צירוף לינארי של עמודות \mathbf{X} - המקרה הלא ריליאנטי. נדרש לבחור ממזרע RSS.

מהי משמעות הביטוי 'נורמלי' ב'ביטויים נורמליים'? קודם לכן רأינו כי $\langle \mathbf{y} - \mathbf{x}\mathbf{w}, \varphi_j \rangle = 0$ לכל $j \leq d$, ולמעשה דבר זה אומר כי $\mathbf{x}\mathbf{w} - \mathbf{y}$ מאונך לכל אחד מ- φ_j או כי $\mathbf{x}\mathbf{w} - \mathbf{y} \in \text{Im}(\mathbf{X})^\perp$. כעת, ניקח $\hat{\mathbf{y}}$ שמהווה פיתרון לשווין הנורמלי שראינו קודם לכך, ונגיד $\hat{\mathbf{y}} = \mathbf{x}\mathbf{w}$.

אם נרצה למזער את RSS, נרצה למעשה למזער את $\|\hat{\mathbf{y}} - \mathbf{y}\|$. נגיד את וקטור השארית בטור $\hat{\mathbf{y}} - \mathbf{y} = \hat{\mathbf{z}}$ וכעת נבהיר כי בהכרח $\hat{\mathbf{z}} \in \text{Im}(\mathbf{X})^\perp$.

במילים אחרות, אם $\hat{\mathbf{z}}$ הוא פיתרון למערכת המשוואות הנורמלית, אז $\hat{\mathbf{z}} = \mathbf{x}\mathbf{w}$ היא ההטלה האורתוגונלית של \mathbf{y} על התמונה של \mathbf{X} ו- $\hat{\mathbf{z}} - \mathbf{y} = \hat{\mathbf{y}}$ הוא הנורמל (האנך) לתמונה. מכאן נובע השם "משוואות נורמליות"¹⁷.

פתרון המשוואות הנורמליות

כפי שראינו מזוית גיאומטרית, במקרה בו $d + 1 \geq m$, פתרון המשוואות הנורמליות הכוונה למציאת וקטור $\hat{\mathbf{z}}$ ש- $\hat{\mathbf{z}} - \mathbf{x}\mathbf{w} = \hat{\mathbf{y}}$ הוא הטלה אורתוגונלית של \mathbf{y} על התמונה של \mathbf{X} . נוכל להסיק מכאן שתי מסקנות על הקיום והיחידות של פיתרון למשוואות אלו:

¹⁷כאן הייתה תמונה שהמטרה שלה הייתה לבחיר את הנושא, אבל מושב טעויות נראה לי שהוא רק מבלבת. אז נותר.

□ קיומ - כיוון שמדובר במערכת משווהות ליניארית, יתכנו שלוש אפשרויות: אין פיתרון למערכת המשווהות, יש פיתרון יחיד, ויש אינסוף פתרונות. מהחנה הגאומטרית עולה כי האפשרות הראשונה לא אפשרית. אם כך, יש פיתרון יחיד או אינסוף פתרונות.

□ ייחדות:

- אם העמודות של X הן בלתי תלויות, אז בפרט \hat{w} יכול להיות מוצג בצורה ייחודית כצירוף של העמודות אלו. כלומר, יש פיתרון ייחוד.
- אם העמודות מכילות וקטורים תלויים, אז ההטלה \hat{y} יכולה להיות מוצגה כאינסוף צירופים ליניארים של עמודות X . מספיק למצוא וקטור אחד שספק את הסchorה.

המקרה הראשון - וקטורים בלתי תלויים

כל להראות כי לכל מטריצה מתקדים כי $\ker(A^\top A) = \ker(A)$ ¹⁸. בעקבות כך, אם הגרען (X) הוא טריוואלי, בהכרח גם $\ker(A^\top A)$ הוא טריוואלי. כלומר, בפרט המטריצה זו הפיכה (לא סינגולרית). כלומר:

$$\begin{aligned} X^\top y &= X^\top X w \\ &\Downarrow \\ [X^\top X]^{-1} X^\top y &= [X^\top X]^{-1} X^\top X w \\ &\Downarrow \\ \hat{w} &= [X^\top X]^{-1} X^\top y \end{aligned}$$

לסיום, נרצה להראות כי כי w הוא אכן מזער של RSS. ניקח את הנזרת השנייה ונקבלו:

$$\frac{\partial^2 RSS(w, X, y)}{\partial w_k \partial w_l} = \frac{\partial -2 \sum_{i=1}^m \left(y_i - \sum_{j=1}^d x_{ij} w_j \right) x_k}{\partial w_l} = 2 \sum_{i=1}^m x_k x_l = 2 [X^\top X]_{kl}$$

אפשר לבדוק אם המטריצה $X^\top X$ היא חיובית למחצה. כיוון שהנחנו כי העמודות של X הן בלתי תלויות, נקבל לכל $i \neq l$ כי:

$$v^\top [X^\top X] v = (Xv)^\top Xv = \|Xv\|^2 > 0$$

דוגמה

הבה ונמצא אומדן \hat{w} לתמונה הבאה. נניח ואנחנו מעריכים בהערכות זמן ריצת 100 מטר, לפי גובה ומשקל. אספנו את הנתונים של האצנים המהירים ביותר באולימפיאדת ריו:

¹⁸בפרט נראה בתרגיל.

אטלט	משקל	גובה	זמן ריצה
יוסיאן בולט	94	195	9.81
ג'סטין גטליין	79	185	9.89
אנדראה דה גראס	70	176	9.91
יוהאן בליק	80	180	9.93

אם כך, התוכנות שלנו הם המשקל והגובה, והתגובה הינה זמן הריצה. תחילה, נארגן את הנתונים (בתוספת h -intercept):

$$\mathbf{x} := \begin{bmatrix} 1 & 94 & 195 \\ 1 & 79 & 185 \\ 1 & 70 & 176 \\ 1 & 80 & 180 \end{bmatrix}, \quad \mathbf{y} := \begin{bmatrix} 9.81 \\ 9.89 \\ 9.91 \\ 9.93 \end{bmatrix}$$

כפי שראינו קודם לכן, האומדן נתון על ידי $\hat{\mathbf{w}} = [\mathbf{x}^\top \mathbf{x}]^{-1} \mathbf{x}^\top \mathbf{y}$. אם כך, קיבל סך הכל מהמידע (מה שנקרה: וודאו בעצמכם) כי $\mathbf{y}^\top (11.38, 0.003, -0.009) \approx \hat{\mathbf{w}}$. אם ניקח מידע חדש, יוכל להעריך זאת על פי האומדן שמצאנו:

$$\hat{y} = \mathbf{x}^\top \hat{\mathbf{w}} = \left\langle \begin{bmatrix} 1 \\ 74 \\ 176 \end{bmatrix}, \begin{bmatrix} 11.38 \\ 0.003 \\ -0.009 \end{bmatrix} \right\rangle = 10.018$$

המקרה השני: וקטורים תלויים - $\dim(\ker(\mathbf{X})) > 0$
אם העמודות של \mathbf{X} תלויות, אז ישנו אינסוף דרכים להציג את \mathbf{y} כצירוף לINIARI של \mathbf{X} . נרצה דרך אחרת, ולכן נחפש פיתרון $\hat{\mathbf{w}}$ שקרוב לראשית \mathbb{R}^{d+1} . נוכל למצוא זאת באמצעות SVD.

הגדרה

תהי $\mathbf{X} \in \mathbb{R}^{m \times d+1}$ ותהי $\mathbf{X} = U\Sigma V^\top$ ה-SVD שלה. הפסאודו-הופכית (מורה-פנורוסה-Moore-Penrose-) של \mathbf{X} מוגדרת על ידי $\mathbf{x}^\dagger = V\Sigma^\dagger U^\top$ כאשר Σ^\dagger היא מטריצה $m \times d$ אלכסונית שמוגדרת על ידי:

$$\Sigma_{i,i}^\dagger = \begin{cases} 1/\Sigma_{i,i} & \Sigma_{i,i} \neq 0 \\ 0 & \Sigma_{i,i} = 0 \end{cases}$$

זו הכללה של המטריצה ההופכית. ואכן, כאשר \mathbf{X} הפיכה, אז $\mathbf{x}^\dagger = \mathbf{x}^{-1}$.

תכונה חשובה של הפסאודו-הופכית, היא כי אם עבור מערכת משווהות לה יש אינסוף משווהות $\mathbf{Ax} = \mathbf{b}$ – אז $\mathbf{A}^\dagger \mathbf{b}$ היא פיתרון עם נורמת ℓ_2 מינימלית. כלומר:

$$\mathbf{A}^\dagger \mathbf{b} = \operatorname{argmin}_{\mathbf{x}} \{ \|\mathbf{x}\|_2 \mid \mathbf{Ax} = \mathbf{b} \}$$

במקרה שלנו, הפתרון למערכת המשוואות הנורמלית עם נורמת ℓ_2 מינימלית, והקרוב ביותר לראשית אם כך, הינו $\hat{\mathbf{w}} = \mathbf{X}^\dagger \mathbf{y}$. ניתן להראות כי כאשר יש לנו מטריצה בת"ל (הגרעין שווה ל-0), אז $\mathbf{y} = \hat{\mathbf{w}}$ (בתרגיל הבית).

טענה

תהי $\mathbf{X} \in \mathbb{R}^{m \times (d+1)}$ בעיית רגרסיה כאשר $\dim(\ker(\mathbf{X})) \neq 0$ ו- $m \geq d + 1$. אז $\hat{\mathbf{w}}$ הוא מזער של RSS.

הוכחה

נסמן $r = \text{rank}(\mathbf{X})$. כיוון שהגרעין של \mathbf{X} לא טריוויאלי, אז $1 \leq r < d + 1$ והערכים הסינגולריים הינם $\sigma_1 \geq \dots \geq \sigma_r > 0$. בicutת תהי $\mathbf{X} = U\Sigma V^\top$ ה-SVD של \mathbf{X} , כאשר העמודות של U, V מהווים בסיסים אורתוגונליים לארבעת תתי המרחבים הבסיסיים:

$$\begin{array}{ll} U_{\mathcal{R}} \in \mathbb{R}^{m \times r} & \mathcal{R}(\mathbf{X}) = \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\} \\ V_{\mathcal{R}} \in \mathbb{R}^{(d+1) \times r} & \mathcal{R}(\mathbf{X}^\top) = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\} \\ U_{\mathcal{N}} \in \mathbb{R}^{m \times (m-r)} & \mathcal{N}(\mathbf{X}) = \text{span}\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_m\} \\ V_{\mathcal{N}} \in \mathbb{R}^{(d+1) \times (d+1-r)} & \mathcal{N}(\mathbf{X}^\top) = \text{span}\{\mathbf{v}_{r+1}, \dots, \mathbf{v}_{d+1}\} \end{array}$$

ותהי $\mathcal{S} \in \mathbb{R}^{r \times r}$ מטריצה אלכסונית עם r ערכים סינגולריים שאינם אפסים (ראינו כי המימד של המטריצה שווה במספר הערכים הסינגולריים שונים מאפס). בicut, נוכל להזכיר בהצעה הקומפקטיבית של ה-SVD:

$$\mathbf{X} := U\Sigma V^\top = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \mathcal{S} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{\mathcal{R}}^\top \\ V_{\mathcal{N}}^\top \end{bmatrix} = U_{\mathcal{R}} \mathcal{S} V_{\mathcal{R}}^\top = \tilde{U} \tilde{\Sigma} \tilde{V}^\top$$

(בתכל'ס, אם נחזור לפרק של ליניארית, זה מה שראינו Katachiachi זה, על מנת להיפטר מהערכים שווים לאפס. הסתכלו שם.). בicut, על מנת למזער את RSS, נקבל:

$$\begin{aligned} \mathbf{X}^\top \mathbf{y} &= \mathbf{X}^\top \mathbf{X} \mathbf{w} \Rightarrow \\ \tilde{V} \tilde{\Sigma}^\top \tilde{U}^\top \mathbf{y} &= \tilde{V} \tilde{\Sigma}^\top \tilde{U}^\top \tilde{U} \tilde{U}^\top \tilde{V}^\top \mathbf{w} \\ \tilde{\Sigma} \tilde{U}^\top \mathbf{y} &= \tilde{\Sigma}^2 \tilde{V}^\top \mathbf{w} \end{aligned}$$

כיוון ש- $\tilde{\Sigma} \in \mathbb{R}^{r \times r}$ היא בעלת מימד מלא, אז $\tilde{\Sigma}^{-1}$ קיים ולכן $\mathbf{w} = \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}^\top \mathbf{y}$. בשימוש, בשימוש במטריצה הפסאודו הופכית, נוכל להרחיב את הצורה הקומפקטיבית של ה-SVD ולקבל:

$$\begin{aligned} \hat{\mathbf{w}} &= \tilde{V} \tilde{\Sigma}^{-1} \tilde{U}^\top \mathbf{y} \\ &= V \Sigma^\dagger U^\top \mathbf{y} \\ &= \mathbf{X}^\dagger \mathbf{y} \end{aligned}$$

מסקנה

הוקטור $y^{\dagger} = \hat{X}$ הוא תמיד פיתרון למערכת המשוואות הנורמלית.

1.3 שיקולים נומריים בעת מימוש אלגוריתם

עד כה התעסקנו באיך לעצב אלגוריתם למידה. בעת נרצה לממש את זה, כלומר לכתוב אלגוריתם יעיל שייממש את האלגוריתם שעיצבנו. נוכל לעשות זאת באמצעות אלגברה נורמלית. במקרה שלנות נדרש לחשב את ה-SVD. בקורסים הבסיסיים של אלגברה ליניארית התעסקנו בדברים יחסית פשוטים ולכן לא עצרנו לחשב איך לחשב את המטריצה ההופכית במחשב. זה לא זהה פשוט כי שזה נראה - מחשבים לא ידעים מה זה \mathbb{R} והם משתמשים בביטים ובריתמטיקה עם דיק סופי.¹⁹

הבה ונראה דוגמה פשוטה לשיקול נומירי באלגברה ליניארית. אנחנו יודעים כי אם הנגרען של X הוא אפס והמטריצה לא סינגולרית, אז נוכל בנוסחה פשוטה למצוא את ההופכית. אבל מה יקרה אם $X^T X$ "כמעט הפיכה"? במצב זה, נוכל הגיעו לביעות נומריות.

למשל, אם השתמש בנוסחה $y^{\dagger} = \hat{X}^{-1} X^T y$, נראה כי הערכים הסינגולרים היכי קטנים של X הם מאוד קטנים, וכך כאשר נרצה לחשב את $\frac{1}{\sigma_i}$ קיבל חוסר דיק. על מנת להימנע מבעיה זאת, אנחנו בוחרים "סף דיק נומירי" $0 > \varepsilon$. נוכל לבחור למשל $\varepsilon = 10^{-8}$. כאשר נשנה את ההגדלה מעט ונקבל:

$$\Sigma_{i,i}^{\dagger,\varepsilon} = \begin{cases} 1/\sigma_i & \sigma_i > \varepsilon \\ 0 & \sigma_i \leq \varepsilon \end{cases}$$

1.4 הוספת רעש

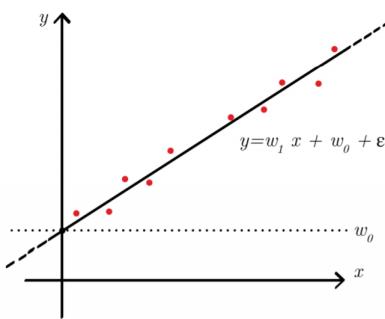
עד כה הוכיחנו כי לוקטור התגובה y יש התנהגות דטרמיניסטית מעל המדגם x ושישנה פונקציה דטרמיניסטית $y \rightarrow \mathcal{X} : f$ שיזכרת את הקשר $y \rightarrow \mathcal{X}$.

זו הנחה שאינה מציאותית, למעשה במציאות תיתכן **רנדומיות**.

על מנת להגדיר את הבעיה ולהתמודד אליה, נגיד מודל הסטרטורי על ה-dataset. נניח כי הקשר $y \rightarrow \mathcal{X}$ הוא ליניארי, אבל נסייף פקטור שיזכר מקרים בקשר הזה. נניח בעת כי ישנה פונקציה $y \rightarrow \mathcal{X} : f(x) = \sum z_i - z(x)$ כאשר z הוא משתנה מקרי. נניח למשל כי הרעש z במדגם הוא מוטפלג בצורה שווה התפלגות. כאמור, מרחיב האימון שלנו הוא S כאשר $S = \{(x_i, f(x_i) + z_i), i=1, \dots, m\}$ בדומה בלתי תלולה ושווה התפלגות. ניקח את המודל שיעיצבנו במקרה הדטרמיניסטי לדוגמה עם הרעש. נבחר מרחב היפוטזה ליניארי \mathcal{H}_{reg} כאשר אלגוריתם הלמידה שלנו מיצא פרדיקציה כלשהי.

קודם לכן הוכיחנו כי $d+1 \geq m$ וגם כי קיימים $w \in \mathbb{R}^{d+1}$, כך שלכל וקטור דגימה x נקבל $z_i^T w + y_i = x_i^T w$:

¹⁹יש תחום שלם שנקרו אלגברה ליניארית נומרית שמתעסק בתחום זה. כאןו שעוסקים במערכות לומדות, עליינו להיות מודעים לדברים אלו כמו שאפשר ובקבות כך להכיר כיצד האלגוריתם עובד בהעמeka.



נסמן את הרעש בתוור וקטור, כלומר, כולם $(z_1, \dots, z_m)^T := z$ ונקבל תצוגה אחרת לוקטור התגובה $z = Xw + \epsilon$.
משמעותו של כי יתכן שהוקטור זה כלל לא נמצא בתמונה של x וכו'. כמו קודם לכן, נוכל להשתמש ב-RSS ולכון

$$\hat{w} := \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \|y - Xw\|^2$$

כלומר, תכל"ס אלגוריתם הלמידה שלנו נשאר אותו הדבר.

1.4.1 עקרון הסבירות המירבית (The Maximum Likelihood principle)

נניח שנרצה לפטור בעיה עם רעש בלבד. נניח כי הרעש מתפלג גausnit, כלומר, $z_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
במציאות וקטורי, נניח כי מרחיב האימון (traning set) שלנו הוא התפלגות גausnit מרובת משתנים: $(Xw, \sigma^2 I_m) \sim \mathcal{N}$.

נניח我们知道 את וקטור המשקל w , אנחנו יכולים לשאול את השאלה הבאה. בהינתן מטריצה X וקטור מקדים w , מה ההסתברות לקבל את וקטור התגובה y ? כיוון שכל דגימה היא בלתי תלויות לחברתה, פונקציית הצפיפות היא מכפלת הצפיפות של כל דגימה (ראינו את זה קודם):

$$p(y | w) = \prod_{i=1}^m \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i^T w - y)^2}{2\sigma^2}\right) \right]$$

זו שאלה בהסתברות. אנחנו ידועים את w 我们知道 את w ונתנו שואלים מה הסיכוי לקבל את y . אמנס, בובאו לעצב אלגוריתם למידה אנחנו למשהו מעתניינים בשאלת החפוכה. יש לנו מרחיב מדגם שכולל את וקטור התגובה y . אנחנו מעוניינים למצוא כל חיזוי ליניארי, H_{reg} וVecto w . אנחנו יכולים לשאול מה הערך w הכי קרוב שאנו יכולים להגיע אליו. דבר זה נקרא "עקרון הסבירות המירבית" (The Maximum Likelihood principle-ML). העיקרון הזה מציע לנו לבחור את w כך שפונקציית הצפיפות שmbiah את y תהיה מקסימלית. על מנת לעשות זאת פורמלית, נתחל בהדרות.

הגדרה

יהי X משתנה מקרי עם התפלגות \mathcal{F} ופונקציית הצפיפות f , שתלויה בפרמטר $\theta \in \Theta$. **פונקציית הסבירות (likelihood function)** מוגדרת על ידי $\mathcal{L}(\theta | X) = f_\theta(X)$

על מנת להציג את ההגדרה, נתבונן בדוגמה במקרה של אחד, כפונקציה של הווקטור x , ובහינתו וקטור תגובה y :

$$\begin{aligned}
 & \underbrace{\text{פירוק לרכיבים}}_{\downarrow} \\
 \mathcal{L}(\mathbf{w} | X, \mathbf{y}) &= \underbrace{\text{אי תלות}}_{\downarrow} \\
 & \mathbb{P}(y_1 = \mathbf{x}_1^\top \mathbf{w}, \dots, y_m = \mathbf{x}_m^\top \mathbf{w} | \mathbf{X}, \mathbf{w}) = \\
 & \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}\right) \\
 & = \frac{1}{(2\pi\sigma^2)^{m/2}} \prod_{i=1}^m \exp\left(-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}\right) \\
 & = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2\right)
 \end{aligned}$$

כעת, נוכל להבחן כי האומדן לשבירות המירבית (maximum likelihood estimator) בוחר את הפרמטר שמקסם את פונקציית השבירות.

הגדרה

תהי \mathcal{L} פונקציית סבירות של פונקציית הסתברות \mathcal{F} כלשהיא שתלויה בפרמטר $\Theta \in \theta$ ויהי X משתנה מקרי

המ��פלג בהתאם ל- \mathcal{F} .

האומדן לשבירות המירבית (MLE) ל- θ מוגדר על ידי:

$$\hat{\theta}^{MLE} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | X)$$

הבה ונתבונן בדוגמה שלנו. נרצה למצוא את ה-MLE של מודל הרגרסיה הлиニアרית:

$$\begin{aligned}
 & \underbrace{\text{מונטוניות}}_{\downarrow} \\
 \hat{\mathbf{w}}^{MLE} &= \operatorname{argmax}_{\mathbf{w}} \mathcal{L}(\mathbf{w} | \mathbf{y}) = \\
 & \underbrace{\text{ביטול לוג}}_{\downarrow} \\
 & = \operatorname{argmax}_{\mathbf{w}} \log \mathcal{L}(\mathbf{w} | \mathbf{y}) = \\
 & \underbrace{\text{מינימום של מינוס פונקציה}}_{\downarrow} \\
 & \operatorname{argmax}_{\mathbf{w}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2\right) = \\
 & = \operatorname{argmin}_{\mathbf{w}} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{w} - y_i)^2
 \end{aligned}$$

אם כך, נוכל לסכם כי MLE (בහינת התפלגות נאותונית *iid*) זהה לחלוטון לשיטת הריבועים הפחותים (Least Squares), שמתќבל מעקרונו שונה לחלוטון.

1.5 משתנים קטגוריים

נתבונן בבעיה הבאה. נניח ואנחנו רוצים לחזות מחיר של בית בהתבסס על הנתונים הבאים:

- מחיר הבית הוא משתנה נומרי שמקבל ערכים חיוביים.
- מחיר הגן הוא משתנה קטgoriy, שמקבל את הערכים: קטן, בינוני וגדול.
- מספר השירותים הוא משתנה קטgoriy נומרי שמקבל מספרים טבעיות.
- סוג הבית הוא משתנה קטgoriy שמקבל את הערכים: בית פרטי, דירה ודירה סטודיו.

נרצה לבנות את בעיית הרגרסיה הבאה:

$$y = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{w} = \begin{bmatrix} w_{\text{house-size}} \\ w_{\text{garden-size}} \\ w_{\text{number-of-bedrooms}} \\ w_{\text{house-type}} \end{bmatrix}$$

נוכל להבחן כי יש לנו **סוגים** שונים וזה לא ברור כיצד להתייחס לכל אחד מהם. גודל הבית ומספר השירותים הם משתנים **כמותיים** שאפשר לסדר את ערכיהם ולכן נוכל לפתור מערכת משוואות ליניארית ומילא לפתור בעיית רגרסיה ליניארית.

במקרה של 'גודל הגן', למשל, למורות שהערכים לא נומריים, נוכל לסדר אותם ממקטן לגודל ולהגדיר למשל קטן בתו 1, בינוני בתו 2, גדול בתו 3.

אך מה לגבי סוג הבית? האם נוכל ליצר מפה לוגית כמו שעשינו למספר השירותים? הדרך הנפוצה לעשות זאת, זה למעשה להוסיף עמודות, כך שהעמודות האחרות הם למעשה משתנים **бинאריים**, כלומר משתנים ששייכים ל- $\{0, 1\}$. במקרה זה, קיבל:

$$\mathbf{x}_{\text{type house}} = \text{'apartment'} \Rightarrow \mathbf{x} = \begin{bmatrix} x_1 & \text{size house} \\ x_2 & \text{size garden} \\ x_3 & \text{bedrooms of number} \\ x_4 & \text{private-house} \\ x_5 & \text{apartment} \\ x_6 & \text{studio-apartment} \end{bmatrix}$$

$$\text{כאשר } x_4 + x_5 + x_6 = 1, x_4, x_5, x_6 \in \{0, 1\}$$

2 התאמה פולינומית (Polynomial fitting)

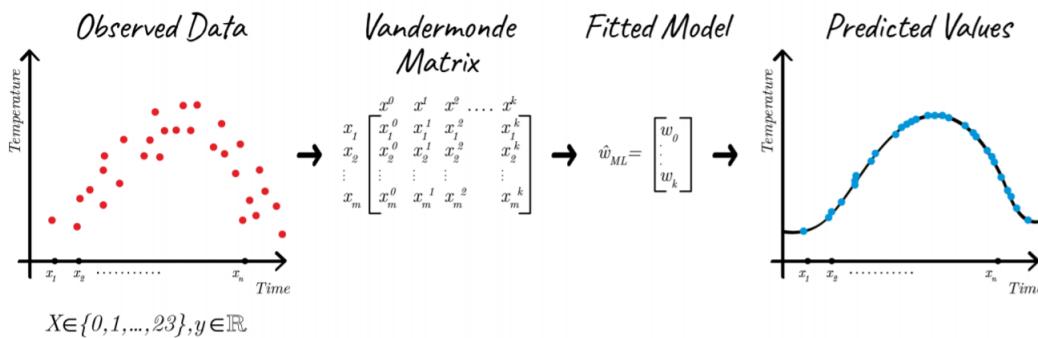
cut נ עבור לדוגמה ספציפית של רגרסיה ליניארית, שתעזר לנו להבין עקרונות כלללים שגם חשובים בפני עצמם. נשתמש באלגוריתם הלמידה שלנו עבור רגרסיה ליניארית, על מנת לחזות את הערך של פונקציה ממשית $g : \mathbb{R} \rightarrow \mathbb{R}$ בנתים ידועה. אנחנו מקבלים נקודות \mathbb{R} נקודות y_1, \dots, y_m (labels). במקרה החסר רוש,

אנחנו יודעים כי $y_i = g(a_i)$ עבור פונקציה g כלשהיא, ונרצה לחזות את הערך של g על נקודות בהתאם לקבוצת האימון.

מחלקת ההיפותזה הינה $\{x \mapsto p_w(x) \mid w \in \mathbb{R}^{d+1}\}$.
 בוגר, כאשר נתון לנו מוגדים אימון $\{(a_i, y_i)\}_{i=1}^m$, נרצה לבחור את וקטור הממקדמים הפולינומי w , בשימוש בשיטת LS.
 כמובן למצוא את w מינימום $\min_{w \in \mathbb{R}^{d+1}} \frac{1}{m} \sum_{i=1}^m (y_i - p_w(a_i))^2$. על מנת לעשות זאת, נבנה בעית גרסה ליניארית מתאימה.
 ניקח את המטריצה X שנבנה אותה כך $(a_i^0, a_i^1, \dots, a_i^d)$

$$X = \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^d \\ 1 & a_2 & a_2^2 & \cdots & a_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_m & a_m^2 & \cdots & a_m^d \end{bmatrix}$$

אפשר לשים לב שמדובר במטריצת נדרמונייה ולכן היא ממיימת מלא. דבר זה אומר כי פתרון מערכת המשוואות הליניארית יכול להיעשות כמו שראינו במקרה הכללי של S-LS - אינטואיטיבית, מה שאנו מוחשים תכל'ס זה פולינום שיטאים לבועיה שלנו. אנחנו לאוראה משתמשים ב-S-LS - מזערם את המרחק מכל הנקודות (תclf יש ציר ואפשר לחבר נקודות, ככלומר לשים קו שעובר בדיק בין הנקודות):
 אנחנו מוחשים פולינום בדרجة כלשהי שעובר בינו כל הנקודות:



ככל שהפולינומים שאנו בוחרים גדול יותר, אויה ההתאמת מדוקית יותר.

2.1 הטיה וסוגות של אומדנים

בhinintu בעיה גרסה ליניארית $y = Xw + \epsilon$, ראיינו כיצד לפתור את הבעיה $\epsilon = y - Xw$, למצוא וקטור w כך ש- $y \approx Xw$.
 כמו כן, הראיינו כי הוקטור שמשמש את סכום המרכיבים הריבועיים (LS) נתון על ידי $\epsilon^\top \epsilon = \|y - Xw\|^2$. חשוב להבחין כי כיוון ש- y הוא משתנה מקרי, אז גם ה-LS הוא משתנה מקרי. ולכן נוכל מאפיינים שונים של כל אומדן.

נזכיר בהגדלה של הטיה ובסוגות של האומדן.

הגדרה

יהי $\hat{\theta}$ אומד כלשהו של θ . **ההטיה** של $\hat{\theta}$ מוגדרת להיות ההפרש בין התוחלת של $\hat{\theta}$ ו- θ .
כלומר, $B(\hat{\theta}) := \mathbb{E}[\hat{\theta}] - \theta$. נאמר כי $\hat{\theta}$ חסר הטיה, אם $B(\hat{\theta}) = 0$.

הגדרה

יהי $\hat{\theta}$ אומדן של θ . **שותות** של $\hat{\theta}$ מוגדרת על ידי $\text{var}(\hat{\theta}) := \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$.

מה השונות והતוחלת של האומדנים הללו? כפי שראינו, אומדן הוא פונקציה מעלה מוגדרת $S = x_1, \dots, x_m \in \mathbb{R}^d$ שנוועדה להעריך את הפרמטר θ $\hat{\theta}(x_1, \dots, x_m) \stackrel{?}{\sim} \mathcal{N}(0, \sigma^2 I)$. במקרה, התוחלת של $\hat{\theta}$ היא לפי בחירת המדגמים. נחזור ל-LS-
- נוכל לשאול מה התוחלת והשותות.

תרגיל

תהי $y = Xw + \varepsilon$ ביעית וגרסיה ליניארית, כך ש-OLS האומדן \hat{w} (האומדן שראינו עם המטריצות הכפיות). הראו כי \hat{w} הוא אומדן בלתי מוטה.

הוכחה

$$\begin{aligned}\mathbb{E}[\hat{w}] &= \mathbb{E}\left[\left[X^\top X\right]^{-1} X^\top y\right] \\ &= \mathbb{E}\left[\left[X^\top X\right]^{-1} X^\top (Xw + \varepsilon)\right] \\ &= \mathbb{E}\left[\left[X^\top X\right]^{-1} X^\top Xw\right] + \mathbb{E}\left[\left[X^\top X\right]^{-1} X^\top \varepsilon\right] \\ &= \mathbb{E}[w] + \left[X^\top X\right]^{-1} X^\top \mathbb{E}[\varepsilon] = w\end{aligned}$$

$$\text{השוויון האחרון נכוון כיון ש-} \mathbb{E}[w] = 0 \text{ ו-} \mathbb{E}[\varepsilon] = 0$$

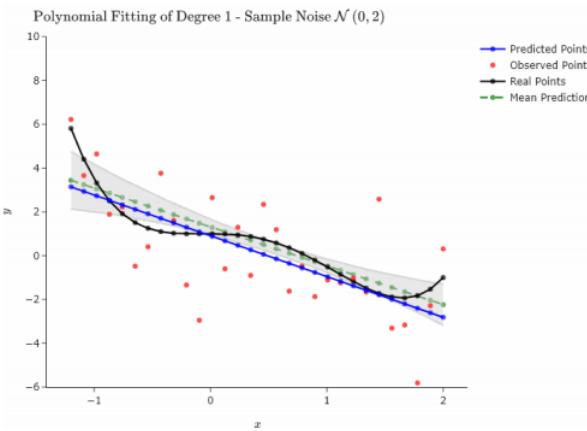
על מנת לקבל מעט אינטואיציה לנושא, נחזיר שוב להערכת פולינומית. נתבונן למשל בדוגמה $Y = X^4 - 2X^3$ – $\varepsilon \sim \mathcal{N}(0, 2)$ כאשר $X = 0.5X^2 + 1 + x$.

נניח כי $x_m, \dots, x_1 \in [-2, 2]$ הינם סט של דגימות כך ש- x_i . נייצר dataset 10 בהסתמך על המודל למעלה. לכל dataset נשתמש ב- x_m, \dots, x_1 ונិיצר את וקטור התגובה y_m, \dots, y_1 בתוספת הרעש. נוכל לראות את ה- \hat{y}_i השוניים שנוצרים על ידי המודל, ובהתאם, גם משפיעים על ערך החיזוי. דבר זה תלוי במרקוריות של המדגמים.

בציר הבא ניתן לבדוק ב佗יפות הבאות:

בירוק, מופיע ממוצע החיזוי של y לכל x בקבוצות המידע. ההבדל בין הקוו השחור והירוק מוגדים את העניין של ההטיה.

באפור ניתן לראות את התוחם של התוחלת של התגובה, ככלומר $\mathbb{E}[\hat{y}] \pm 2 \cdot \text{Var}(\hat{y})$ עבור x כלשהו. דבר זה נקרא גם רוח בר סמך (confidence interval).



כל שקלט החלטה רגש יותר, ישנה שונות גבואה יותר. כאשר נבוא לушות פרדיקציה, הפרדיקציה תלולה במידה במדד שראינו.

כל שהמודל עשיר יותר, השונות תגדל, אבל ההטיה תקטן. לעומת זאת, אם אנחנו מגדילים את הרעש, אז למעשה תהיה שונות מאוד גבואה כי יהיה קשה לחזות על פי זה.

אנו נחפש את הפונקציה שמצוות את האיזון בין השונות ובין ההטיה. מבחינה מתמטית, ניתן לראות כיצד השונות וההטיה מתחברים.

נסמן $(S) \hat{y} = \hat{y}$ האומדן של y אם נסמן ols ו- \hat{y} ערכי y האמתיים. כאשר נפתח את בעיית הרגריסיה, נרצה למזער את MSE בין שני הערכים הללו.

אם נסמן $\mathbb{E}[\hat{y}] = \bar{y}$, נקבל:

$$\begin{aligned}\mathbb{E}[(\hat{y} - y^*)^2] &= \mathbb{E}[(\hat{y} - \bar{y} + \bar{y} - y^*)^2] \\ &= \mathbb{E}[(\hat{y} - \bar{y})^2] + 2(\hat{y} - \bar{y})\mathbb{E}[\hat{y} - \bar{y}] + (\bar{y} - y^*)^2 \\ &= \mathbb{E}[(\hat{y} - \bar{y})^2] + (\bar{y} - y^*)^2 \\ &= \text{var}(\hat{y}) + \text{Bias}^2(\hat{y})\end{aligned}$$

כלומר (ועוד נראה זאת בהמשך), ניתן לפרק את שגיאות ההכללה (generalization error)²⁰ למרכיב השונות ורכיב התוחלת. לעומת זאת בהמשך), ניתן לפרק את שגיאות ההכללה (generalization error)²⁰ למרכיב השונות ורכיב התוחלת. לעומת זאת בהמשך), ניתן לפרק את שגיאות ההכללה (generalization error)²⁰ למרכיב השונות ורכיב התוחלת. לעומת זאת בהמשך), ניתן לפרק את שגיאות ההכללה (generalization error)²⁰ למרכיב השונות ורכיב התוחלת.

$$\text{MSE}(\hat{y}) = \text{Var}(\hat{y}) + \text{Bias}^2(\hat{y})$$

אם כך, נוכל לסכם בשלב זה כי כאשר נשימוש באומדן על המידע שלנו, שגיאת ההכללה מושפע משני פקטוריים אלו - דבר זה נקרא Bias-Variance Trade-off.

²⁰במקרה שלנו - הממוצע של ההפרש בין החיזוי ובין הערך האמתי. בהמשך נכליל זאת גם לפונקציות הפסד (loss) אחרות.

חלק III

בעיות סיווג (Classification)

1 הקדמה לשיווג

מהי בעיית סיווג? נרצה לסווג כל מיני פיצ'רים ולחזות על פיהם בעתיד. תחילה, נתמקד בתת מרחב עם d פיצ'רים, $\mathbb{R}^d = \{\pm 1\}^d$. ישנו הרבה סוגים סיווג (קליסיפיקציה): לחזות למשל האם מטופל יפתח בעיה רפואית או לא. האם משתמש יאהב מוצר חדש וכן הלאה. בנוסף, לבדוק האם תעבורת רשות היא התקפת סייבר, האם יוצרה היא זיווג, האם יש ספרם, האם תעבורת הרכטיס אשראי תקינה או לא.

דוגמה

נתבונן בדוגמאות של התקפי לב לגברים בדרום אפריקה. נניח שלקחו כל מיני מדדים (האם יש היסטוריה של מחלות לב וכוכו):

id	idcode	age	adiposity	height	type	density	glucose	spx	chd
0	100	32.00	5.73	23.71	Present	49	25.20	97.20	82 1
1	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63 1
2	118	0.08	3.48	32.88	Present	52	29.14	3.81	46 0
3	170	7.50	6.41	38.03	Present	51	31.99	24.28	58 1
4	134	19.90	3.50	27.78	Present	60	25.99	57.94	49 1
487	214	13.40	5.98	31.72	Absent	64	28.45	0.00	58 0
488	182	4.20	4.41	32.10	Absent	52	28.61	18.72	52 1
489	106	3.00	1.91	15.23	Absent	40	20.09	26.64	55 0
490	110	5.60	11.61	30.79	Absent	64	27.32	23.84	40 0
491	132	0.00	4.82	33.41	Present	62	14.70	0.00	40 1

התגובה היא "האם יש מחלת לב".

האם נוכל לדעת מותווך המידע זהה האם מישחו חשוף למחלת לב? במקרה שלנו, נתבונן בדוגמה של $d=2$ (כי לא ניתן לראות ממשו מממד גדול).

1.1 פונקציית הפסד (Loss Functions)

כיצד נחליט מהו מסווג טוב?
נגיד:

$$L_5(h) := \sum_{i=1}^m 1_{y_i \neq h(x_i)} = |\{i \mid y_i \neq h(x_i)\}|$$

בפשטות, מדובר במספר מס' פעמים בהם כלל החלטה 'טעה'. דבר זה נקרא 'דיוק'. כמובן, אנחנו מעוניינים במסווג שהייה הטוב ביותר בהתבסס על המידע שלו. אמנם, הדבר לא מדויק, ונרחיב קצת.

1.2 שגיאות מסווג ראשון וסוג שני

קודם כל, נבחר מהו נחשב 'שלילי' ומהו נחשב 'חיובי'. ואז נקבל:

	$y = -1$, $y = 1$ true
$\hat{y} = -1$.	error Type-II
$\hat{y} = 1$	error Type-I	.

למשל, אם חיזינו שבנ adam מסוים קיבל תוצאה חיובית לكورونا והוא קיבל תוצאה שלילית (False-positive) וכן החפק. בדרך כלל טעות מסווג ראשוני תהיה חמורה יותר.

דוגמה

נניח ונרצה לבדוק האם תרופה מסוימת בטוחה לשימוש או לא. נגיד כי 1 – זה שהתרופה מסוכנת ו-0 זה תרופה בטוחה.

במקרה זה, טעות מסווג ראשוני היא לחתת תרופה שחורגת אנשים, וטעות מסווג שנייה היא לא לחתת את התרופה כלל.

עלינו להכريع איזו בעיה חמורה יותר. דבר זה משתנה מסווג המידע.

1.3 מדדיות של ביצועים

הבה ונגיד את המונחים של הביצועים.

◻ חיובי - כמה חיובי יצא בסך הכל.

◻ שלילי - כמה שלילי יצא בסך הכל.

◻ חייתי חיובי וצדקתי - True positive.

◻ חייתי חיובי וטעהתי - False positive

◻ חייתי שלילי וצדקתי - True negative.

◻ חייתי שלילי וטעהתי - False negative.

נוכל לחתת את כמה הדברים הללו ולספר אותם.

למשל:

◻ 'כמה שגיאותعشית': $\frac{FP+FN}{P+N}$. מוגדר על ידי Error Rate

◻ 'כמה פעמים צדקתי': Accuracy - $\frac{TP+TN}{P+N}$

◻ Precision - $\frac{TP}{TP+FP}$

◻ Specificity - $\frac{TN}{N}$

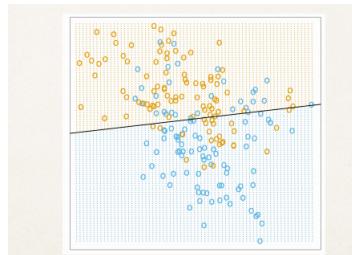
◻ ממד ה-False-Positive-Rate $\frac{FP}{N}$ - FP/N

1.4 גבולות החלטה (Decision Boundaries)

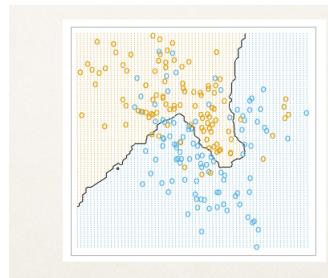
יהי h ככלל החלטה ביןארי ב- \mathbb{R}^d . אנחנו יכולים לקחת כל נקודה $x \in \mathbb{R}^d$ ולהכניס ול- h ולקבל שתי מחלקות.

$$\mathbb{R}^d = \{x \mid h(x) = 1\} \bigcup \{x \mid h(x) = 0\}$$

ואז נקבל את התמונה הבאה:



הגבול בינהם יכול לקבל כל מיני צורות (יכול להיות על משור ועדי):



כעת, בכל פעם שנקבל 'מסווג', נרצה לשאול מה הגבול שלו.

למעשה, יהיו לנו המונחים שאלות על כל מסווג:

1. מהי מחלוקת ההיפותזות H ? מהי קבוצת כל הפונקציות שמתוכם צריך לבחור אחת. כיצד הגבול נראה.
2. מה העיקרונו של פיו אני בוחר $H \in \mathcal{H}_s$ קלומר, אך אני בוחר את ההיפטזה?
3. כיצד נמשח חשיבות העקרונות? כיצד ניתן לכתוב זאת במחשב?
4. כיצד ניתן לישם את המודל מבחינה אלגוריתמית?
5. כיצד ניתן לאחסן את המודל שלמדנו במחשב?
6. בהינתן מודל $H \in \mathcal{H}_s$, כיצד ניתן לחשב את הפרדיקציה החדשה (x) עבור דגימה x חדשה?
7. האם המודל הוא בר פירוש? (נראה בהמשך)
8. האם המודל מספק מידע על הסיכוי להשתתיק למחלוקת מסוימת.
9. האם רואים משפחה אחת של מודלים או כמה?
10. מתי משתמש בכלל אחד מהם?

2 מסווג-חצ-מרח (Half-Space Classifier)

מדובר במחלקת המסווגים פשוטה ביותר. בדומה לגרסיה ליניארית גם כאן נרצה להפריד את המידע לשני חלקים באמצעות מפריד ליניארי. נבוד עם קבוצת התווים $\{ \pm 1 \} = \mathcal{Y}$.

הגדרה

תהי $w \in \mathbb{R}^d$, $b \in \mathbb{R}$. על מישור מוגדר על ידי (w, b) שהינה הקבוצה:

$$\{x \mid \langle w, x \rangle = b, x \in \mathbb{R}^d\}$$

הגדרה

יהי (w, b) על מישור, חצ-המרחב של (w, b) מוגדר על ידי הקבוצה:

$$\{x \mid \langle w, x \rangle \geq b, x \in \mathbb{R}^d\}$$

הפונקציה $x \rightarrow \text{sign}(\langle x, w \rangle + b)$ מתחילה כל נקודה $-1 \leq \langle x, w \rangle + b \leq 1$ על המישור. במקרה ההומוגני, בו $b = 0$, על המישור עובר דרך הראשית, ואז מדובר על הפונקציה שמקבלות 1 בצד אחד, ו-1 בצד השני.

מחלקת ההיפותזה במקרה זה הינה $\{h_w \mid w \in \mathbb{R}^d\} = \mathcal{H}_{\text{half}}$ - קבוצת כל הפונקציות הללו. כיצד נבחר $h_w \in \mathcal{H}_{\text{half}}$? כמובן, מהו כלל הבחירה? במקרה זה "נספור טעויות", על אף כי קודם לכן אמרנו שדבר זה לא מומלץ - ננסה לבצע אופטימיזציה לדבר זה. פורמלית, אם $\langle x, y_i \rangle > 0$, אז קיבלנו מצב בו אין שגיאה (אם המכפלת הפנימית כפול הערך גדולים מאפס, משמע ששניהם 'באותו צד'). במקרה זה נניח כי מרחב המוגנים מופרד ליניארית (ריליאלי!), כלומר שיש על מישור שבו כל "הפלסים" בצד אחד, וכל "המנוסים" בצד שני. מתמטית, נוכל להגיד זאת כי:

$$\exists w \in \mathbb{R}^d, b \in \mathbb{R} \quad \text{s.t.} \quad \forall i \in [m] \quad y_i \cdot \text{sign}(\langle x, w \rangle + b) = 1$$

אם כך, נוכל לקחת את העל מישור הזה ולחפש אותו.

2.1 הפסד אempiri מינימלי (Learning Linearly Separable Data Via ERM)

כפי שראינו, על מנת למצוא את השגיאה המינימלית במקרה ההומוגני, כאמור, אנחנו בודקים האם $y_i \cdot \text{sign}(x^\top w) = 1$ או כי $0 < x^\top w < y_i$. נוכל להגיד את ה'הפסד' בתוור $L_S(h_w) := \sum_{i=1}^m \mathbf{1}_{[y_i \cdot x^\top w < 0]} S$. כיוון שאנו מוחים כי מדובר במקרה הריליאלי, כלומר S היא מופרdat ליניארית, נרצה למצוא $h_w \in \mathcal{H}_{\text{half}}$ שמחלקת את המידע באופן 'מושלם'.

כלומר למצוא היפותזה כזה שתקיים $L_S(h_w) = 0$. בambilים אחרות, אנחנו מפעלים את עקרון ERM ומ Chapman על משורר w^\perp בהתאם להיפותזה h_w שמצוירת את $L_S(h_w)$.

משמעות

כיצד נוכל למצוא את $0 = L_S(h_w)$ מבחן חשיבות? לפי הנחת הריליאנטיות, קיימים וקטור $w_0 \in \mathbb{R}^d$ כך $y_i \cdot x^\top w_0 > 0 \quad i = 1, \dots, m$, קיימים וקטור $w_1 \in \mathbb{R}^d$ כך $y_i \cdot x^\top w_1 \geq 1 \quad i = 1, \dots, m$. נוכל פשוט לנормל את w_0 באמצעות האיבר הקטן ביותר, כלומר $w_0 := \frac{1}{\min_i \{y_i \cdot x^\top w_0\}} w_1$.

2.2 פיתרון ERM בחזאי מרחבים

כאמור, נרצה למצוא את העל-משורר שמצויר את הסיכון. נרצה למצער את 0 בהינתן האילוצים $y_i \cdot x^\top w \geq 1 \quad i = 1, \dots, m$. מדובר בעיית פיזבליות, כי אנחנו רק מchapisms וקטור שמקיים את כל האילוצים האלה בו זמינות ותו לא. למעשה, ניתן לפתור זאת באמצעות אלגוריתם תכנון ליניארי פשוט.

אמנם, ניתן לפתור זאת גם באמצעות אלגוריתם ה-Perceptron הבא.

2.2.1 אלגוריתם ה-Perceptron

אלגוריתם זה הוא אלגוריתם איטרטיבי, כאשר כל וקטור מתקבל בעקבות הווקטורים הקודמים לו. נראה את האלגוריתם:

אלגוריתם 1 אלגוריתם ה-Perceptron

$$w^{(0)} \leftarrow 0.$$

2. לכל $i \leq 1$:

$$(a) \text{ אם } y_i \langle w^{(t)}, x_i \rangle \leq 0$$

$$w^{(t+1)} = w^{(t)} + y_i x_i \quad i.$$

(b) אחרת:

$$w^{(t)} \text{ תחזיר את}.$$

המטרה של האלגוריתם הזה היא למצוא וקטור w כך $y_i \cdot x_i^\top w > 0 \quad \forall i = 1, \dots, m$. דרך העדכון האיטרטיבי של האלגוריתם גורמת לעל המשורר להיות יותר מדויק ככל שמתקדמים. האלגוריתם מניח ריליאנטיות ומשתמש בעקרון ERM. הרעיון של האלגוריתם הוא לעבור דגימה ולבדוק האם טעינו בה או צדקנו. אם טעינו, נתקן אותו בהתאם - אם טעינו "לכיוון ה+", נתקן אותו לכיוון ה"-". במקרה נתקן אותו לפי $-x_i$ (נחוור לה המשך). לבסוף האלגוריתם ייעזר ויחזיר לנו את $w^{(t)}$ (דבר זה נובע מהנחה הריליאנטית). נבחן כי:

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

בכל פעם נתקן מעט את השגיאה, באמצעות הנורמה.

3 המטוווג SVM-Support Vector Machine

מדוע שנרצה להשתמש ב-SVM? לא **תמייד** המידע מוחלק בצורה ליניארית. בנוסף, יתכונו כל מיני 'חצאי מרחבים', שנרצה לדעת כיצד לבחור אחד מהם. מדובר במטוווג שמשתמש באותה מחלוקת היפותזות, אך עקרון הלמידה שלו שונה - והוא איננו משתמש ב-ERM. גם במקרה זה, נתחשב במקרה בו $b = 0$ בלבד.

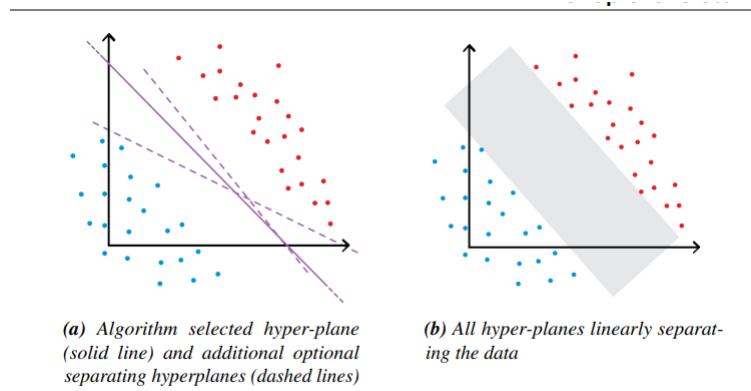


Figure 3.7: Illustration of the existence of multiple separating hyper-planes

3.1 עקרון הלמידה של מקסום השול (Maximum Margin)

הגדרה

יהי $\mathbf{w} \in \mathbb{R}^d$ ויהי $\mathbf{x} \in \mathbb{R}^d$. נגדיר את המרחק בין (\mathbf{w}, b) ו- \mathbf{u} בתוור:

$$d((\mathbf{w}, b), \mathbf{u}) := \min_{v: \langle v, \mathbf{w} \rangle + b = 0} \|\mathbf{u} - v\|$$

טענה

$$\text{אם } d(\mathbf{x}, L) = |\langle \mathbf{w}, \mathbf{x} \rangle + b| \text{ אז } \|\mathbf{w}\| = 1$$

הוכחה

על מנת לפתור זאת, נגדיר מספר נקודות על 'על המשור', נחשב את המרחק מ- \mathbf{x} ונראה מינימליות. ניקוזה $\mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \cdot \mathbf{w}$ היא אכן על המשור:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{v} \rangle + b &= \langle \mathbf{w}, \mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \cdot \mathbf{w} \rangle + b \\ &= \langle \mathbf{w}, \mathbf{x} \rangle - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \|\mathbf{w}\|^2 + b \\ &= \langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle - b + b = 0 \end{aligned}$$

המרחק מ- \mathbf{x} הינו:

$$\|\mathbf{x} - \mathbf{v}\| = |\langle \mathbf{w}, \mathbf{x} \rangle + b| \cdot \|\mathbf{w}\| = |\langle \mathbf{w}, \mathbf{x} \rangle + b|$$

כעת, נסכם כי \mathbf{v} היא הנקודה בעל המישור הקרובה ביותר ל- \mathbf{x} . תהי \mathbf{u} נקודה כלשהיא בעל המישור. אז מתקיים:

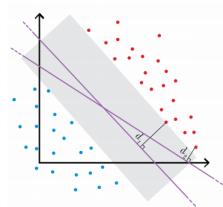
$$\begin{aligned} \|\mathbf{x} - \mathbf{u}\|^2 &= \|\mathbf{x} - \mathbf{v} + \mathbf{v} - \mathbf{u}\|^2 \\ &= \|\mathbf{x} - \mathbf{v}\|^2 + \|\mathbf{v} - \mathbf{u}\|^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &\geq \|\mathbf{x} - \mathbf{v}\|^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &= \|\mathbf{x} - \mathbf{v}\|^2 + 2(\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \\ &= \|\mathbf{x} - \mathbf{v}\|^2 + 2(\langle \mathbf{w}, \mathbf{x} \rangle + b)(\langle \mathbf{w}, \mathbf{v} \rangle - \langle \mathbf{w}, \mathbf{u} \rangle) \\ &= \|\mathbf{x} - \mathbf{v}\|^2 \end{aligned}$$

הגדרה

יהי $S = \{(\mathbf{w}, b)\}$ על מישור ותהי $\mathbf{u}_1, \dots, \mathbf{u}_m \in \mathbb{R}^d$ קבוצה של נקודות. **השול (margin)** של S קובץ של נקודות. **הוֹרְקוֹטּוֹרִים (Support Vectors)** הם הנקודות שמשוגדר על ידי:

$$M((\mathbf{w}, b), S) := \min_{i \in [m]} d((\mathbf{w}, b), \mathbf{u}_i)$$

עקרון הלמידה הוא זה: נבחר $h_{\mathbf{w}, b} \in \mathcal{H}_{\text{half}}$ שלו יש את **השול הגדול ביותר** ביחס למינימום שמן. הווקטורים הקרובים ביותר לעל מישור נקראים **Support Vectors** ומכאן השם של עקרון הלמידה.



3.2 המקרה הריאלי - Hard-SVM

תחילה נתבונן במקרה הריאלי (ניתן להפריד בין חלקים מידע באופן ליניארי).

mbin כל על המישור האפשריים, נצטרך לחפש את על המישור עם השול המקסימלי - המרחק למידע הוא הגדל ביותר.

בצורה פורמלית, נרצה כי:

$$\begin{aligned} & \text{maximize}_{(\mathbf{w}, b)} && M((\mathbf{w}, b), S) \\ & \text{to subject} && y_i \cdot (\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 \quad i = 1, \dots, m \end{aligned}$$

מדובר בבעיית אופטימיזציה. האם מדובר בבעיית אופטימיזציה קמורה?

טענה
יהיו (\mathbf{v}^*, c^*) פתרון אופטימלי של:

$$\begin{aligned} & \text{argmin}_{(\mathbf{w}, b)} && \|\mathbf{w}\|^2 \\ & \text{to subject} && y_i \cdot (\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 \quad i = 1, \dots, m \end{aligned}$$

אזי \mathbf{w}^* עברו $\frac{1}{\|\mathbf{v}^*\|}$ והוא פתרון אופטימלי ל- $|\langle \mathbf{w}, \mathbf{x} \rangle + b| := \gamma$ שהראינו שהוא שcolaה לבעית השול.

מכאן עולה כי מקסום השול הוא למעשה מעשה מזעור הגודל של העל-מישור.
מדובר בבעיית אופטימיזציה קמורה, ולא סתם אלא בבעיית אופטימיזציה עם תכונן ריבועי.

3.3 בלתי ניתן להפרדה ליניארית - Soft-SV

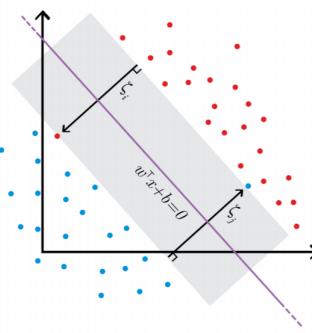
עד כה דיברנו על מקרה בו ניתן לפצל את המידע בצורה ליניארית, אך בדרך כלל זה לא קורה.
אנחנו ניקח את הרעיון שפיתחנו קודם לכן ונרשא בו 'הפרות' - קלומר נרצה כי יש נקודות שעוברות ב'על מישור'
לצד הלא נכון.
כלומר, מעשה:

$$\exists \xi_i > 0 \quad \text{s.t.} \quad y_i \cdot (\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i$$

למעשה, אנחנו משנים את בעית האופטימיזציה ל:

$$\begin{aligned} & \text{minimize} && \|\mathbf{w}\|^2 \\ & \text{to subject} && \begin{cases} y_i \cdot (\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i & i = 1, \dots, m \\ \xi_i \geq 0 \quad \wedge \quad \frac{1}{m} \sum_{i=1}^m \xi_i \leq C \end{cases} \end{aligned}$$

כאשר C הוא קבוע שאנו מוגדרים. המשתנים ξ_1, \dots, ξ_m הם משתני עזר שאנו מוגדרים. נבחן כי ככל שאנחנו מגדילים את C , אנחנו מושרים יותר הפרות של השולדים. מצד אחד, נרצה להרשות הפרות של 'רעש' כדי שה'על מישור' יתעלם מהם. מצד שני, אם נרצה יותר מדי הפרות, נאבד את הצורה של 'על מישור'. זה למעשה מקיים בדיקת הטריד אוף של ההטייה והשונות - ככל ש- C גדול, אנחנו מושרים יותר חופש ל"אלגוריתם הלמידה", "לרווח אחרி הדגימות".



לפעמים, במקומות C , נעדיף לעבוד עם הגדרה מעט שונה. נרצה למזער את הנורמה של w והמוצע של ξ_i . בrama הפורמלית:

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

אם λ מאד גדול, כל הפרה קטנה ממשמעותית ומילא ההטיה נמוכה אך השונות גדולה (בדומה למזה שראינו ברגרסיה לINIARITY). אם λ קטן, אז לכל הפרה יש משקל, ומילא ההטיה גדולה יותר. דבר זה נקרא רגוליזציה. למעשה, כיון שיש תלות ב- λ - מדובר במשפחה של פתרונות ולא בפתרון יחיד.

על מנת לפשט את בעיית האופטימיזציה הזאת, נגידיר את פונקציית hinge loss:

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

טענה

בהתנחת מרחיב מודגם $\{(x_i, y_i)\}_{i=1}^m$ ועל מישור (w, b) , האופטימיזציה של על מישור שcola ל:

$$\operatorname{argmin} \left(\lambda \|w\|^2 + L_S^{\text{hinge}}(w, b) \right)$$

כasher $L_S^{\text{hinge}}(w, b) := \frac{1}{m} \sum \ell(y_i \cdot x_i^\top w)$
הוכחה
בתרגיל.

4 רגרסיה לוגיסטיבית

4.1 מודל הסטברותי עם רעש

אזכיר כי ברגרסיה דיברנו על הקשר $y \sim \mathcal{N}(\mathbf{x}\mathbf{w} + \mathbf{\epsilon}, \sigma^2 I_m)$ כאשר $\mathbf{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_m)$. נשים לב כי כיוון ש- $\mathbf{\epsilon}$ הוא משתנה מקרי, גם y הוא משתנה מקרי שמתפלג נאוסניט:

$$y \sim \mathcal{N}(\mathbf{x}\mathbf{w}, \sigma^2 I_m)$$

אם נתמקד בזוג (\mathbf{x}_i, y_i) נוכל לחשב על זה כמו ההסתברות המותנית של y_i בהינתן x_i . כמובן:

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \mathcal{N}(y_i | \phi_{\mathbf{w}}(\mathbf{x}_i), \sigma^2) \quad \text{where} \quad \phi_{\mathbf{w}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$$

כאשר $(\mathbf{w} | \mathbf{x}_i, y_i)$ מייצגת את הסיכוי שתתקבל התוצאה y_i עבור פיצ'ר x_i ו- \mathbf{w} כלשהוא. אנחנו מתנים זאת גם ב- \mathbf{w} (למרות שלא מדובר במשתנה מקרי), על מנת לייצג במפורש את ההתלות בתוווי המודל. במקרים אחרים, אנחנו מניחים כי כל דוגמה (y, \mathbf{x}) היא זאת כך שהתחולת של y היא ליניארית ביחס ל- \mathbf{x} . כיוון שאנו עוסקים במודל רגרסיה וב- \mathbb{R} התומך²¹ של המשתנה המקרי w , $y_i | \mathbf{x}_i \in \mathbb{R}$ הוא.

הבה ונשדרג את המודל מלמעלה עבור בעיות קליספיכיה. נניח כי y מותפלג **ברונולי** עם סיכוי $(\mathbf{x}_i) p$ שמתקשרת אליו ונתקשה \mathbf{x}_i ולכן נקבל:

$$p(y_i | \mathbf{x}_i) = \text{Ber}(y_i | p(\mathbf{x}_i))$$

מה הקשר בין $p(\mathbf{x}_i | \mathbf{x})$ לשונה ממודל הרגרסיה הליניארית, איננו יכולים להניח כי קיימת פונקציה ליניארית $\mathbf{w}^\top \mathbf{x} = p$ כאשר $(\mathbf{x}_i) p$ בהכרח מצומצם ל- $[0, 1]$. במקומות זאת, נבחר פונקציית קישור (link function) $\phi_{\mathbf{w}} : \mathbb{R} \rightarrow [0, 1]$ שהינה מונוטונית עולה ושממפה $(-\infty, \infty) \rightarrow (0, 1)$. נגידר זאת כך:

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \text{Ber}(y_i | \phi_{\mathbf{w}}(\mathbf{x}_i)), \quad \phi_{\mathbf{w}} := \text{sigm}(\mathbf{x}^\top \mathbf{w})$$

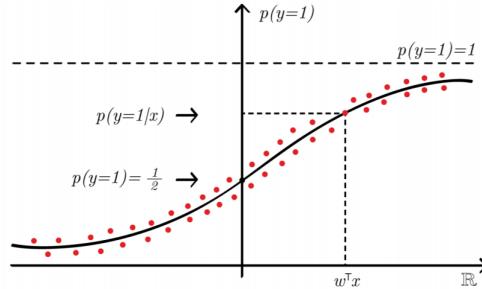
כאשר sigm היא פונקציית זיגמוד, שידוע גם בתחום פונקציית logit (ומכאן השם הפונקציה הלוגיסטיבית):

$$\text{sigm}(\mathbf{a}) := \frac{e^{\mathbf{a}}}{e^{\mathbf{a}} + 1}$$

²¹התוחום שהסתברות עליו גדולה מ-0

- פונקציה זו היא אכן מונוטונית עולה והיא מمفה $(-\infty, \infty) \rightarrow (-1, 0)$ כפי שאנו רוצים:
- אם $\mathbf{w}^\top \mathbf{x} \rightarrow -\infty$ אז $p(y_i = 1 | \mathbf{x}_i, \mathbf{w}) \rightarrow 0$ sigm $(\mathbf{x}^\top \mathbf{w})$ - כלומר הסיכוי שהי $y_i = 1$ מאוד רחוק.
 - אם $\mathbf{w}^\top \mathbf{x} \rightarrow \infty$ אז $p(y_i = 1 | \mathbf{x}_i, \mathbf{w}) \rightarrow 1$ sigm $(\mathbf{x}^\top \mathbf{w})$ - כלומר הסיכוי שהי $y_i = 1$ מאוד קרוב.

בתמונה, זה נראה כך:



4.1.1 מחלוקת ההיפතוזה

כעת, מחלוקת ההיפתוזות מקבלת ערכים בין 0 ל-1. כלומר:

$$\mathcal{H}_{\text{logistic}} := \{h_{\mathbf{w}}(\mathbf{x}) = \text{sigm}(\mathbf{x}^\top \mathbf{w}) \mid \mathbf{w} \in \mathbb{R}^{d+1}\}$$

4.1.2 עקרון הלמידה - עקרון הנראות המירביה (Maximum Likelihood)

כיוון שאנו משתמשים בהסתברות, הגיוני שנשתמש בעקרון הנראות המירביה. תהי $S = \{(\mathbf{x}_i, y_i)\}_{y=1}^m$ מבחן בלחתי תלי, נניח כי $y_i \sim \text{Ber}(\phi_{\mathbf{w}}(\mathbf{x}))$ כאשר ϕ היא פונקציה לוגיסטיית. אז, הנראות (Likelihood) של $\mathbf{w} \in \mathbb{R}^{d+1}$ הינה:

$$\begin{aligned} \mathcal{L}(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) &= \mathbb{P}(y_1, \dots, y_m \mid \mathbf{X}, \mathbf{w}) \\ &= \prod \mathbb{P}(y_i \mid \mathbf{x}_i, \mathbf{w}) \\ &= \prod_{i:y_i=1} \mathbb{P}(y_i \mid \mathbf{x}_i, \mathbf{w}) \cdot \prod_{i:y_i=0} \mathbb{P}(y_i \mid \mathbf{x}_i, \mathbf{w}) \\ &= \prod_{i:y_i=1} p_i(\mathbf{w}) \cdot \prod_{i:y_i=0} (1 - p_i(\mathbf{w})) \\ &= \prod p_i(\mathbf{w})^{y_i} (1 - p_i(\mathbf{w}))^{1-y_i} \end{aligned}$$

כאשר $\phi_{\mathbf{w}}(\mathbf{x}_i) = p_i(\mathbf{w})$. כיוון שפונקציית הלוג היא מונוטונית עולה, אנחנו יכולים למקסם את הלוג של הפונקציה הקודמת במקומות, כלומר $\ell(\mathbf{w}) := \log \mathcal{L}(\mathbf{w})$ ולכן:

$$\begin{aligned}\ell(\mathbf{w} | \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^m [y_i \log(p_i(\mathbf{w})) + (1 - y_i) \log(1 - p_i(\mathbf{w}))] \\ &= \sum_{i=1}^m \left[y_i \log\left(\frac{e^{\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{\mathbf{x}_i^\top \mathbf{w}}}\right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\mathbf{x}_i^\top \mathbf{w}}}\right) \right] \\ &= \sum_{i=1}^m \left[y_i \cdot \mathbf{x}_i^\top \mathbf{w} - \log\left(1 + e^{\mathbf{x}_i^\top \mathbf{w}}\right) \right]\end{aligned}$$

ולכן בחירת הפונקציה $h \in \mathcal{H}_{\text{logistic}}$ על ידי עקרון הנראות המירבית משמעה:

$$\hat{\mathbf{w}} := \underset{\mathbf{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{i=1}^m \left[y_i \cdot \mathbf{w}^\top \mathbf{x}_i - \log\left(1 + e^{\mathbf{w}^\top \mathbf{x}_i}\right) \right]$$

4.2 מימוש חישובי

מדובר למעשה בבעיית אופטימיזציה קעורה ולכן ניתן לפתור זאת. הרבה חבילות של machine-learning מומשאות זאת, ובפרט glment. חלק מהאלגוריתמים מבוססים על ניוטון ופסון.

4.3 פרשנות (Interpretability)

היתרון של הרגרסיה הלוגיסטי היא כי ניתן לאחר לשאול "מה התרחש בפייצ'רים" שגורם לשינוי זה ואיילו פייצ'רים תרמו תרומה משמעותית יותר. למשל, דגימות שקרובות לאפס משפיעות באופן דיבטן על המודם, בשונה מדוגמאות גדולות. תכונה זו היא תחוצה חשובה במסוגים.

4.4 מחלוקת ההיפותזה וכייזד חזים

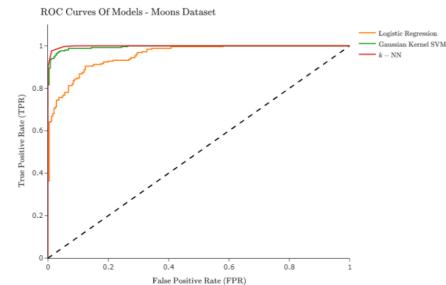
חלוקת ההיפותזה מוגדרת על ידי $\mathcal{H}_{\text{logistic}}^d = \{x \mapsto \pi(\langle x, w \rangle) \mid w \in \mathbb{R}^{d+1}\}$. למעשה דבר זה גורר כי היא מכילה פונקציות $\mathbb{R}^{d+1} \rightarrow [0, 1]$ ולא פונקציות $\{0, 1\} \rightarrow \mathbb{R}$. מצד אחד, יש יתרון במספרים בין $[0, 1]$, אך מצד שני צריך להגיע מתיישבו להחלטה. נגידר מונחים (cutoff) שמוגדר על ידי $\alpha \in [0, 1]$. מחלוקת הפרדיקציה שלנו תהיה:

$$\hat{y} := \begin{cases} 1 & h(x) > \alpha \\ 0 & h(x) \leq \alpha \end{cases}$$

כייזד נוכל לבחור את α ? נוכל לבחין כי אם α קרוב ל-0 או ל-1 הרף מאוד גבוה (כלומר, רוב הדוגמאות החדשות יסווגו כשלילות או חייבות בהתאם). דבר זה משפייע על ה-FP או ה-FN והינו רצוי "לשלוט" בהם.

אנו יודעים כי האלגוריתם מוחזיר לנו מספרים בין [0, 1]. נוכל לספר את ה-TP, ולמצוא את מדד ה-False-Poisitive rate. כיצד נעשה זאת? ניקח גוף שהולך מ(0, 0) ל-(1, 1). על מנת לקבל אינטואיציה, נבחן כי בנקודה (0, 0) הכל שלילי - כלומר אין 'חיוביים' נכונים' ואיו' 'חיוביים שגויים'. בcontra דומה אך הפוכה ב-(1, 1). עוקמה זו נקראת ROC Receiver Operating Characteristic ובקיצור ROC. ניתן להבחן כי אם מדובר בגרף ליניארי, המסובב די גרווע. אם ישנה 'קפיצה' מ-0 קרוב ל-1, מדובר במסובב די טוב (כי יש הרבה דגימות שבחרנו טוב, וקצת שבחרנו גרווע).

נוכל לראות זאת בדוגמה הבאה:



כל שהתיליות יותר גבוהה, אז מחיר ה-FP יותר נמוך.
באמצעות המודל הזה ניתן לבחור את α על מנת לקבל את המיזוג.

5 השכנים הקרובים ביותר Nearest Neighbors

5.1 חיזוי באמצעות k-NN

מדובר ב-learner ביליאר מחלוקת היפතויות ושלב אימון. הפרמטר היחיד שיש לנו הוא **פרמטר כוונון - tuning parameter** שמודגדר על ידי k כאשר $n \leq k \leq 1$. על מנת לבצע פרדיקציה עבור דוגמה $\mathbf{x} \in \mathbb{R}^d$, נמצא את k השכנים הכי קרובים שלו, ולפי "רוב" השכנים נבחר את הסיווג. ניתן לראות גם כאן את ממד השונות-הטיטיה - אם ניקח "שכן" אחד בלבד, השונות תהיה מאוד גבוהה (אם הנקודה תhapeox את הסימן, נשנה את הפרדיקציה). מאידך, אם ניקח אלף שכנים, השונות תהיה מאוד נמוכה. מה נחשב קרוב? אפשר להשתמש במרקח האוקלידי, או לתת משקל לכל אחד מהפיצרים:

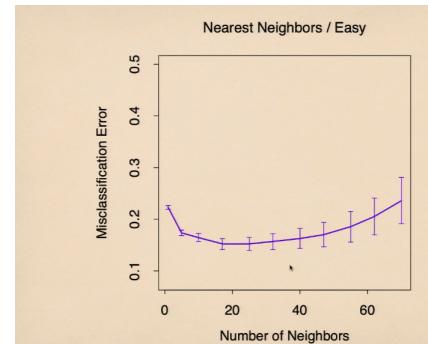
אלגוריתם 2 k-NN

1. קח דוגמה \mathbf{x} לחזות עבורה.
 - (א) תחשב את המרחק מכל אחד מהדוגמאות.
 - (ב) תסדר את המרחקים מהשכנים לפי הגודל.
 - (ג) תבחר את k הדוגמאות הקרובות ביותר ותחזה לפי הרוב.

בעקבות כך, נקודת המפתח היא כיצד לבחור את k .

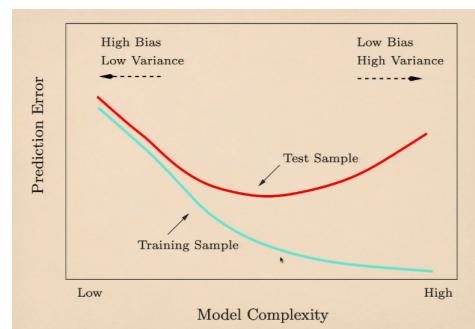
5.2 בחירת k

אם נבחר נקודה אחת, אויל לא ניתן לסמוך על המידע - אנחנו מאוד חשופים לכל השפעה של נקודה. למעשה, מדובר במסווג טוב ויציב שלא 'מתרגש' מנקודות. נוכל לראות שבקצוטות, ככל שה- k נמוך יותר או קטן יותר, ה- Error missclassification גובה יותר:



נחשף k אופטימלי, שהוא לא גבוה מדי ולא נמוך מדי, בדומה למה שראינו לגבי המודל - שייהי מרכיב אבל לא מדי.

ניתן לראות זאת גם בתמונה הבאה:

**5.3 חישובית**

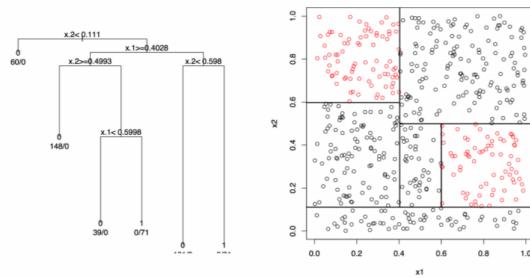
כיצד ניתן לפטור את המודל מבחינה חישובית? כיצד אפשר לחזות את המידע? אפשר באמצעות Brute force - כלומר, לחשב ולמיין את (\mathbf{x}, \mathbf{y}) לכל דוגמא מבחן \mathbf{x} . ישנים אלגוריתמים אחרים, שלא נלמד בקורס, שעובדים בצורה קשה בשלב המקדים, מייצרים מבנה נתונים שמאפשר חיפוש של שכנים קרובים.

6 עצי החלטה

המודל random forest הוא אחד המודלים הפופולריים ביותר. אחד מסוגיו מבצע רגרסיה, ואחד אחר מבצע קלאסיפיקציה.

העקרון של random forest הוא כזה: נרצה לחתות את המרחב האוקלידי ה- d - מימדי ולחבל אותו לפונקציות קבועות למוקוטעים.

ניחס את המרחב 'ונפצל' אותו לkopfsאות:



6.1 מחלקה ההפואתית

בהתינו חלוקה של עצים $\mathbb{R}^d = \bigcup_{i=1}^N B_i$, נאחסן תגית $\{0, 1\}$ לכל תיבת B_j , כך שכל אחד מהם מייצג פונקציה קבועה, כלומר, $h(\mathbf{x}) = \sum_{j=1}^N c_j \mathbf{1}_{B_j}(\mathbf{x})$. מחלוקת ההפואות היא מחלוקת כל הפונקציות הללו. אנחנו יכולים להגביל את גובה העץ ל- k ולהגדיר כי \mathcal{H}_{CT}^k - פונקציות עם לכל יותר k שלבים. כל פונקציה $h \in \mathcal{H}_{CT}^k$ שיכת לעץ החלטה כלשהו.

6.2 עקרון הלמידה

כיצד נוכל לבחור h מתוך מחלוקת ההפואות שלנו? לכaura נוכל להשתמש ב-ERM. מצד שני, ברור כי علينا להגביל את גובה העץ - אחרת, כל דוגמה תהיה בקבוצה בפני עצמה ונקבל overfit מטורף. אם נגביל את עצמו ל- k שלבים ונרצה למצוא את הפונקציה הממצערת בעז זה, מדובר בעיית NP קשה. לכן לא נשמש בעית אופטימיזציה אבל בחיפוש חמדני. אחד המודלים שקיים הוא Classification And Regression Trees (CART).

6.3 שתילת עצים

נתחילה מוקופה בודדה, ונחותוך בכל פעם לשתי קופסאות עד שנגיע ל- k צעדים. נחותוך כל קופסה לפי הקוארדינטה הטובה ביותר והערך הטוב ביותר לשים לה. עליינו ליסט את כל הקוארדינטות ולחפש את כל הערכים הטובים ביותר (באמצעות ERM). הסיבוכיות לא מרכיבת מדי. למה? בכל פעם עבורים על d פיצרים $-n$ דוגמאות ומחפשים את ה- t הטוב ביותר. לכן הסיבוכיות היא $O(dn)$. לכaura יש 2^k חלוקות, אבל כיוון שאנו לא מרשימים חלוקות ריקות, יש סך הכל n חלוקות אפשריות (אחד לכל מידע) ולכן נקבל סך הכל סיבוכיות של $O(dn^2)$.

חלק IV

תיאוריות PAC של למידה סטטיסטיות

1 הקדמה תיאורטיבית

אחד השאלות הבסיסיות בלמידה מוכנה היא: אלו מושימות הן נלמדות? כיצד אנו יכולים ללמידה מושימות אלו וכמה דגימות נצרכן על מנת ללמידה אותן?

בפרק זה נפתח את תיאוריית-h-PAC של למידה, שתאפשר לנו (בהתמוך על ההנחה וההגדרות) תשובה מלאה לשאלות אלו, עבר supervised learning.

מודל יצירת מידע

נבחן כי עליינו להניח שתי הנחות ברשות ה-PAC:

◻ ישנה פונקציה (דטרמיניסטיבית) f שהיא המסוגת (classifier). כלומר, לכל x יש תווייה אחת מתאימה, הנטונה על ידי $y = f(x)$.

◻ כל הדגימות, הן למרחב האימון שלו או בכל מרחב מבחר בעתייד, מותפלגים בצורה iid , כלומר, משתמשים בהסתברות \mathcal{D} מעל מרחב דגימות \mathcal{X} . למעשה, עולה מכאן כי ההסתברות $(S) \mathbb{P}$ של סדרת הדגימות $\mathbb{P}(S) = \prod_{i=1}^m \mathcal{D}(x_i)$ נתונה על ידי x_1, \dots, x_m

הבה ונשווה את ההנחות של מודל ה-PAC למודל הרגרסיה הליניארית שראינו קודם. בשני המקדים קיבלו סדרת דגימות x_m, \dots, x_1 . במקורה של רגרסיה ליניארית, אנחנו מניחים כי לכל הדגימות יש חשיבות שווה, למשל, עבור תרומותם לשגיאה הכלכלית. כתעת כל הדגימות הן iid וכאן יש להן הסתברויות שונות להופע ומילא הן בעלי משקלים שונים בפונקציית loss. הבדל נוסף הוא שבקרה של המודל הליניארי, התחלו עם ההנחה כי $(x_i) = f$, כאשר f היא דטרמיניסטיבית וליניארית. כתעת אנחנו מניח כי f היא דטרמיניסטיבית, אבל לא מדובר דוקא בפונקציה ליניארית.

במודל הלינאי כמעט חלשו את הנחת הדטרמיניסטיביות, ואנו מניחים כי y הוא פונקציה רדונומלית של x מהצורה $y_i = f(x_i) + z$. כתעת זאת מעיט יותר ונוריד את הנחה זו. בנוסף, על אף כי פרק זה מתעסק במסוגים, הרבה מהרעיון יכולם לבוא לידי ביטוי בפתרון בעיות רגרסיה.

הכללת שגיאה עבור מסוגים (Generalization Error for Classifiers)

עבור מושימות סיווג בלבדי, נגדיר את Generalization Error (נקרא לה הכללת השגיאה) עבור היפותזה h , כשהסתברות להציג x עבורו (x) h שונה מהתווית האמיתית (x):

$$L_{\mathcal{D},f}(h) \equiv \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \equiv \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq f(x)\})$$

כאשר \mathcal{D} ו- f הן בלתי ידועות. הכללת השגיאה ידועה גם בתור הסיכון או השגיאה האמיתית. נבחן כי דבר זה יכול להיות חשוב עבור ספירת בעיות מיסקלסיפיקציה. נזכיר בהבנה שעשינו בין שגיאות מסווג 1 ומסוג 2, כאשר לעיתים האחת גורעה באופן משמעותית מהאחרת. כתעת, נתיחס לבעית הכללה ביחס לבעית מיסקלסיפיקציה, ולכן לא נבחן בין שני הסוגים של השגיאות.

(The Fundamental Theorem of Statistical Learning)
 אם כך, המשימה שלנו היא לעצב אלגוריתם למידה'', שקיבל מרחב דגימות בגודל m ויצא כלל חיזוי (פרדיקציה) $\mathcal{Y} \rightarrow \mathcal{X} : h$. עברו בעית סיווג, למשל $\{\pm 1\} = \mathcal{Y}$, אבל נוכל להרחב זאת לתחום גדול הרבה יותר.

נניח כי נקודות המידע, הן בקבוצת האימון והן בקבוצת המבחן, מיצירות באופן בלתי תלוי על ידי הסטברות \mathcal{D} שאיננה ידועה לנו. התווויות (labels) y הן קבועות: בהינתן x ספציפי, ישנה פונקציה f דטרמיניסטיבית כך $y = f(x)$, כאשר f איננה ידועה לנו.
 לבסוף, הביצועים של כל החלטה (candidate rule) שהינו $\mathcal{Y} \rightarrow \mathcal{X} : h$ שהלומד (learner) שלנו ייצר, יושפעו על פי האיות שהוא ייצר עבור דגימות בלתי ידועות בעתיד. דבר זה נוכל להעריך באמצעות מד המיקסליספקציה. כתת נתמקד בהגדרות הבסיסיות של דבר זה.

הגדרה

מחלקת היפותזה \mathcal{H} נקראת מחלוקת היפותזה-ניתנת ללמידה PAC אם קיים אלגוריתם למידה \mathcal{A} ופונקציה $m_{\mathcal{H}, \mathcal{A}} : \mathbb{N} \rightarrow (0, 1)^2$ עם התכונות הבאות:

$$\square \text{ לכל } \varepsilon, \delta \in (0, 1)$$

\square לכל התפלגות \mathcal{D} מעל \mathcal{X}

\square לכל פונקציית תיוג (labeling function) $L_{\mathcal{D}, f} (h^*) = 0$ כך שקיים $h^* \in \mathcal{H}$ שמקיימת $h^* \in \mathcal{H} \rightarrow \{\pm 1\}$ (labeling function) על (ε, δ) כך שמייצירות באופן iid על ידי \mathcal{D} ומוגדים על כאשר נרים את אלגוריתם הלמידה \mathcal{A} . על $m_{\mathcal{H}, \mathcal{A}}(\varepsilon, \delta, m) \geq m$, דגימות שמייצירות באופן iid על ידי \mathcal{D} ומוגדים על ידי f , האלגוריתם יחזיר היפותזה $h_S = \mathcal{A}(S)$ כך שעם הסטברות של לפחות $\delta - 1$ קיבל כי $L_{\mathcal{D}, f} (h_S) \leq \varepsilon$ וגם נקבע כי:

$$\mathcal{D}^m (S | L_{\mathcal{D}, f} (h_S) \leq \varepsilon) \geq 1 - \delta$$

נסמן את גודל מרחב המדגם המינימלי הנדרש בשביל תנאים אלו, ביחס ל- δ, ε ואלגוריתם כלשהו, כ:

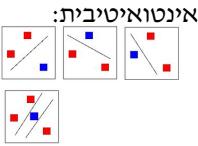
$$m_{\mathcal{H}}(\varepsilon, \delta) = \min_{\mathcal{A}} m_{\mathcal{H}, \mathcal{A}}(\varepsilon, \delta)$$

הפונקציה $\mathbb{N} \rightarrow m_{\mathcal{H}}$ נקראת סיבוכיות המדגם של מחלוקת ההיתפוזה-הניתנת ללמידה PAC.

הגדרה

תהי $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ מחלוקת היפותזה. עבור תת קבוצה $C \subset \mathcal{X}$ יהיו \mathcal{H}_C הelts של \mathcal{H} ל- C .
 כלומר, $\mathcal{H}_C = \{h_C : h \in \mathcal{H}\}$, כאשר $f : \mathcal{X} \rightarrow \mathcal{Y}$, $h_C : C \rightarrow \mathcal{Y}$, $h_C(x) = h(x)$ היא הפונקציה כך ש- $x \in C$ לכל $x \in X$.
 נגידר את מימד VC-dimension (VC-dimension) של \mathcal{H} בתורה:

$$VC\dim(\mathcal{H}) = \max \left\{ |C| \mid C \subset \mathcal{X} \text{ and } |\mathcal{H}_C| = 2^{|C|} \right\}$$



בדוגמה כאן (מתוך ויקיפדיה), עולה כי ניתן לבצע הפרדה עבור 3 נקודות בלבד באמצעות מفرد ליניארי, עבור 4 נקודות לא ניתן לבצע הפרדה. לכן ממד ה-VC כאן הוא 3. על ידי בחירת מידת PAC בתור הפירוש שלנו ללמידה, ההדרות דלעיל מספקות לנו תנאי הכרחי ומספיק מתי למידה אפשרית, ומהו גודל מרכיב האימון המיניימי הנדרש. מסגרת זו גם מאפשרת לנו "לומד" אוניברסלי שלומד בצורה מוצלחת בהינתן מרחיב אימון גדול מספיק. במילים אחרות, יש לנו מידע מספיק מתי נוכל להכליל מקבוצות אימון לדגימות חדשות. תוצאה זו ידועה בתור "המשפט היסודי של הלמידה הסטטיסטי" שוטען כי:

1. מחלקת היפותזה \mathcal{H} היא נלמדת-*PAC* אם ורק אם $\text{VC dim} \{\mathcal{H}\}$ הוא סופי.

2. סיבוכיות המדגם של מחלקת היפותזה עם מימד VC סופי, נתון בערך על ידי:

$$m_{\mathcal{H}}(\varepsilon, \delta) \sim \frac{V \text{CDim}(\mathcal{H}) + \log(1/\delta)}{\varepsilon}$$

3. כלל ה-ERM מושג את המינימום, כאשר למידה אפשרית.
הצעד הראשון שבהנחת מושגים אלו יוביל אותו לזווית ראייה שונה של מסגרת זו.

1.1 למידה כמשחק - ניסיון ראשון

ניתן להסתכל על המסתגרת שהציגנו קודם כמשחק ביןינו לבין הטבע, עם תשלום רנדומלי. המשחק מועבר כדלהלן. קודם כל, מספר הדגימות m מתקבל כפרמטר המשחק. לאחר מכן, נבחר "לומד" A שמאמן m דגימות. זאת האסטרטגיה שלנו. לאחר מכן, הטבע בוחר פונקציית הסתברות \mathcal{D} ופונקציית "תיגוג" f - זאת האסטרטגיה של הטבע. חשוב לשים לב כי הטבע יודע מה האסטרטגיה שלנו בוחר את האסטרטגיה שלו. על מנת לחשב את תוצאות המשחק, אוסף דגימות המתפלגות iid בגודל m נלקחות בהתאם להתפלגות \mathcal{D} , ומתיוגות לפי הפונקציה f , שניהם נבחרו על ידי הטבע.

קובוצה זו נכנסת לאחר מכן לתוך הלומד A שבחרנו על מנת לקיים את כלל החיזוי (S) $h_s = \mathcal{A}(S)$. הנוטציה h_S כולה כי כלל החיזוי שבחרנו תלוי באופן משמעותי במרקם המדגם הרנדומלי שנבחר. התשלומים הינו $L_{\mathcal{D}, f}(h_S)$. המטריה שלנו במשחק היא לסיים עם $L_{\mathcal{D}, f}(h_s)$ מינימלי ככל שניתנו, בעוד הטבע רוצה שהוא יהיה הגודל ביותר. כיוון שדבר זה תלוי בمدגם שנוצר רנדומלית, אם אנו חסרים מיל והمدגם S לא משווה, קלומר לא מייצג את \mathcal{D} בצורה טובה, אז ה-"לומד" שלנו לא יוכל בצורה מספק טוב, וההפסד שלנו יהיה גדול. אם כך, מה האסטרטגיה הטובה ביותר עבורנו? כיצד נבחר את A , הטוב ביותר? הזכרו כי הטבע יוציא מה נבחר, ולכן ינסה לבחור \mathcal{D} ו- f שה-"לומד" שלנו לא התכוון אליהם כמו שצריך. כמו כן, תמיד יש סיכוי שנבחר מדגם לא מספק מייצג. מה הדרך שלנו להתמודד מול בעיות אלו?

לסיכום, נראה את ההגדלה למשחק זה.

הגדירה - משחק הלמידה (גרסה ראשונה)

□ מרחב מדגם m הוא קבוע.

□ נבחר אסטרטגיה ("לומד") A . מדובר בפונקציה שמתאימה כל דוגמה S לכל חיוי $\gamma \in \mathcal{X} \rightarrow \mathcal{A}$.

□ הطبع ידע את בחירתנו, ובוחר אסטרטגיה שכוללת התפלגות \mathcal{D} מעל \mathcal{X} ופונקציית תיוג $\gamma \in \mathcal{X} \rightarrow \mathcal{A}$.

□ קבוצת דוגמות S בגודל m מיוצרת לפי \mathcal{D} ומוגנת לפי f .

□ הדוגימות ננסות בתוך \mathcal{A} על מנת ליצר כלל חיוי $h_S = A(S)$.

□ התשלום, $L_{\mathcal{D}}$, שלמעשה קובע את מספר שגיאות המיסקלטייפיקציה.

□ הطبع עושה כמעט יכולתו על מנת לנצל. נרצה להבטיח כי פונקציית ההפסד ($L_{\mathcal{D},f}$) (h_S) תהיה מינימלית.

1.2 לומדים "בערך נכונים" ו"קרוב לוודאי נכונים" (Learners)

כפי שראינו, פונקציית ההפסד $L_{\mathcal{D},f}$ (h_S) היא רנדומלית, כיוון שהיא תלולה בדוגימות שנוצרות רנדומלית. לכן, ניתן לדבר על על הסתברות שפונקציית ההפסד תקבל ערך ספציפי. כמו כן, ניתן להתייחס להז במנוחים אחרים ולומר כי הסיכוי שפונקציית ההפסד תהיה קטנה מערך מסוים, הוא 1. כך מתאפשרת ההגדירה הבאה:

הגדרה

יהי $\varepsilon \in (0,1)$. נאמר כי "לומד" A הוא בערך נכון, עם דיוק ε אם לכל מרחב אימון S שנוצר באמצעות \mathcal{D} , ה"לומד" A יוצר כלל חיוי h_S עם הפסד קטן או שווה ל- ε :

$$\mathcal{D}(S | L_{\mathcal{D},f}(h_S) \leq \varepsilon) = 1$$

האם קיים $\varepsilon \in (0,1)$ כך שנוכל למצוא "לומד" שיהיה קרוב לוודאי נכון? לצערנו - לא. הطبع תמיד יוכל לבחור פונקציית הסתברות שלרבות מרחבי המדגם לא תיצג את \mathcal{D} כלל, ותגרום לכך שפונקציית ההפסד תהיה קרובה מאוד ל-1, וגודלה מכל ε שנבחר. נוכל להבין זאת על ידי הדוגמה הבאה.

דוגמה

יהי $\varepsilon \in (0,1)$ וניקח שתי נקודות x' , x . נסמן את S' בטור מרחב אימון שאינו מכיל את x כלל. ככלומר $(x_1, y_1), \dots, (x_m, y_m)$ כאשר $x'_i = f(x_i) - x$. למרות שמרחב מדגם זהה הינו נדיין, עדין דבר זה יכול להתקיים ולכן האלגוריתם שלנו A יאמר לנו מהי התגובה שהתקבלה, וגם $h_{S'}(x)$ יראה את שאר הנקודות חוץ מ- x' ובפרט יראה את x , במקרה ש- S' מתקבלת בקטלט. במקרה בלי הגבלת הכלליות כי בחרנו את $x = +1$ ו- $x' = -1$.

כלומר כי אם היא S' הוא מרחב האימון שלנו, A יראה $+1$ על x . בשלב זה, האסטרטגיה של הطبع מכילה שני חלקים. ראשית, הוא מצמצם את \mathcal{A} למרחב דו-מימדי, על ידי בחירת פונקציית הסתברות \mathcal{D} שמעיפה כל דבר חוץ מ- x ו- x' . מעבר לכך, הוא בוחר \mathcal{D} כך ש- $\gamma(x) = 1$ ו- $\gamma(x') = 0$, כאשר γ מקיימת כי $1 < \gamma < \varepsilon$.

שנית, הطبع בוחר פונקציית תיוג f כך ש- $f(x) = -1 = h_{S'}(x)$. נוכל לבדוק כי הסתברות לקבל את S' אינה אפס, שהרי הסתברות $0 < \varepsilon^m < 1$. לכן, על מנת להראות כי האסטרטגיה של הطبع מנوع מ- A להיות "קרוב

לוודאי נכון" עם דיק ε , כל שנטורך הוא להראות כי הפרש (או הפסד - loss) במקרה שנבדוק את הקבוצה S' הוא גדול מ- ε . וכן, יוכל לראות כי מתקיים:

$$L_{\mathcal{D},f}(h_{S'}) = \mathcal{D}(\mathbf{x}) \{\mathbb{1}\} [(\mathbf{x}) \neq h_{S'}(\mathbf{x})] + (\mathbf{x}) \{\mathbb{1}\} [(\mathbf{x}') \neq h_{S'}(\mathbf{x}')]$$

כאשר $\mathbb{1}[a \neq b]$ זה הפונקציה המczyינית שמקבלת 1 כאשר $a \neq b$ ו-0 אחרת. כיוון שני הביטויים הימניים הם אי שליליים, נקבל:

$$\geq \mathcal{D}(\mathbf{x}) \mathbb{1}[f(\mathbf{x}) \neq h_{S'}(\mathbf{x})] = \gamma > \varepsilon L_{\mathcal{D},f}(h_{S'})$$

לכן, לכל $\varepsilon \in (0, 1)$ ולכל אסטרטגיה \mathcal{A} שנשחק, לטעו יש אסטרטגיה f, \mathcal{D} כך שישנה הסתברות אי שלילית מעל בחירות מרחב הדגימות m , כך שנקבל $\varepsilon > L_{\mathcal{D},f}$. כיוון ש- ε היא שרירוטי (arbitrary בעז), עולה מכך כי הטעו יכול בהסתברות לא מבוטלת לגרום למשחק להסתטים בהפסד קרוב ל-1.

בדוגמה לעיל כי גם אם בהסתברות נמוכה של $(\gamma - 1)^m$, נקבל מרחב מודגם "גרוע", S שאנו מייצג את \mathcal{D} מספק טוב ושלא מאפשר לנו להכליל את המדגמים. מכאן נוכל לסכם כי בהינתן פרמטר דיק $\varepsilon \in (0, 1)$, הטבע תמיד יוכל למצוא אסטרטגיה אין "לומד" \mathcal{A} שיכל להבטיח כי בהסתברות 1, ההפסד/הפרש לא עבר את ε . הטבע תמיד יוכל למצוא אסטרטגיה כך ש- ε -> $L_{\mathcal{D},f}(h_S)$ יקבל הסתברות שונה מפאס.

אם כך, נוכל להסיק מהאפשרויות של דגימות גרועות כי איןנו יכולות לשאוף לוודאות מוחלטת בהשגת הפסד מינימלי. לכן נדרש רק וודאות מוגבלת, ש惕חון, שתיקרא **ביטחון** (confidence) להשגת הפסד מוגבל. ככלומר, נרצה כי ההסתברות לקבלת מודגם גרוע לא תעבור איזשהו סף $\delta \in (0, 1)$. הפרמטר זהה גם יתקבל עם תחילת המשחק, יחד עם ε . כיוון שאנו דורשים ביטחון מוחלט, האם יתכן שנבחר דיק מוחלט (הפסד 0), אך עם ביטחון מוגבל?

הגדרה

תהי $\delta \in (0, 1)$. נאמר כי "לומד" \mathcal{A} הוא **קרוב לוודאי נכון** (Probably Correct) עם מקדם ביטחון δ אם הסיכוי לקבל מודגם S , ש- \mathcal{A} תיצור עבורו כלל חייזי h_S עם דיק מושלם (הפסד שווה ל-0) הוא גדול או שווה $1 - \delta$.
כלומר:

$$\mathcal{D}^m(S | L_{\mathcal{D},f}(h_S) = 0) \geq 1 - \delta$$

משמעותו לב להבדלים בין ההגדירות, בהגדירה הקודמת דרשנו ביטחון מוחלט לגבי ערך ספציפי, ובהגדרה כאן דרשנו דיק מושלם עם ביטחון לא מוחלט). למעשה, המושג ביטחון מייצג חוסר ביטחון. האם ישנו מקדם ביטחון $\delta \in (0, 1)$ שנוכל למצוא עבורו "לומד" כלשהוא? ברור שההתשובה אינה לא. לכל δ , הטבע תמיד יוכל לבחור אסטרטגיה עם הסתברות גדולה מ- $\delta - 1$ כך ש- $0 > L_{\mathcal{D},f}(h_S)$ מודיע? נזכיר כי הפסד-אפס, משמעותו כי בהסתברות 1 (ביחס ל- \mathcal{D}), התגית החוויה היא תמיד נכונה. הטבע תמיד יכול לשחק את \mathcal{D} עם הסתברות זעירה לקבלת $\mathcal{A} \in \mathcal{X}$. אם כך, **בהתברות גדולה**, מרחב המדגמים לא יכול את

א ולכן, לא משנה מה התగית ש- x קיבל, מדובר בבדיקה ולכן יכול להיות שתתקבל טעות והפסד סופי.שוב, נוכיח דבר זה באמצעות דוגמה של אסטרטגיה שהطبع יכול ליצר.

דוגמה

יהי $\delta \in (0, 1)$ ונניח כי ישנו שני נקודות $x' \in x$. בדומה לדוגמה הקודמת, נניח כי מרחב האימון הינו $(x_1, y_1), \dots, (x_m, y_m)$ עם $x'_i = x_i$ ו- $y'_i = f(x')$ לכל $1 \leq i \leq m$. כמו כן, נניח כי כאשר נכניס את הדוגמה לתוך האלגוריתם, נזהה כי $h_{S'}(x) = +1$ כי שראינו קודם, האסטרטגיה של הطبع הינה לבחור פונקציה f כך ש- $f(x) = -1 = h_S(x)$ ופונקציית הסתברות D כך ש- $D(\gamma) = 1 - \gamma$. במקרה שלנו, γ נבחרת כך ש- $\gamma^m < \delta$. מכך עולה כי ככל m -היותר, ההסתברות לקבל את x (כלומר γ) היא מאוד קטנה. אם כך, במקרה שלנו S' היא מוגדרת כך שההסתברות שללה להופיע, $(\gamma^m - 1) < \delta$. בדומה לחישוב בדוגמה הקודמת, נקבל כי $\gamma \geq L_{D,f}(h_S) > \gamma$, כלומר הפסד שונה מ-0. כיוון שהסיכוי לקבל S' הוא גדול מ- δ , נקבל כי הסיכוי לקבל הפסד שונה מ-0. לעומת זאת, ה"לומד" שלנו אינו "קרוב לוודאי נכוויס".

אם כך, נוכל לסכם כי לכל מועד ביטחון $\delta \in \delta$, אין שום "לומד" A שיכל להבטיח כי בהסתברות של לפחות $1 - \delta$, הפסד ייעלם: הטענה תמיד יכולה למצאו אסטרטגיה כך ש- $L_{D,f}(h_S) < \delta$.

הדוגמאות לעיל מלמדות אותנו כי לא משנה איך "לומד" נבחר, לעיתים לא נצליח וודאות מוחלטת שההפסד לשונו יצומצם, גם אם נבחר ביטחון מצומצם לקבעת דיוק מסוים. מה שכן נוכל לעשות הוא לקבל ביטחון מסוים כך שההפסד שלנו לא יעבור סף כלשהו. דבר זה מביא אותנו להגדירה הבאה:

הגדרה

תהיה $\delta \in (0, 1)$. נאמר כי "לומד" A הוא קרוב לוודאי בערך נכוויס (Probably Approximately Correct) עם מועד ביטחון δ ומועד דיוק ϵ אם ורק אם הסיכוי לייצר מרחב אימון S שעבورو A תיציר כלל חייזר h_S עם הפסד שאינו עובר את ϵ , גדול או שווה ל- $\delta - \epsilon$:

$$\mathcal{D}^m(S | L_{D,f}(h_S) \leq \epsilon) \geq 1 - \delta$$

חשוב להבין את ההבדל בין הבדיקה ובין הביטחון:

◻ נזכר צרנו את דוגמאות האימון S בצורה רנדומלית. במקרה זה, האלגוריתם פועל בצורה רנדומלית ולכן החיזוי הוא רנדומי. אם S הינו מזוזר (לא מייצג את D בצורה טובה), אז כלל החיזוי h_S יהיה שונה, ככלומר לא יכול היה. המספר δ מייצג את ההסתברות לטעות בעקבות דוגמאות אימון "מוזירות".

◻ אחרי שה"לומד" מיצא כלל החלטה h_S , הוא בודק זאת על פי דוגמה חדשה. הדוגמה החדשה גם היא מתתקבלת בצורה רנדומלית. $L_{D,f}(h_S)$ היא כמות השגיאות היחסית ש- h_S תעשה. ככלומר, הבדיקה על המידע. המספר ϵ מתייחס לדיוק זה.

1.3 **למידה כמשחק - ניסיון שני**

כיוון שאנו מנסים לeyer "לומד" קרוב לוודאי בערך נכון, נעדכן גם את הגדרת המשחק. גודל המשחק m כבר לא יהיה קבוע. במקומות, נקבע את פרמטר הדיקט ε ומקדם הביטחון δ , שה"לומד" שלנו נדרש להציג, כערך המשחק.

נצרך להחליט על **m** כחלק מהסטרטגייה שלנו. שימו לב שכעת ישנו ניוואנס עדין בסימונים שלנו.

כאשר נכתב את A שהינו הוליך, נדבר למשמעותו של סדרה של לומדים, שלכל אחד יש מודגש בגודל m . לכן נוח יותר לכתוב זאת כך: $\mathcal{X}^m \rightarrow (\mathcal{Y} \times \mathcal{A}_m)$.

נגיש ואלטנטואה צואנה פלונית גט ב-2

הגדירה - משחק הלמידה (גרסה שנייה)

נקבע את הpermטרים $\varepsilon, \delta \in (0, 1)$ ו אז:

מ נבהיר יותר מדגם m ולומד A . ששתיים תלויות ב- (δ, ε) .

הטבּ יודע את האסטרטגיה שלנו ובעקבות כך בוחר אסטרטגיה שכוללת הסתרות D מעל \mathcal{X} ופונקציית $T_{\text{יג}}(\cdot, \cdot) : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$. האסטרטגיה של הטבּ עלולה להיות תלויה גם ב- (δ, ε) וגם בבחירה של m ו- \mathcal{A} .

מבחן S מוגדר m מיוצר לפי \mathcal{D} ומתייג לפי f .

מוכניס את S לתוך A וניצר כלל חיזוי h_S .

כעת התוצאה הסופית (payoff) הוא $L_{\mathcal{P}, f}$. מדובר בתוצאה הנדמית כיון ש- S הינו רנדומלי וממילא גם h_S .

הטבע עושה כמיibe יכולתו על מנת לנצל. על מנת לעשות זאת, הוא יחש A כך שהסתברות של לפחות $1 - \delta$ הפסד ($L_{D,f}$) לא יעבר את ε . לא משנה איך אסטרטגיה D , f הטבע יבחר לשחק.

על מנת לוודא שニיצחנו במשחק, נshallק בו המון המון פעמיים. בכל פעם גם אנחנו וגם הטיבע השחק את האסטרטגיה שלנו. אולם, בכל פעם הדגימות הנוצרות הינן שונות. נחשב את ההסתברות (מעל התוצאות הרנדומיות של מוגדים האימון) של המאווער $\varepsilon \leq \{S \sim \mathcal{D}^m \mid L_{\mathcal{D},f}(hs) \leq \delta\}$. אם ההסתברות שנמצאה גדולה מ- $\delta - 1$ אז הלומד A שנבחר הינו קרוב לוודאי בערך נכoon, עם דיוק ε ומקדם ביטחון δ - נגד האסטרטטגיה של הטיבע. יוכל להגיד במדויק זה כי ניצחנו.

זווית המשחק, למרות שימושו בקורס שונה, שcolaה לגרסה סטנדרטית יותר של אתגר הלמידה שלנו: ה"לומד" אינו מכיר את \mathcal{D} ואת f . הלומד מקבל מוקדם דיווק ϵ ומוקדם ביטחון δ . לאחר מכן, הוא מבקש נתוני אימון S (training data) שמכילים דוגמאות (δ, ϵ) (m מספר הדוגמאות יכול להיות תלוי בערך של ϵ ו- δ אך לא בא- \mathcal{D} -ו- f) ש אינם ידועים לנו). לבסוף, הלומד צריך ליצור היפותזה h_S שתלויה רק ב- δ, ϵ ובمدגם האימון S שנוצר, כך שהסתברות של לפחות $1 - \epsilon$ נקבע כי h_S מתקיימת. כלומר, ה"לומד" צריך להיות קרוב לוודאי בערך נכון עם מקדם דיווק ϵ ומוקדם ביטחון δ .

כיוון שכתעת אנחנו בוחרים את גודל המדגם m וכיוון שמידע עולה זמן וכסף, נעדיף כי m יהיה קטן ככל שניתן, ולמעשה נקבל כת טרייד-אוף בין δ, ϵ ובין בחירת m .

2 אין ארכות חינס ומחלות היפותזה

עם כל הבאה, אין לנו דרך כלל לנצל את הטבע, אפילו בגרסתו השנייה של המשחק. בלי הגבלות על בחירת D או f אין לנו דרך להיות בטוחים מסווגים שמספר S_h מדויקת מספיק. דבר זה נכון, לא משנה כמה m יהיה גדול.

בדוגמאות לעיל, הטבע בחר \mathcal{D} שמעלימה הכל במרחב המודגם \mathcal{X} חז' משתי הנקודות. בפרט, האסטרטגיה של הטבע הספיקה על מנת למנוע מאייתנו בנייה של לומד בעל מקדם ביטחון $= \delta$ או $0 = \varepsilon$ לכל \mathcal{X} עם 2 נקודות או יותר. בדוגמה הבאה, האסטרטגיה של הטבע תנצח רק אם לא- \mathcal{X} יש מספר אינסופי של נקודות (למרות שמספרם שווה למספר בן מניה שלהם).

דוגמה

נניח כי $\infty = |\mathcal{X}|$ ונעקב אחרי הצעדים של הגרסה השנייה של המשחק עברו $(0, 1) \in \delta < \varepsilon < 0$ כלשהו.

□ נקבע את m .

□ נחליט מה התגית שנזהה עברו נקודה שלא מופיעה במרחב המודגם S . נוכל לדוגמה, להחליט כי דוגמה כלשהיא, נניח x_4 שלא מופיעה ב- S אבל x_3, x_2, x_1, x כן מופיעות, כולל עם תגית +1.icut, נבחר את $x_4 = h_S(x_4)$. אך אם כל הנקודות האחרות יהיו עם תגית -1, אז נקבל כי $1 = h_S(x_4)$. ככלומר, נוכל לראות כי הערך שנבחר עברו x_4 תלואה למעשיה ב- S . לעומת זאת, נוכל להגביל את עצמנו לקבלת **אotto Urz**, h_S , **כל עוד S לא מכילה את x_4** , ללא תלות ב- S שנבחר. במילים אחרות, נבחר פונקציה $(x) = g$ ואם נקודה $\mathcal{X} \in x$ איננה מופיעה ב- S , אז בנקודה זו הכלל $(x) = h_S(x) = A$ מיצא, יקיים כי $(x) = g$, $h_S(x) = A$ ללא תלות ב- S שנבחר.

□ הטבע מכיר את m , אז הוא לוקח תת קבוצה $\mathcal{X} \subset C \subset |C| > 2m$ עם $\mathcal{D}(x) = \frac{1}{2m} < \frac{1}{|C|}$ כשבתוכה ההסתברות הינה איחוד: לכל $C \in \mathcal{X}$ מתקיים כי $\frac{1}{2m} \leq \mathcal{D}(x)$.

□ הטבע גם מכיר את $(x) = g$, ובקבות כך, כצד מלולך למדיו הוא בוחר פונקציית תיוג f שמקיימת כי $(x) = -g$ לכל $\mathcal{X} \in x$ - ככלומר בדיקת ההיפוך ממה $-h_S$ תראה עבור נקודות שטרם נחזרו.

□ כעת, יהיו S מרחב אימון כלשהו. יהיה $C \subset \text{supp}(S)$ (הקבוצה של כל איבר בה גודלה מ-0) קבוצות הנקודות השונות $\mathcal{X} \in x$ ש모ופיעות ב- S . נזכיר כי אותן x יכול להופיע ב- S מספר פעמים ולכן $|\text{supp}(S)|$ יכול לקבלת כל ערך בין 1 ל- m . מכיוון $sh \leq |\text{supp}(S)|$ ומכיון ש- \mathcal{D} היא איחוד מעל C , אז נקבל כי $\geq 1/2$ $\in \mathcal{X} \setminus \text{supp}(S)$. ככלומר, ההסתברות שדוגמה חדשה כלשהיא לא מופיעה ב- S הינה לפחות $\frac{1}{2}$.

□ לפי כללי המשחק, נכנס את A לתוך S ונקבל כי $A(S) = h_S$. מה ההפסד כעת? בלי קשר למה ש- h_S יראה, לנקודת המבחן יש הסתברות שגדולה מ- $\frac{1}{2}$ להיות ב- $(S) \setminus \mathcal{X}$ ולכן בהסתברות שגדולה מ- $\frac{1}{2}$, נקבל כי $(x) = g$ ונקבל עם כל הבאה כי $(x) = -g$. לכן נקבל כי $L_{\mathcal{D}, f}(h_S) \geq \frac{1}{2}$.

□ דבר זה קורה לכל מודגם אימון S . אם כך, האסטרטגיה של הטבע מבטיחה כי בהסתברות 1, תוצאות המשחק יקיים כי $\frac{1}{2} \geq L_{\mathcal{D}, f}(h_S)$.

□ אם כך, לא נוכל למצוא לומד \mathcal{A} כך שייהיה קרוב לוודאי בהכרח נכון, ללא תלות באסטרטגיה של הטבע. בחירת מודגם גדול יותר לא תעזר לנו. אם נגדיל את m , הטבע יבחר C גדול יותר והסתברות \mathcal{D} שהינה איחוד מעל C .

איפה היכנו לאיוב? הטבע יכול לבחור **כל** פונקציית תיוג שירצה, ואילו אנחנו מנסים ללמד את f מתוך מודגם שהינו קטן באופן יחסי למספר הפונקציות האפשריות. דבר זה ידוע בתורת טענת "אין ארוחות חינם" ("No Free Lunch" Theorem). - בלי להניח שהוא על פונקציית התיוג f , במידה היא בלתי אפשרית. בצורה שוקלה, נאמר כי אם מספר פונקציות התיוג גדול מדי, הטבע יכול ליצור פונקציה שאיננו יכולים ללמד.

הדוגמא לעיל מראה את משפט זה, אבל היא לא הוכחה, כי הגבלנו את עצמנו ל- $(x)g$ שאינה תלולה ב- S . הגבלה זו הפכה את החיכים של הטבע לקלים יותר כי היא אפשרה לו לבחור $(x)g = -g$ שמקסמת את השגיאות של הלומד שלנו. במקרה, נוכל לבחור $(x)g_S$ לכל אחד מה- S השונים שלא מכילים את S , ובכך נקשה על הטבע. אמנם, דוגמה זו מסבירה מעט את האינטואיציה.

כפי שהזכרנו לעיל, ישנו מספר טענות שיכولات להיות מכוונות בתור "אין ארכות חינוך" - למעשה כל טענה שטראה כי בלי מידע מוקדים על פונקציית התיאוג, ככלומר שיש יותר מדי אפשרויות $-f$, למידה היא בלתי אפשרית. למטה נראה את טענת "אין ארכות חינוך" הוכחה, שימושה ברעיון של PAC-Shenfeld בהמשך, נמצא בספר Understanding Machine Learning.

טענה (אין ארכות חינוך)

יהי \mathcal{A} מרחב מדגם. $\infty = |\mathcal{A}|$. לכל $\frac{1}{2} < \varepsilon < 0$ ישנו $0 > \delta$ כך שלכל אלגוריתם A ישנה פונקציית הסתברות D מעל \mathcal{A} ופונקציה $h \rightarrow \mathcal{A}$: f שכשר נרץ את A על S מגודל סופי שנוצרת iid מ- D , אז בהסתברות של לפחות δ לפלאט של A , h_S , יש הפסד גדול יותר מ- ε . ככלומר $\varepsilon \geq L_{D,f}(h_S) \geq \varepsilon$.

מחלקות היפותזה נדרשות (Needing Hypothesis Classes)

בהתמך על טענת "אין ארכות חינוך", הנה חיבים להניח כי f מתאפשר מחלוקת היפותזה $\mathcal{H} \subset \mathcal{H}$ או לפחות היא קרובה למספר פונקציות במחלוקת זו. מחלוקת זו יכולה להיות גדולה מדי, אחרת נגיעה לבעה שראינו קודם לכן.

הנחה הרילזבליות

אם אנחנו מניחים כי מחלוקת היפותזה מסוימת משוויכת למשחק שלנו, דבר זה אומר שאנו מגבילים את הכללי הבחירה שהאלגוריתם שלנו יכול לבחור למשפה מסוימת של פונקציות. **הנחה הרילזבלי** הוא כאשר הטבע חייב לשחק עם פונקציה $\mathcal{H} \in f$. למעשה, לא באמת אכפת לנו אם הטבע משחק עם פונקציה f שלא- \mathcal{H} אך יש פונקציה אחרת $\mathcal{H} \in h^*$ שהינה זהה ל- f בכל הנקודות ש- D לא מעיפה. ככלומר, לעומת לא נקבע כי $(x)h^* \neq f$ ש- h^* פורמלית, נאמר כי הנחת הרילזבליות הינה: הטבע משחק עם פונקציה f כך שישנה פונקציה $\mathcal{H} \in h^*$ כך ש- $L_{D,f}(h^*) = 0$.

הلومד מקבל \mathcal{H} לפני שהלמידה מתחילה ויוצא רק $\mathcal{H} \in h_S$. בambil אחריות: עברו דוגמאות אימון מגודל m , אלגוריתם הלמידה הוא העתקה $\mathcal{H} \rightarrow h \in \mathcal{H} \times \mathcal{A}$ כך ש- $\mathcal{H} \in h \rightarrow S$ גם כאן נמשיך בנטציה המוצבנת לורות שמדובר בסדרה של אלגוריתמי למידה).

ראינו קודם לכן כי אם $\infty = |\mathcal{A}|$ אין \mathcal{H} גדול מדי ללמידה. אם כך, נוכל לשאול:

◻ מהי מחלוקת ההיפותזה \mathcal{H} שהינה "קטנה מספיק" שנוכל למצוא עבורה לומד קרוב לוודאי בערך נכו? מה נחשב כמחלקה היפותזה גדולה מדי?

◻ נניח שבחרנו מחלוקת היפותזה קטנה מספיק. ונניח שלכל δ, ε יש לפחות אסטרטגיה אחת (m, A) שמקיימת את מה שאנו רוצים. אם כך, נוכל לשים את מספר דוגמאות האימון המינימליות בתוור $(\delta, \varepsilon)_m$. האם נוכל למצוא את הפונקציה המינימלית $(\delta, \varepsilon)_m$? האם יש קשר בין הגודל של מחלוקת היפותזה \mathcal{H} ו- $(\delta, \varepsilon)_m$?

◻ נניח שבחרנו מחלוקת היפותזה קטנה מספיק. האם נוכל למצוא לומד A קונקרטי שמציל תמיד? אם כך, כמה דוגמאות m הוא צריך בשבייל להצליח תמיד?

◻ האם נוכל למצוא את האלגוריתם הייעיל ביותר? (לומד עם מספר דוגמאות מינימלי)

2.1 משחק הלמידה (גרסה שלישית)

נקבע את הפרמטרים $(\delta, \varepsilon) \in (0, 1)^2$. נקבע $\mathcal{X} \subset \mathcal{H}$ ואז:

□ נבחר גודל מוגן m ולומד $\mathcal{H} \rightarrow \mathcal{X}^m$: \mathcal{A} שנייהם תלויים ב- (δ, ε) .

□ הطبع יודע את האסטרטגיה שלנו ובבקבוקת כך בוחר אסטרטגיה שכוללת הסתרות \mathcal{D} מעל \mathcal{X} ופונקציית תיוג $\mathcal{U} \rightarrow \mathcal{A}$: f מחלוקת ההיפותזה. האסטרטגיה של הطبع עלולה להיות תלואה גם ב- (δ, ε) וגם בבחירה של m ו- \mathcal{A} ובחלוקת ההיפותזה (שים לב שביקשנו כען תנאי קשוח יותר, השווו למקרה שביקשנו קודם לכן).

□ מוגן S מוגן m מיוצר לפי \mathcal{D} ומתיוג לפי f .

□ נכנס את S לתוך \mathcal{A} וניציר כלל חיזוי $\mathcal{H} \in h_S$.

□ כעת התוצאה הסופית (payoff) הוא $L_{\mathcal{D}, f}$. מדובר בתוצאה רנדומית כיוון ש- S הינו רנדומי וממילא גם h_S .

□ הطبع עושה כמעט יכולתו על מנת לנצל. על מנת לעשות זאת, הוא יחש \mathcal{A} כך שהסתבות של לפחות $\delta - 1$ הפסד ($L_{\mathcal{D}, f}(h_S)$ לא עבר את ε). לא משנה איזו אסטרטגיה f , הطبع יבחר לשחק.

□ על מנת לוודא שניצחנו במשחק, נשחק בו המון המון פעמיים. בכל פעם גם אנחנו וגם הطبع נשחק את האסטרטגיה שלנו. כאמור, בכל פעם הדגימות הנוצרות הין שונות. נחשב את ההסתבות (על התוצאות הרנדומות של מוגני האימון) של המאורע $\{S \sim \mathcal{D}^m \mid L_{\mathcal{D}, f}(h_S) \leq \varepsilon\}$. אם ההסתבות שנמצאה גדולה מ- $\delta - 1$ אז הלומד \mathcal{A} שנבחר הינו קרוב לוודאי בערך נכון, עם דיק ε ומקדם ביחסו δ - נגד האסטרטגיה של הطبع. נוכל להגיד במקרה זה כי ניצחנו.

דוגמה - פונקציות סף (Threshold Functions)

ראינו קודם לכן מוגן שהינה גודל מדי (חלוקת הפונקציות גדולה מדי) ולכן לא מתאים לגרסה השלישית של המשחק. ברור כי גרסה קטנה מדי תהיה טרוייאלית ללמידה. כיצד נוכל לבצע גודל מתאים של \mathcal{H} ?

בדוגמה הבאה נראה דוגמה למרחב מוגן אינסופי (בר מנייה) וחלוקת היפותזה פשוטה, שתצליח לכל δ, ε . נניח כי מרחב המוגן הוא $\mathbb{R} = \mathcal{X}$ - כלומר יש פיצ'ר אחד בלבד. ניקח מחלוקת סיווג, כך שקבוצת התיוג תהיה $\{0, 1\}$ במקומות $\{\pm 1\}$, בשביל להשתמש בגרסה הסטנדרטית של פונקציות סף.

הגדרה
קבוצת הפונקציות:

$$\mathcal{H}_{th} = \{x \mapsto h_\theta(x) : \theta \in \mathbb{R}\}$$

כאשר $[x > \theta]$ לכל $x \in \mathbb{R}$, נקראת מחלוקת היפותזה של פונקציות הסף (Threshold Functions) מעל \mathbb{R} .

עבור מחלוקת היפותזה h_{th} , משחק האלגוריתם הלמידה שלנו נראה בזורה הבא: פונקציה בחלוקת ההיפותזה דלעיל מסובגת נקודות על הישר המשני בין 0 לנקודת הסף המסוימת. מעבר לסק', היא מסובגת את כל הנקודות ל-1. הطبع בוחר אחת מהפונקציות הללו, שבודחת סף θ כשהיא ידוע לנו (שאנו ידוע לנו) והסתבות \mathcal{D} מעל הישר

המשי. אנחנו מקבלים מרכיב אימון S של נקודות מתויגות, כנדרשה לחזות בהצלחה דגימות עתידיות. אם כך, נרצה לחזות (מדויק ככל שניתן) את θ .

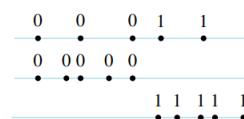


Figure 4.1: An example of a threshold function

אחרי שתוננו לנו מחלוקת ההיפותזה, נctrיך להגדיר את מחלוקת האסטרטגיות, שמכילה את מספר הדגימות m שנctrיך על מנת שאלגוריתם הלמידה A ייצור כל החלטה. כאמור, תהי $S = \{(x_i, y_i)\}_{i=1}^m$ קבוצת האימון. כפי שראתה, הלומד לא יהיה תלוי ב- δ , אך m אכן תהיה תליה בהם.

אלגוריתם למידה עבור פונקציות סף

אפשר לראות דוגמאות לבחירת דגימות אימון:

Figure 4.2: Valid training data for H_{th} Figure 4.3: Invalid training data for H_{th} - violates the Realizability Assumption

השורה הראשונה בוחרת מגם תקין (מחלקת ההיפותזה), והשורה השנייה בוחרת מידע שאינו תקין (לא שייך למחלוקת ההיפותזה הרצוי). אחד האלגוריתם, שמתאים לעקרון ERM הינו:

אלגוריתם 3 מצא פונקציית סף

אם $y_i = 1$ לכל $i \leq m$, אז $\theta_{alg} = -\infty$ (אלגוריתם הלמידה מסוווג את הנקודות כ-1). אם $y_i = 0$ לכל $i \leq m$, אז $\theta_{alg} = +\infty$ (אלגוריתם הלמידה מסוווג את הנקודות כ-0).

בשאר המקרים, קיבל כי $\theta_{alg} = \max_i \{x_i : y_i = 0\}$.

הערך של האלגוריתם הינו כזה:



מספר הדגימות

בהתנתו $(0, 1) \in \delta, \varepsilon$, נרצה למצוא מהו m שmbטich שהסתברות של לפחות $\delta - 1$ השגיאה האמיתית תהיה במקסימום ε .

טענה

יהיו $\delta, \varepsilon, m \geq \frac{\log(\frac{1}{\delta})}{\varepsilon}$. אם $\delta, \varepsilon \in (0, 1)$ ו- $L_{D, f_\theta}(h_{\theta_{alg}})$ עם האלגוריתם לעיל, הינו במקסימום ε :

$$\mathcal{D}^m \{S \mid L_{\mathcal{D}, f_\theta}(h_{\theta_{alg}}) \leq \varepsilon\} \geq 1 - \delta$$

הוכחה

יהי \mathcal{D} פונקציית הסתברות מעל \mathbb{R} . תהי $f_\theta \in \mathcal{H}_{th}$ פונקציית הסיווג האמיתית. מהגדרות האלגוריתם, נקבל כי $\theta_{alg} \leq \theta$ (למה? כי האלגוריתם בוחר את ה- y_i המקסימלי שה- y_i שלו שווה ל-0). יכול להיות שההפרדה הינה "בצד ימי יותר"?)

כל החזוי שנתקבל מהאלגוריתם יהיה נכון עבור דוגמאות $\theta_{alg} \leq x$ או $\theta > x$ ולא נכון עבור $\theta_{alg} < x \leq \theta$ (כי למעשה לא התחשבנו בנקודות אלו). אם כך, השגיאה האמיתית מתקבלת מ:

$$L_{\mathcal{D}, f_\theta}(h_{\theta_{alg}}) = \mathcal{D}(x : \theta_{alg} < x \leq \theta)$$

נחלק למקרים:

□ אם $\varepsilon < \mathcal{D}(x : -\infty < x \leq \theta)$ אז מדריך האלגוריתם יתקיים גם כי $\varepsilon < \mathcal{D}(x : \theta_{alg} < x < \theta)$ ולכן השגיאה האמיתית תהיה תמיד קטנה יותר מ- ε .

□ אם $\varepsilon \geq \mathcal{D}(x : \theta' < x \leq \theta)$ בפרט ישנו θ' כך ש- $\varepsilon = \mathcal{D}(x : \theta' < x \leq \theta)$. במקרה נבחן כי אם ישנה נקודה S כך ש- $\theta' < x_i \leq \theta$ אז $y_i = 0$ ולכן בפרט $\theta' \leq \theta_{alg} \leq \theta$ וכן קיבל סך הכל:

$$L_{\mathcal{D}, f_\theta}(h_{\theta_{alg}}) = \mathcal{D}(\{x : \theta_{alg} < x \leq \theta\}) \leq \mathcal{D}(\{x : \theta' < x \leq \theta\}) = \varepsilon$$

הסיכוי לא לקבל דוגמאות כאלה במדגם האימון הינו $(1 - \varepsilon)^m$ (לכל דוגמה). אם כך, הסיכוי שנבחר שגיאה אמיתית שגדולה מ- ε קטן או שווה ל- $(1 - \varepsilon)^m$. נשים לב כי $e^{-m\varepsilon} < 1 - \varepsilon$ ולכן בפרט קיבל כי $< (1 - \varepsilon)^m$ ²² נקבע בפרט כי $\delta \geq \frac{\log(\frac{1}{\delta})}{\varepsilon} e^{-m\varepsilon}$. אם נבחר

סיכום

אם כך, ראיינו שעבור $\mathcal{X} = \{0, 1\}$ ומחלקת ההיפותזות של פונקציות הסוף, מצאנו אסטרטגיה שתמיד מנחת את הטבע!
דבר זה מביא אותה להגדרה המפורשת של PAC.

3 למידת PAC

הגדרה

מחלקת היפותזה \mathcal{H} נקראת "ניתנת ללמידה" ("PAC learnable") אם ישנו אלגוריתם למידה \mathcal{A} ופונקציה $m_{\mathcal{H}, \mathcal{A}}$:

$$(0, 1)^2 \rightarrow \mathbb{N}$$

□ לכל $(\varepsilon, \delta) \in (0, 1)^2$

$$\text{כיצד? } -m\varepsilon \text{ וモונוטונית } \log(e^{-m\varepsilon}) = -m\varepsilon^{22}$$

□ לכל הסתברות \mathcal{D} מעל \mathcal{X} .

□ לכל פונקיות תיוג $f : \mathcal{X} \rightarrow \{\pm 1\}$ כך שישנו $h^* \in \mathcal{H}$ שמקיים כי $0 = L_{\mathcal{D},f}(h^*)$.

כאשר נريץ את האלגוריתם \mathcal{A} על $m \geq m_{\mathcal{H},\mathcal{A}}(\varepsilon, \delta)$ דגימות שנוצרות iid על ידי \mathcal{D} ומתויגים על ידי f , האלגוריתם יחזיר היפותזה $(S) = \mathcal{A}(S)$ כך שבסתברות של לפחות $\delta - 1$ קיבל כי $L_{\mathcal{D},f}(h_S) \leq \varepsilon$.

$$\mathcal{D}^m (\{S \mid L_{\mathcal{D},f}(h_S) \leq \varepsilon\}) \geq 1 - \delta$$

נסמן את גודל המדגם המינימלי שנדרש על מנת שיתקיים הביטוי, ביחס לכל אלגוריתם, על ידי:

$$m_{\mathcal{H}}(\varepsilon, \delta) = \min_{\mathcal{A}} m_{\mathcal{H},\mathcal{A}}(\varepsilon, \delta)$$

הפונקציה $\mathbb{N} \rightarrow \mathbb{N}_m$ מכונה בתור **סיבוביות המדגם** (Sample Complexity) של מחלוקת היפותזה הניתנת ללמידה PAC.

3.1 למידת PAC של מחלקות היפותזה סופיות

תחילה, לשם הפשטות והבהנה, נתיחס למקרה בו \mathcal{H} הינה מחלוקת היפותזה סופית. עדין נתייחס למקרה בו מחלוקת היפותזה מסוימת גדולה. למשל, את \mathcal{H} בתור כל הפונקציה מ- $\mathcal{U} \rightarrow \mathcal{X}$ שיכולות להיות ממומשות בתוכנת פיתוח באורך של מקסימום b , עבור b קבוע וגדול. או שניקח את \mathcal{H} להיות כל הפונקציות $\mathcal{U} \rightarrow \mathcal{X}$ כך ש- $|\mathcal{X}| - |\mathcal{U}|$ הינו סופי.

לכארה, זה לא אומר לנו הרבה. אבל לרובה ההפתעה (!) נגלה כי ישנו סוג פשוט של לומד, שנקרא "מצוער הסיכון האמפירי" (Empirical Risk Minimization) שתמיד מציליח מחלוקת היפותזה סופיות. מדובר בReLUון טבעי למדי: תנסה להיות מדויק ככל שאתה יכול. לעומת זאת, נבחר פונקציה $h \in \mathcal{H}$ שנותנת את התיאוג הטוב ביותר עבור מספר מקסימלי של נקודות ב- S .

הגדרה

עבור \mathcal{H} ו- $h \in \mathcal{H}$ ו- $S = \{(x_i, y_i)\}_{i=1}^m$ נגידיר את הסיכון האמפירי (Empirical Risk) בתורו:

$$L_S(h) = \frac{1}{m} |\{i : h(x_i) \neq y_i\}|.$$

כלל h עם $y_i = h(x_i)$ לכל $i \leq m$ (כלומר $L_S(h) = 0$) נקרא **עקבי** (Consistent) עם מודל האימון S .

הגדרה

אלגוריתם שלכל S מייצר כלל חיזוי שמצויר את הסיכון האמפירי נקרא **לומד** "מצוער הסיכון האמפירי" (Empirical Risk Minimization) או $\text{ERM}_{\mathcal{H}}$. ובקיצור ERM. למשמעות:

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

כיצד אנחנו יודעים כי אכן יש מינימום זהה? כיוון שאנו יודעים כי $L_S \geq 0$ ונתנו ממצאים מעלה קבוצה סופית, בהכרח יש מינימום. למעשה, לעיתים יתכן כמה $h \in \mathcal{H}$ שיספקו את המינימום, ולכן $ERM_{\mathcal{H}}$ יכולה להתיחס לקבוצת אלגוריתמים שמחזירה אחת מהפתרונותים אם נניח ריזלביות, או $0 = L_S(f)$, אפשרי, כי הרי $\mathcal{H} \in f$ - וכך במקרה זה תמיד מקבל כלל החלטה עקי (consistent rule) וכך במקרה הריזלבי, כלל עקי- ERM הוא מיילים נרדפות.

למידת מחלקות סופיות

טענה
 יהי $\epsilon \in (0, 1), |\mathcal{H}| < \infty, m \geq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\epsilon}$ ונסמן את h_S^{ERM} בתור כלל החיזוי של לומד $ERM_{\mathcal{H}}$ כלשהו. אזי לכל $f \in \mathcal{H}$ ולכל \mathcal{D} , בהסתברות של לפחות $\delta - 1$ קיבל כי $\epsilon \leq L_{\mathcal{D}, f}(h_S^{ERM})$.

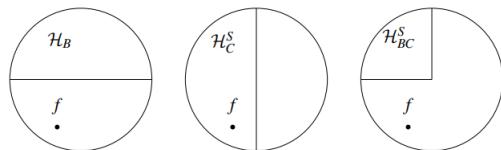
הוכחה
 יהי $\epsilon \in (0, 1)$ ותהי פונקציית הסתברות \mathcal{D} מעל \mathcal{X} ופונקציית תיוג f . על ידי בחרת \mathcal{D} ו- f , אנחנו מגדירים (באופן לא מפורש) תת קבוצה של \mathcal{H} שתסמן בתור \mathcal{H}_B (מלשון bad يعني). שמכילה את כל h הרעות, שאינם מערכיים את f כראוי:

$$\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{\mathcal{D}, f}(h) > \epsilon\}$$

כל שעליינו להראות הוא כי $\delta < \mathcal{D}^m(\{S \mid h_S^{ERM} \in \mathcal{H}_B\})$ נקבע בבדיקה ש- S מגדירה תת קבוצה של \mathcal{H} , שנסמנה בתור \mathcal{H}_C^S , שמכילה את כל h - h שהין עקיות עם f ב-

$$\mathcal{H}_c^S = \{h : L_S(h) = 0\} = \{h : h(\mathbf{x}_i) = f(\mathbf{x}_i), i = 1 \dots, m\}$$

כל קבוצה S גם מגדירה קבוצת חיתוך $\mathcal{H}_{BC}^S = \mathcal{H}_B \cap \mathcal{H}_C^S$ (כל h - h שהין עקיות ורעות). כיוון שהאלגוריתם שלנו הוא לומד ERM , בהכרח מתקיים כי $h_S^{ERM} \in \mathcal{H}_C^S$ ולכן כל הדגימות S שבתוך הקבוצה $\{S : h_S^{ERM} \in \mathcal{H}_B\}$ גם בחיתוך \mathcal{H}_{BC}^S שאינו ריק כיון שהוא מכל לפחות פונקציה אחת. כיוון ש- $\mathcal{D}^m(\{S : \mathcal{H}_{BC}^S \neq \emptyset\}) < \delta$ מספיק להוכיח כי $\mathcal{D}^m(\{S : h_S^{ERM} \in \mathcal{H}_B\}) \subseteq \{S : \mathcal{H}_{BC}^S \neq \emptyset\}$ דוגמאות לשולשות הקבוצות:



נבחן כי נוכל לסמן זאת גם בצורה הבאה:

$$\{S \mid \mathcal{H}_{BC}^S \neq \emptyset\} = \bigcup_{h \in \mathcal{H}} \{S \mid h \in \mathcal{H}_{BC}^S\} = \bigcup_{h \in \mathcal{H}_B} \{S \mid h \in \mathcal{H}_C^S\}$$

אם כך נדרש להוכיח כי:

$$\mathcal{D}^m \left(\bigcup_{h \in \mathcal{H}_B} \{S \mid h \in \mathcal{H}_C^S\} \right) < \delta$$

ומיחסם האיחוד קיבל כי:

$$\mathcal{D}^m \left(\bigcup_{h \in \mathcal{H}_B} \{S \mid h \in \mathcal{H}_C^S\} \right) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m (\{S \mid h \in \mathcal{H}_C^S\})$$

נניחו כי למעשה הדרישה בכל אחת מההסתברויות כי h מושלמת לנמרוי (כל התగיות נכונות). כיוון שכל הדוגמאות נוצרות בצורה בלתי תליה, נרצה כי h תהיה נכונה לכל x . ההסתברות כי h לא תהיה נכונה עבור x כלשהו הינה $(1 - L_{\mathcal{D},f}(h))^m$. לכן קיבל כי:

$$\mathcal{D}^m (\{S \mid h \in \mathcal{H}_C^S\}) = (1 - L_{\mathcal{D},f}(h))^m$$

בפרט מתקיים כי אם $\mathcal{D}^m (\{S \mid L_S(h) = 0\}) < (1 - \varepsilon)^m$, ובפרט קיבל כי $\varepsilon \text{ אי } h \in \mathcal{H}_B$ או $h \in \mathcal{H}_B$. לכן קיבל:

$$\mathcal{D}^m \left(\bigcup_{h \in \mathcal{H}_B} \{S \mid h \in \mathcal{H}_C^S\} \right) \leq \sum_{h \in \mathcal{H}_B} (1 - \varepsilon)^m < |\mathcal{H}_B| (1 - \varepsilon)^m \leq |\mathcal{H}| \cdot (1 - \varepsilon)^m$$

בשימוש בא"ש $1 - \varepsilon \leq e^{-\varepsilon}$ קיבל כי:

$$\mathcal{D}^m (\{S \mid L_{\mathcal{D},f}(\text{ERM}_{\mathcal{H}}(s)) > \varepsilon\}) < |\mathcal{H}| \cdot e^{-\varepsilon m}$$

אם נבחר $m \geq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon}$ קיבל כי הביטוי לעיל קטן מ- δ כפי שרצינו.

מהטענה זו עולה כי כל מחלוקת היפותזה סופית ניתנת ללמידה PAC שאיננה גדולה מ- $\frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon}$. מעבר לכך לכל m גדול או שווה לחסם זה, כל $\text{ERM}_{\mathcal{H}}$ הוא ניתן ללמידה PAC.

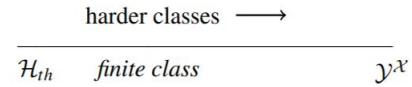
נותרו לנו עוד מספר שאלות שלא שאלנו:

$$\square \text{ האם החסם } m_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon}$$

\square מה קורה כאשר יש רעש? כלומר, y לא נוצר בהכרח לפי $-x$?

\square מה קורה בחלוקת היפווטה אינסופיota?

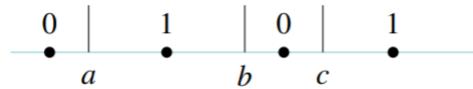
קודם לכן רأינו כיחלוקת היפווטה של פונקציות הס' הינה ניתנת למינית PAC, אך מחלוקת אחריות (למשל אינסופיota), לא ניתנת למינית. האם נוכל להסביר למה? למעשה אפשר לסדר את בעיות אלו בתווך "סדרל סיבוכיות", שנראה כך:



לדוגמא, אם נתבונן בחלוקת היפווטה של כל המקטעים. ככלומר:

$$\mathcal{X} = \mathbb{R}, \quad \mathcal{H} = \{h_{a,b,c} : a < b < c \in \mathbb{R}\}, \quad h_{a,b,c} = \mathbb{1}[x \in [a, b] \vee x \geq c]$$

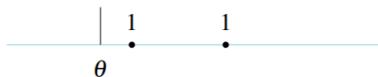
ונניח בדוגמה הבאה:



נניח נתבונן בדוגמה הבאה:

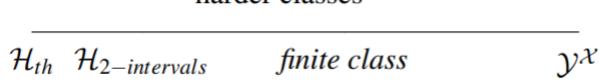


נוכל לחלק אותה כך:



אמנם, לא נוכל ללמד את הדגימות שמקיימות כי \mathcal{H}_{th} עם $x_1 < x_2, y_1 = 0, y_2 = 1$ עם $\mathcal{H}_{2-intervals}$ עם $x_1 < x_2, y_1 = 1, y_2 = 1$.

בקיצור, אנחנו מבינים שהוא אמין לא אינסופי, אבל זה עדין בסיבוכיות גדולה יותר:



$$\mathcal{H}_{th} \quad \mathcal{H}_{2-intervals} \quad \text{finite class} \quad \mathcal{Y}^{\mathcal{X}}$$

האם יש לנו דרך להעריך את הסיבוכיות? מסתבר שכן!

3.2 ממד ה-VC (VC-Dimension)

מדד ה-VC (VC Dimension) הוא סוג של סרגל סיבוכיות לחלוקת היפווטה. מדובר בסרגל קומבינטורי כיוון שהוא תלוי במספר האפשרויות של סיוג נקודות ב- \mathcal{X} . למעשה, סרגל זה מציע מבחן מדויק האםחלוקת היפווטה

הינה "פשוטה" יחסית במנחים של למידת PAC. היא גם מאפשרת לחשב את סיבוכיות המודגש של מחלוקת היפווצה זו.

נשים לב לבעה הבאה. נניח והצליחנו למצוא פונקציה $\mathcal{H} \in h$ שמייצגת את הדגימות שקיבלו (עם שגיאה אפסית), ולאחר מכן שינו בידים חלק מהתווות וקיבלו עבור הדגימות חדשות פונקציה $\mathcal{H} \in h'$ שמייצגת את הדגימות החדשות. יש כאן בעיה! אנחנו צריכים לדעת להקליל את המודגם זהה למוגדים אחרים, ללא קשר לתווות הספציפיות שקיבלו! מכאן המוטיבציה להגדרות הבאות.

הגדרה

תהי $\mathcal{X} \subseteq C$ תת קבוצה של מרחב המוגדים, \mathcal{Y} ותהי $\mathcal{H} \in h$ היפווצה. הפונקציה $h_C : C \rightarrow \mathcal{Y}$ מוגדרת על ידי: $(x) \in \mathcal{H}_C = \{h_C | h \in \mathcal{H}\}$, והיא נקראת הצמצום (restriction) של h ל- C . הקבוצה \mathcal{H}_C נקראת הצמצום של \mathcal{H} ל- C .

הגדרה הזאת חשובה, כי אם נסתכל היכן המשחק שלנו קודם, נבחן שההפסד במשחק נבע מכך ש- \mathcal{H} הכללה כל קבוצה C גדול $2m$ כך ש- $\mathcal{H}_C = \mathcal{Y}^C$. אם כך, הגודל המקסימלי של C ב- \mathcal{H} הוא קריטי! הוא גם נותן לנו חסם תחתון של $m_{\mathcal{H}}$ ואם הוא אינסוף, כלומר לכל $N \in m$ מתקיים כי \mathcal{H} מכילה קבוצה $|C| > m$, אז \mathcal{H} ניתנת למידת PAC.

הגדרה

תהי $C = \{x_1, \dots, x_{|C|}\}$, $\mathcal{Y} = \{\pm 1\}$ ו- \mathcal{H}_C הצמצום של \mathcal{H} ל- C . נאמר כי \mathcal{H} מנפצת את C (shatter) אם $\mathcal{H}_C = \mathcal{Y}^C$. דבר זה שקול לומר כי $|\mathcal{H}_C| = 2^{|C|}$.

הגדרה

ממד ה-VC (VC-dimension) של מחלוקת היפווצה \mathcal{H} מוגדר בטור:

$$\text{VC dim}(\mathcal{H}) = \max\{|\mathcal{S}| : \mathcal{H} \text{ shatters } \mathcal{S}\}$$

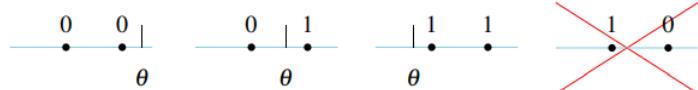
כלומר, הגודל המקסימלי של תת קבוצה $\mathcal{S} \subset C$ כך ש- $\mathcal{H}_C = \mathcal{Y}^C$.

דוגמה

נתבונן במחלוקת ההיפוציות של פוקציית הסף \mathcal{H}_{th} אותה ראיינו קודם. הקבוצה $\{0\}$ מנפצת על ידי \mathcal{H}_{th} , כי למשל:



אבל שתי נקודות x_1 ו- x_2 לא יכולות להיות מנופצות, כיון שייתכן מצב בו 0 מופיע בצד ימין:



דוגמה נוספת

נתבונן במחלקה המקטע הבודד (One-Interval hypothesis class) מעל \mathbb{R} :

$$\mathcal{H} = \{h_{a,b} : a < b \in \mathbb{R}\}, \quad h_{a,b}(x) = \mathbb{1}[x \in [a, b]]$$

נראה לדוגמה שני נקודות $\{0, 1\}$. אנחנו יכולים למקם את המקטע על שנייה על אחד מהם או מחוץ לו $[0, 1]$ ולכן $\{0, 1\}$ מנווץ. אמנם, כל שלוש נקודות אין יכולות להתנפץ (אין מקטע ובדד שיכל לכיסות את x_1 ו- x_3 בלי לכיסות את x_2 ויש דוגמה נגדית כ- $y_2 = 0$).

3.3 המפט היסודי של למידה סטטיסטית (The Fundamental Theorem of Statistical Learning)

בסעיפים הקודמים עסוקנו בלמידת PAC ובממד VC . ראיינו כל מיני חסמים מעניינים שתכף נתעסק בהם שוב.

למעשה, קודם לכן דיברנו על כך שהיכולת לבדוק האם מודל מסוים הינו ניתן ללמידה PAC ומזה סיבוכיות המדגם שלו, באמצעות ממד VC (VC-dimension). דבר זה נובע מהמשפט היסודי של הלמידה הסטטיסטית, שלמעשה אומר את הדברים הבאים:

- למידת PAC ניתנת להסקה על ידי VC-dimension.
- מחלוקת היפותזה הינה ניתנת ללמידה PAC אם ורק ממד ה- VC -dimension הינו סופי.
- אם $\text{dim } (\mathcal{H})$ הינו סופי, אז \mathcal{H} יש סיבוכיות מוגבלת, שנתונה על ידי $m_H(\varepsilon, \delta) \sim \frac{VC \cdot \dim(\mathcal{H}) + \log(1/\delta)}{\varepsilon}$.
- כלל הלמידה הוא "לומד" (כמעט) אופטימלי, כשהמקרה בו מחלוקת היפותזה ניתנת ללמידה PAC, לומד ERM משמש בכך ש- $m(\varepsilon, \delta) \sim \frac{V \cdot \dim(\mathcal{H}) \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$.

משפט (המשפט היסודי של הלמידה הסטטיסטית)

תהי \mathcal{H} מחלוקת היפותזה של מסוגים ביןaries עם $VC \dim d \leq \infty$. אז \mathcal{H} ניתנת ללמידה PAC אם ורק אם $d < \infty$. במקרה זה ישנו קבועים C_1, C_2 כך שסיבוכיות המדגם של \mathcal{H} מקיימת:

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}$$

מעבר לכך, החסם העליון של סיבוכיות המדגם ניתן להשגה על ידי לומד ERM.

כפי שראינו, האינטואיציה לחסם התחרתו מבוססת על הקשר בין ממד ה- VC ובין המספר המינימלי של דוגמאות הנדרשות על מנת להגיע ללמידה PAC. על מנת להבין את החסם העליון, علينا להבין את הקשר בין העבודה שהוא לומד גנרי ובין הממד של ה- (\mathcal{H}) .

4 למידת Agnostic PAC

בשלד התיאורטי שפיתחנו ישנו מספר הגבלות:

- ◻ הוא איננו מתייחס לרעש. לעיתים יכולות להיות שגיאות בדגימות (הוא אמור להיות 1 אבל מתואג כ-1-).
 - ◻ אנחנו רוצים מודל שיאפשר בהסתברות נמוכה, לצפות בנקודה או פעמיים ולקבל שתי תוצאות שונות.
 - ◻ הנחת הריאלבליות אינה מציאותית.
 - ◻ הוא מצומצם לשגיאת misclassification. נרצה להרחיב את המודל לכל פונקציית הפסד (loss)
- על מנת להתמודד עם בעיות אלו, נציג את מודל ה- Agnostic PAC. בנוסף, נראה כיצד המשפט היסודי עובד גם לגבי!

4.1 הקדמה לפונקציית התפלגות משותפת מעל $\mathcal{X} \times \mathcal{Y}$

קודם לכן, \mathcal{D} הייתה פונקציית התפלגות מעל \mathcal{X} . במקרה שלנו (שהינו כללי יותר), \mathcal{D} תהיה פונקציית התפלגות מעל $\mathcal{Y} \times \mathcal{X}$.

כלומר, אנחנו מייצרים כתה דוגמה (y, x) כך שני המשתנים מיוצרים בצורה רנדומלית, אך הם תלויים זה זה! נוכל להגדיר את \mathcal{D} בשתי דרכים - מותנים ב-y או מותנים ב-x. נסתכל על שניהם לשם ההבנה:

- ◻ (x) אנחנו מייצרים כתה דוגמה $(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y | X = x)$. מהזווית הזאת, מדובר למעשה בעשיה בהכללה ישירה של מה שעשינו קודם, כאשר x הוא משנה מקרי ו- $y = f(x)$. לאחר שיצרנו את x רנדומלית באמצעות $\mathbb{P}(Y = y | X = x)$ אנחנו בוחרים תווית מתאימה באמצעות פונקציית ההתפלגות המותנית $\mathbb{P}(Y = y | X = x)$. נוכל להסתכל על זה האינטואיטיבית: המשנה המקרי Y הוא משתנה מקרי ברנולי, נוכל לחושב עליו בהתאם למשתנה $p = \mathbb{P}(Y = 1 | X = x)$. אם $p = 0$ אז $y = 0$, אזי התווית נוצרת בצורה דטרמיניסטית וזכורנו לקרה הקודם. אבל במקרה الآخر, בו מתקבים ערכים אחרים של x, זה תלוי בעשיה ב-x ולכן נוצר רעש.²³

- ◻ (y) מזווית זו, אנחנו קודמים כל מייצרים את התווות ורק אז מייצרים את הדוגמאות. זה תהליך דומה ל-LDA.

קודם לכן הגדכנו את שגיאת הסיווג (misclassification loss) בטור ההסתברות שהתקבל x שהתגית האמיתית שלו שונה מהתגית החזויה. כתה, כיון ש- \mathcal{D} היא הסתברות מעל $\mathcal{Y} \times \mathcal{X}$, נכיל זאת ונגדיר את השגיאה בטור ההסתברות שהתקבל x כך שהתגית שלו שונה מהחזויה:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} \{h(\mathbf{x}) \neq y\} = \mathcal{D}\{(\mathbf{x}, y) | h(\mathbf{x}) \neq y\}$$

²³מדובר בדבר דומה למה שעשינו ברגסיה לוגיסטיבית, אלא שם לא התייחסנו להסתברות של x. אנחנו שמתקבלות דוגימות וניסינו להעריך את ההסתברות.

4.2 עידון הנחת הריזולבליות (Relaxing Realizability Assumption)

קודם לכן הנחנו ריזולבליות. נזכיר שהה פשטוט אומר כי אנחנו מוגעים לשגיאת הכללה ששויה לאפס, כלומר כי יש h כלשהו שימושה את f המקורי.

עכשו אנחנו לא ממש יכולים לעשות את זה. למה? אנחנו משווים פונקציה דטרמיניסטית לפונקציה רנדומית. משמע - אי אפשר לצפות להפסד אפסי. בקיצור, אין לנו באמות f אמייתית, אי ברור שאי אפשר להעריך אותה. הדבר הכי קרוב שיש לנו הוא פונקציית התפלגות מותנית שמודגרת על ידי $\mathbb{P}(Y = y | X = x)$.

עם כל הצער שבדבר, הרילזובליות כבר לא רלוונטית כרגע (במובן שאיננו יכולים לצפות לפונקציית הכללה אפסית).

אם כך, מכאן נובע כי נctrיך גם לשנות את ההגדירה של הדיקט (accuracy).

הגדרה

תהיה \mathcal{D} פונקציית התפלגות מעל $\mathcal{Y} \times \mathcal{X}$. נגידר את מסוג הביס האופטימלי (Bayes Optimal Classifier) בתורו²⁴:

$$f_{\mathcal{D}}(\mathbf{x}) = \begin{cases} 1 & \mathbb{P}(y = 1 | \mathbf{x}) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

נבחן כי למרות שאנו מגדירים את ההיפותזה שלנו בתורו $\mathcal{Y} \rightarrow \mathcal{X}$, היא תלואה ב- \mathcal{D} שאינה ידועה לנו (לפי חוקי המשחק). אם כך, $f_{\mathcal{D}}$ הינה למעשה כמות נבואה (Oracle Quantity) : אם יש לנו נבואה שאומרת לנו מהו \mathcal{D} נוכל לסוג עם $f_{\mathcal{D}}$. באופן כללי נשתמש בדבר זה על מנת להשוו את הפסד של כלל כלשהו לכל החלטה הטוב ביותר האפשרי.

הגדרה

יהי $0 > \varepsilon$. נאמר כי כלל $\mathcal{H} \in \mathcal{H}$ הוא בערך נכון (Approximately Correct) עם דיוק ε , ביחס לפונקציית התפלגות \mathcal{D} מעל $\mathcal{Y} \times \mathcal{X}$ אם $L_{\mathcal{D}}(h) \leq L_D(h) + \varepsilon$ מההפסד הטוב ביותר במרקח ε מההפסד הטוב ביותר האפשרי, לכל היפותזה ב- \mathcal{H} :

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

נבחן כי מדובר בהכללה של ההגדירה הקודמת. למה? אנחנו כי קיימת $\mathcal{H} \in \mathcal{H}$ עם $h' = 0$ ($L_D(h') = 0$) ולכן ההפסד המינימלי היה תמיד 0. כאן זה כבר לא המצב. אם כך, הטבע גם לא בוורח פונקציית תיווג f אלא רק פונקציית התפלגות משותפת \mathcal{D} מעל $\mathcal{Y} \times \mathcal{X}$.

4.3 פונקציית הפסד כללית (General Loss Function)

ההרחבה האחרונה שאנו נדרשים לעשות היא ליצור פונקציית הפסד כללית.

הגדרה

פונקציית הפסד (A Loss Function) הינה פונקציה $\ell : \mathcal{H} \times Z \rightarrow [0, \infty)$ כאשר \mathcal{Y} כאמור

דיברנו כבר על פונקציית הפסד המפורסמת - שגיאת הסיווג שידועה בתור שגיאת 0-1 :

²⁴ להרחבה - רואו בתרגיל 3 בשאלות הראשונות.

$$\ell_{0,1}(h, (\mathbf{x}, y)) := \begin{cases} 1 & h(\mathbf{x}) \neq y \\ 0 & h(\mathbf{x}) = y \end{cases}$$

למעשה, ניתן לכתוב את ההגדרה של $L_{\mathcal{D}}(h)$ גם ככזה:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}} [\ell_{0,1}(h, (\mathbf{x}, y))]$$

כאשר הסימון $\mathbb{E}_{\mathcal{D}}$ מסמן כי התוחלת הינה ביחס ל- \mathcal{D} .

בהתחרש בהגדרות אלו, נוכל ליזור הגדרה חדשה לשגיאת ההכללה.

הגדרה

בහינתן פונקציית הסתברות \mathcal{D} מעל $\mathcal{Y} \times \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, היפותזה $\mathcal{Y} \rightarrow \mathcal{X}$: h ופונקציית שגיאה כללית loss (general loss) המוגדרת על ידי $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow [0, \infty)$ (function) נגדיר את שגיאת ההכללה $L_{\mathcal{D}}(g)$ של היפותזה הנוצרת על ידי ℓ ביחס ל- \mathcal{D} כתוחלת של ℓ

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}[\ell(h, z)]$$

כיוון שההגדרה של "בערך נכון" (Approximately Correct) לא תתייחס מפורשת לבירהה של פונקציית ההפסד, היא נשארת רלוונטיות גם במקרה של שגיאת הההפסד הכללית.

4.4 למידת Agnostic-PAC

המודל החדש והכללי יותר שאנו מפתחים נקרא Agnostic-PAC.

הגדרה

היו $\varepsilon \in (0, 1)$. נאמר כי לומד \mathcal{A} הוא אגנסוטי קרוב לוודאי בערך נכון (Probably Approximately Correct) אם מקדם ביטחון δ ומקדם דיקוק ε ביחס לפונקציית ההפסד ℓ , מחלוקת היפותזה \mathcal{H} והסתברות \mathcal{D} מעל $\mathcal{Y} \times \mathcal{X}$ אם הסתברות שנתקבל דוגמאות אימון S כך ש- \mathcal{A} - S יצא כל חייזר $h_s \in \mathcal{H}$ עם הפסד $L_{\mathcal{D}}(h_s) \leq \delta$ בפרק 1 – לכל היוטר של ε מההפסד הטוב ביותר האפשרי בכל היפותזה ב- \mathcal{H} הינה קטנה או שווה $\delta - \varepsilon$:

$$\mathcal{D}^m \left(\left\{ S \mid L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon \right\} \right) \geq 1 - \delta$$

הגדרה

מחלקת היפותזה \mathcal{H} הינה ניתנת ללמידה PAC אגנוטטי, ביחס לפונקציית הפסד (\cdot) אם $\ell : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \rightarrow [0, \infty)$ אמ' יש פונקציה $\mathbb{N}^{\tilde{m}_{\mathcal{H}}} \text{ ואלגוריתם למידה } \mathcal{A} : (\mathcal{X} \times \mathcal{Y}^m) \rightarrow \mathcal{H} \text{ עם התכונות הבאות:}$

$$\square \text{ לכל } \varepsilon, \delta \in (0, 1) \text{ }$$

$$\square \text{ לכל הסטברות } \mathcal{D} \text{ מעל } \mathcal{Y} \times \mathcal{X}$$

כאשר נريץ את האלגוריתם \mathcal{A} על m דוגימות שנוצרות iid על ידי \mathcal{D} , האלגוריתם יחזיר היפותזה $\mathcal{A}(S) = h_S$ כך שיתקיים, בהסתברות של $\delta - 1$:

$$\mathcal{D}^m \left(\left\{ S \mid L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon \right\} \right) \geq 1 - \delta$$

אפשר להראות בקלות כי לומד PAC אגנוטטי הוא לומד PAC רגיל.

הבה ונכתב כעת את משחק הלמידה בצורת ה-PAC האגנוטטי.

יהו $(\varepsilon, \delta) \in (0, 1)$ מקדמי דיוק וביטחון, תהי \mathcal{H} מחלקה היפותזה כך ש- $\mathcal{Y} \subset \mathcal{X}$ ופונקציית הפסד ℓ . נחקק נגד הטבע, בתשלום רנדומלי:

\square נבחר גודל מבחן m ולומד $\mathcal{H} \rightarrow \mathcal{A} : (\mathcal{X}, \mathcal{Y})^m \rightarrow \mathcal{A}$ כאשר m ו- \mathcal{A} יכולים להיות תלויים ב- ε, δ .

\square מבחן $S \in \mathcal{D}$ של m דוגימות נוצרת בהתאם ל- \mathcal{D} .

\square נכניס את המבחן S לתוך \mathcal{A} ונניצר כלל החלטה $h_S = \mathcal{A}(S)$. נבחן כי $h_S \in \mathcal{H}$.

\square כעת התשלום הינו $L_{\mathcal{D}}(h_S) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\ell(h, (\mathbf{x}, y))]$. הוא רנדומלי כיון ש- S נוצר רנדומלי ואז גם h_S רנדומלי.

\square הטבע עושה את הטוב ביותר שלו לניצח, אז אנחנו מחרשים לומד \mathcal{A} שיש לו את ההפסד המקסימלי והבטווח $L_{\mathcal{D}}(h)$ לכל אסטרטגיה f שהטבע עלול לבחור. (a guaranteed maximal loss)

\square על מנת לוודא אם ניצחנו, אנחנו משחקים את המשחק המון פערניים. אנחנו סוברים ומחשבים את $\left\{ S \sim \mathcal{D}^m \mid L_{\mathcal{D}}(h_S) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon \right\}$. אם ההסתברות שמעל דוגימות שנוצרות רנדומלית, S , של המאורע $L_{\mathcal{D}}(h) < \varepsilon$ אז ניצחנו.

על מנת לחשב את המאורע הנ"ל, علينا להניח כי יש בידינו את הידע על ε (כמפורט נבואית זהה). האם ללמידה PAC אגנוטטי אפשרה לנו יותר בחירות של \mathcal{D} ? וואלה כן! מעבר לכך, נוכל להבחן בקשר שבין ללמידה PAC ובין ללמידה PAC אגנוטטי.

טענה

יהי \mathcal{X} מרחב מבחן ותהי $\mathcal{Y} \subset \mathcal{X}$ מחלקה היפותזה. \mathcal{H} ניתנת ללמידה PAC אם ורק אם היא ניתנת ללמידה PAC אגנוטטי.

4.5 המשפט היסודי של הלמידה הסטטיסטית

קודם לכן הגדרנו את המזעור האמפירי המינימלי (ERM) בטור $\mathcal{H} \in h$ שהינה עקבית עם מוגן האימון שלנו. אנחנו לא רילזבלים ולא נחנכו צריכים הגדרה אחרת.

הגדרה

יהי $\mathcal{Y} : \mathcal{X} \rightarrow h$ כלל חיזוי. נגדיר את הסיכון האמפירי (empirical risk) של h ביחס לפונקציית ההפסד ℓ והדגימות על ידי $S = \{\mathbf{x}_i, y_i\}$:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i), \quad z_i = (\mathbf{x}_i, y_i)$$

נאמר כי אלגוריתם הינו אלגוריתם למידה ERM אם הוא אלגוריתם שמייצא היפותזה שמצועתת את הסיכון האמפירי:

$$\mathcal{A}_{ERM} : S \mapsto h, \quad h \in \left\{ \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) \right\}$$

כמובן מה-ERM לא חייב להיות יחיד. יכולות להיות מספר היפותזות שימושיות תוצאה זו.

הרענון של לומדי ERM מובן לפי העיקרונו של החוק החלש של המספרים הגדולים²⁵. חוק זה למעשה אומר כי אם הינה סדרה של משתנים מקריים בלתי תלויים ושווים התפלגות ו- $(X_i) = \mu$ אז מתקיים כי:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m X_i = \mu$$

כשההתפלגות הינה בהסתברות. כמובן, לכל $0 < \delta$ מתקיים כי:

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

דבר זה שקול לומר כי לכל $0 < \delta$ קיים $m_0 \in \mathbb{N}$ כך שלכל $m > m_0$ מתקיים כי:

$$\mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^m X_i - \mu \right| > \delta \right\} < \varepsilon$$

מהנתונים שיש בידינו עולה כי לכל h מותקיים כי $L_D[h] = \mathbb{E}_D[L_S(h)] = L_S(h)$. מהחוק החלש מותקיים כי $L_D(h)$ (בהסתברות וכאשר S נוצרת בדומה בלתי תלוי ושוות התפלגות). כמובן, לכל $0 < \delta$ קיים $N \in \mathbb{N}$ כך שלכל $m > m_0$ מותקיים כי:

²⁵ככלל, חוק זה אומר: ככל שתעשה יותר ניסויים תקרב יותר לתוחלת.

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

האם נתונים אלו מספיקים בידינו כדי לומר כי $\operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ קרוב ל- $\operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$

4.5.1 המשפט היסודי

משפט (המשפט היסודי של הלמידה הסטטיסטית)

תהי \mathcal{H} מחלקה היפותזה של מסווגים בינהירים עם $\dim_{VC} d \leq \infty$. אזי \mathcal{H} ניתנת ללמידה PAC אגנוטית אם ורק אם $\infty < d$.

במקרה זה ישנו קבועים C_1, C_2 כך שסיבוכיות המודגם של \mathcal{H} מקיימת:

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$

מעבר לכך, החסם העליון של סיבוכיות המודגם ניתן להשגה על ידי לומד ERM.

שיםו לב כי המחיר שאנו משלמים עבור למידת PAC אגנוטי היא כי סיבוכיות המודגם פרופרציונלית ל- $\frac{1}{\varepsilon^2}$ ולא ל- $\frac{1}{\varepsilon}$.

על מנת להשלים פרק זה, ניגע מעט בהוכחה של המשפט, שתשופך לנו אוור על האינטואיציה והקשר של ממד ה- VC ללמידה PAC אגנוטי.

כבר רأינו בעיקרונו "אין ארוחות חינס" קצר אינטואיציה על טענה זו. ראיינו כי אם יש $\mathcal{X} \subset C$ שמנופצת על ידי \mathcal{H} , אין אלגוריתם למידה שיכול להיות PAC אם יש בו פחות מ- $\frac{|C|}{2}$ דוגמאות. הרעיון של $\dim_{VC} = \infty$ אומר למעשה כי ישנו תת-קבוצה של \mathcal{X} מוגדל שירוטי שהין מנופצת על ידי \mathcal{H} ולכן מספר סופי של מודגים.

על מנת להבין מעט יותר את הקשר בין סופיות ממד ה- VC ובין למידת PAC אגנוטי, נctrיך להגדיר התכנסות במידה שווה של מחלקות היפותזה.²⁶

4.5.2 תוכנות התכנסות במידה שווה

נזכר כי לומד ERM בוחר כלל (s) L_S שמצויר את (h) עבור מודגם S .

²⁶ חלק זה לא מדן במלל לימוד אינפוי לממד "ח' וכו'. דרושה אינטואיציה על מנת להביןcosa זה טוב יותר. כיוון שלא ניגע כאן בתוכנות נקודתיות של פונקציית, לא נדרש על הנושא יותר מדי. אבל הרעיון כאן הוא שהוא מרחיב שלנו הוא מרחיב פונקציות - אילו הם למעשה במרחב. כיוון שאנו דורשים התוכנות לאייררים במרחב נרצה להבין מהי משמעויות התוכנות במרקחה זה. סדרת פונקציות הינה למעשה תלתן כלשהיא של הפונקציה $b-a$. למשל x^n זאת סדרה פונקציית, כאשר $\{ \dots, x^2, x, 1 \}$ הם האיברים בסדרה. הרעיון של התוכנות במידה שווה הוא כי החל מ- n מסוים כל הפונקציות מוכנסות לאותו ערך. מבחינה אינטואטיבית, נאמר כי סדרת פונקציה מוכנסת לפונקציה מסוימת, אם לכל n קיים שרוול ε , כך שהחל מאותו n כל הפונקציות מוציאות בשרוול זה.

אנחנו מוקווים כי $h_S \in \text{ERM}_{\mathcal{H}}(S)$ תמצער היטב. למעשה, נרצה להוכיח כי:

$$\mathcal{D}^m \{S \in (\mathcal{X} \times \mathcal{Y})^m | |L_{\mathcal{D}}(h_S) - L_S(h_S)| \leq \varepsilon\} \geq 1 - \delta$$

דבר זה יכול לקרות ורק אם S היא חתיכת דגימה מיוחדת - ציאת שלכל $\mathcal{H} \in h$ הסיכון האמפירי קרוב יותר לפונקציית ההכללה $(S, L_{\mathcal{D}})$. אין נוכל להוכיח את זה? תכל'ס זה די קשה. בואו ונחזר רגע לקשר שראינו בין שגיאת ההכללה $L_{\mathcal{D}}(h)$ והחומר החלש של המספרים הגדולים. - ראיינו כי $L_S(h)$ מתכנסת בהסתברות ל- $L_{\mathcal{D}}(h)$ כאשר $m \rightarrow \infty$:

$$\forall \mathcal{D} \forall h \in \mathcal{H} \forall \varepsilon, \delta \in (0, 1) \quad \exists m_0 \in \mathbb{N} \quad \text{such that} \quad \mathbb{P}\{|L_S(h) - L_{\mathcal{D}}(h)| < \varepsilon\} > 1 - \delta$$

אממה, m_0 תלוי כאן ב- \mathcal{D} וב- h , ואנחנו רוצים m_0 שמתכנס במידה שווה, בלי קשר ל- \mathcal{D} ו- h .

הגדרה

סדרה של פונקציות $f_n : X \rightarrow \mathbb{R}$ מתכנסת במידה שווה ל- $f : X \rightarrow \mathbb{R}$ אם ורק אם:

$$\forall \varepsilon > 0 \quad \exists m_0 \in \mathbb{N} \quad \forall x \in X |f_n(x) - f(x)| < \varepsilon$$

אם כך, אנו רוצים לוודא כי $L_S(h)$ מתכנס ל- $L_{\mathcal{D}}(h)$ במידה שווה ב- \mathcal{D} וב- h (ללא תלות בהם. תלות ב- ε בעצם).

הגדרה

מודם אימון S נקרא נציג- ε עבור $\mathcal{D}, \mathcal{H}, \ell$ אם מתקיים:

$$\forall h \in \mathcal{H} \quad |L_S(h) - L_{\mathcal{D}}(h)| < \varepsilon$$

בגדיול, נוכל לומר כי אם יש לנו מודם אימון S שהינו נציג- ε , אז $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq \text{ERM}_{\mathcal{H}}(S) + \varepsilon$ ישיג כמעט את $\text{ERM}_{\mathcal{H}}(S)$.

למה

יהי S נציג- $\frac{\varepsilon}{2}$ עבור $\mathcal{D}, \mathcal{H}, \ell$. יהי $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$. קלומר $L_S(h_S) \leq \text{ERM}_{\mathcal{H}}(S) + \frac{\varepsilon}{2}$.

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

הגדרה

מחלקת היפותזה \mathcal{H} הינה בעלת תכונת ההתקנסות במ"ש, אם ורק אם ישנה פונקציה $N : (0, 1)^2 \rightarrow$ כך שלכל $m_{\mathcal{H}}^{\text{UC}} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$ וכל הסטברות \mathcal{D} מעל $\mathcal{X} \times \mathcal{Y}$, לכל $m \in (0, 1)$ מתקיים:

$$\mathcal{D}^m(\{S \in (\mathcal{X} \times \mathcal{Y})^m \mid S \text{ representative is } S\}) \geq 1 - \delta$$

ניתן להראות כי אם מחלקת היפותזה בעלת תכונת ההתקנסות במ"ש, עם פונקציית $m_{\mathcal{H}}^{\text{UC}}$, אז \mathcal{H} היא ניתנת ללמידה PAC אגנוטטי, עם סיבוכיות מדגם $.m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}(\varepsilon/2, \delta)$.

מדד VC גורר תכונת ההתקנסות במידה שווה

אם כך, בהסתמך על מה שאמרנו, מספיק שnochich כי $\dim(\mathcal{H}) < \infty$ כדי $\text{VC dim}(\mathcal{H})$ איזו מתקיימת תכונת ההתקנסת במ"ש. הבה ונזכיר את ההדרות שראינו קודם לכן לדברים שקשורים יותר אלינו. על מנת להשיג "אחדות" (תכל"ס במ"ש, התרגומים המתמטיים לא צלחו כאן את מבחן התרגומים), גם על \mathcal{D} וגם על $\mathcal{H} \in h$, נגדיר את הפונקציה הבאה:

$$F_m^{\mathcal{D}} : (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathbb{R}$$

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

פונקציה זו מומפה למעשה דגימות בגודל m במספר שמסמל את הטיעות המקסימלי ביוון, את השגיאה הכי גרועה. נבהיר כי כיוון ש- $F_m^{\mathcal{D}}$ -היא פונקציה של המשתנה הרנדומלי S היא גם רנדומלית, כשההתפלגות שלה תליה למעשה בהתפלגות \mathcal{D}^m על מרחבי אימון מגודל m .

למעשה, נרצה (כפי שראמנו) להראות כי הפונקציה $F_m^{\mathcal{D}}$ היא קטנה. כמובן, נרצה להראות כי לכל $\varepsilon, \delta \in (0, 1)$ קיים N כך שלכל התפלגות \mathcal{D} מתקיים כי:

$$\mathcal{D}^m \{ F_m^{\mathcal{D}}(S) > \varepsilon \} < \delta$$

המקרה בו \mathcal{H} סופי

על מנת להבין את הרעיון, נתבונן תחילה במקרה הפרשוני בו מחלקת היפותזות שלנו \mathcal{H} הינה סופית.

טענה

יהיו $\varepsilon, \delta \in (0, 1)$, איזי ישנה $N \in \mathbb{N}$ כך שלכל $m_0 > m$ מתקיים כי:

$$\forall \mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y} \quad \mathcal{D}^m \{ F_m^{\mathcal{D}}(S) > \varepsilon \} \leq \delta$$

הוכחה

מההגדירה עולה:

²⁷אם זה מזכיר לכם משהו - אז זה בהחלט אמרת לזכור! ההגדירה שskolla להתקנסות במ"ש היא אם הגבול של הפונקציה הזאת שווה ל-0.

$$\begin{aligned}
 \mathcal{D}^m \{F_m^{\mathcal{D}}(S) > \varepsilon\} &\stackrel{\text{def.}}{=} \\
 &\underbrace{\text{חסם האיחוד}}_{\downarrow} \\
 \mathcal{D}^m \{S | \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} &\leq \\
 \sum_{h \in \mathcal{H}} \mathcal{D}^m \{S | |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} &\leq \\
 |\mathcal{H}| \cdot \max_{h \in \mathcal{H}} \mathcal{D}^m \{S | |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}
 \end{aligned}$$

מהחוק החלש של המספרים הגדולים וממה שראינו קודם נקבל כי:

$$\forall \varepsilon > 0 \quad \mathcal{D}^m \{|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \xrightarrow{m \rightarrow \infty} 0$$

אבל זה לא מספיק, כי אנחנו הרি רוצים אי תלות ב- \mathcal{D} .
 העשא כאן משהו מעניין. אנחנו מוחפשים דרך לחסום את המרחק בין הממוצע האמפירי ובין התוחלת.
 מי יבוא לעוזתנו? הופדינג, כМОון.
 נזכיר בו: יהיו $\theta_1, \dots, \theta_m$ סדרה של משתנים מקרים בלתי תלויים ושווי התפלגות, ונניח כי לכל i מתקיים כי $\mu = 1$ ו- $\mathbb{E}[\theta_i] = 1$, איזי מתקיים:

$$\forall \varepsilon > 0 \quad \mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^m \theta_i - \mu \right| > \varepsilon \right\} \leq 2 \exp \left(-2 \frac{m\varepsilon^2}{(b-a)^2} \right)$$

לעניןנו, נגיד ראת $L_S(h) = \frac{1}{m} \sum_{i=1}^m \theta_i$ וכי $L_{\mathcal{D}}(h) = \mathbb{E}[\theta_i]$. נזכיר כי $\theta_1 = \ell(h, (x_i, y_i))$.

$$\begin{aligned}
 \mathcal{D}^m \{S | |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} &\leq 2 \exp(-2m\varepsilon^2) \\
 &\Downarrow \\
 |\mathcal{H}| \cdot \max_{h \in \mathcal{H}} \mathcal{D}^m \{S | |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} &\leq 2|\mathcal{H}| \exp(-2m\varepsilon^2) \\
 \text{על ידי בחירת } m, \text{ נקבל כי:} &\geq \frac{\log(\frac{2|\mathcal{H}|}{\delta})}{\varepsilon^2}
 \end{aligned}$$

$$\mathcal{D}^m \{F_m^{\mathcal{D}}(S) > \varepsilon\} \leq 2|\mathcal{H}| \exp(-2m\varepsilon^2) \leq \delta$$

כנדרש.

המקרה של \mathcal{H} אינסופי
 עם כל הבאה, איננו יכולים להפעיל את חסם האיחוד על מספר אינסופי של היפותזות.

במקום זאת, נזכיר ביצומם של \mathcal{H} ל- C , אותו הגדרנו על ידי \mathcal{H}_C . המפתח בהוכחה הוא להבין **מה מהר** היצום \mathcal{H}_C גדול ביחס ל- $|C|$. אם $\text{VC dim}(\mathcal{H}) \leq |C|$ יתכן כי \mathcal{H} מנצח את $|C|$ ולכנו יתכן כי $|\mathcal{H}_C| = 2^{|C|}$. אמנם, אם $|\mathcal{H}| > \text{VC dim}(\mathcal{H})$ זה לא יכול להתקיים.

הגדרה

עבור מחלקה היפותזה \mathcal{H} נגידר את $(m) \tau_{\mathcal{H}}$ על ידי:

$$\tau_{\mathcal{H}}(m) = \max \{ |\mathcal{H}_C| \mid C \subset \mathcal{X}, |C| = m \}$$

למעשה, מדובר בהגדרה שמתיחסת לצירוף קומבינטוררי של \mathcal{H} : המספר המקסימלי של פונקציות שנוכל להשיג על ידי יצום של \mathcal{H} לכל תת קבוצה מוגדר m . ככל ש- \mathcal{H} יהיה מורכב יותר, ביטוי זה יהיה גדול יותר. אם למשל ממד ה-VC הוא $= 2^m$, אי- ∞ , כלומר, מדובר בסדר גודל אספוננציאלי.

הגדרה

תהיי $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. נניח כי קיים $0 < \beta < \infty$ ו- $m_0 \in \mathbb{N}$.

$$\forall m > m_0 \quad \tau_{\mathcal{H}}(m) \leq b \cdot m^{\beta}$$

נאמר כי \mathcal{H}_C גדול פולינומיאלית ב- $|C|$.

נרצה אם כך להוכיח שני דברים:

ראשית, כי אם $|\mathcal{H}_C|$ גדול פולינומיאלית ב- $|C|$ אי-מתקיים ב- \mathcal{H} תכונות ההתקנסות במידה שווה. כמו כן, אם $\text{VC dim}(\mathcal{H}) < \infty$, מתקיים כי $|\mathcal{H}_C|$ גדול פולינומיאלית ב- $|C|$.

טענה

תהיי \mathcal{H} מחלקה היפותזה כך ש- $|\mathcal{H}_C|$ גדלתה פולינומיאלית ב- $|C|$, אי-ל- \mathcal{H} יש את תכונות ההתקנסות במש"ש. הוכחה

נזכיר כי אנחנו רוצים להראות כי $\delta < \epsilon$ ($\{F_m^{\mathcal{D}}(S) > \epsilon\}$ ב- \mathcal{D}^m בצורה "אחדה" ב- \mathcal{D} (במש"ש למעשה). כיוון ש-

הוא משתנה מקרי אי שלילי, נוכל להשתמש בא"ש מורכב.

למה

תהיי $F_m^{\mathcal{D}}$ המוגדרת על ידי $F_m^{\mathcal{D}}(S) = \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ מתקיים כי:

$$\mathbb{E}_{\mathcal{D}^m} [F_m^{\mathcal{D}}(S)] \leq O \left(\frac{\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}} \right) + o(m)$$

כעת, כיוון שעל פי ההנחה $|\mathcal{H}_C|$ גדול פולינומיאלית ב- $|C|$, לפי ההגדרה קיים m_0 כך שכל $m > m_0$ מתקיים כי $\tau_{\mathcal{H}}(m) \leq b \cdot m^{\beta}$ ועוד מותקיים מהנתנו כי:

$$\mathbb{E}_{\mathcal{D}^m} [F_m^{\mathcal{D}}(S)] \leq O \left(\frac{\sqrt{\beta \cdot \log(2m)}}{\sqrt{2m}} \right) + o(m)$$

כעת, נוכל להשתמש בא"ש מركוב ויתקדים:

$$\mathbb{P}_{\mathcal{D}^m} \left\{ \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| > \varepsilon \right\} = \mathbb{P}_{\mathcal{D}^m} \{ F_m^{\mathcal{D}}(S) > \varepsilon \} \leq \frac{\mathbb{E}_{\mathcal{D}^m} [F_m^{\mathcal{D}}(S)]}{\varepsilon}$$

הביטוי למעלה מתכנס ל-0, כלומר קיבלנו ישרות מהгадרה כי \mathcal{H} יש את תכונת ההתקנסות במ"ש.

טענה

תהי \mathcal{H} מחלקת היפותזה עם ממד VC סופי. אז $|\mathcal{H}_C|$ גדול פולינומיאלית ב- $|C|$.

הוכחה

לפי ההגדרה, אם $m \leq \text{VC dim}(\mathcal{H})$ אין קיימת תת-קובוצה $\mathcal{X} \subset \mathcal{H}$ מוגדר m שמנופצת על ידי \mathcal{H} . מכאן עולה כי אם $m \leq \dim(\mathcal{H})VC$ ניתן להראות כי אם $m > \text{VC dim}(\mathcal{H})$ אז $(\frac{em}{d})^d \leq (\tau_{\mathcal{H}}(m))^m$ עבור d כלשהו. אם כך על אף כי $\tau_{\mathcal{H}}(m)$ גדול אקספוננציאלית ב- m , הוא רק רך גדול פולינומיאלית ב- m עבור $m > \text{dim}(\mathcal{H})VC$.

חלק V

שיטת אנסמבל - Ensemble Methods

בפרק זה לא נדבר על אלגוריטם ספציפי או שהוא כזה. במבט על נציג מספר שיטות שנוכל לישם בכל אחד מאלגוריתמי הלמידה שהכרנו, שיפרו ממשות את ביצועיהם. השימוש הקדוש הינו: אתחול-אווני נעלים²⁸ (Boosting), צבירת-אתחול (Bootstrapping), Bagging (bagging) והאצה (Boosting).

סביר כיצד כל אחת מהשיטות הללו מאפשרת לנו לשלוט על ה-Bias-Variance Trade-off.

1. יחסיו הכוחות בין ההטייה והשונות (Bias-Variance Trade-off)

דיברנו רבות על הקשר בין השונות ובין ההטייה. בקצרה, נעשה כאן סקירה שתתחבר גם לפרקים הקודמים. בගודל, מה אנחנו מתחשים כל הזמן? היפותזה שתתאר לנו טוב ככל הנימן את המדגמים הנוכחי ותאפשר לנו לחזות מדגמים נוספים בצורה מדויקת, על סמך המידע הזה.נו, אז מה הבעיה? נוכל לחתוך מחלוקת היפותזה ממש גדולה, שבוטוח תהיה 'קרובה ככל הנימן' לפונקציית התיאוג 'אמיתית' של המדגמים הנוכחי. אכן, ההטייה (Bias) תהיה סופר נמוכה, אבל מושב שהנכنسנו את עצמנו לתוך הפיקסלים של התמונה, כבר לא נוכל לראות את הציור. מה הכוונה? בסוף המטריה שלנו היא לא למצוא היפותזה למוגם שלנו (כאילו כנ), אלא לחזות. ומרוב שנותבון במדגים

²⁸מצטער על זה.

הנוכחי, נמצא את עצמו מתפתלים סביב פונקציה הזיהה שמתארת את המודם הנוכחי (שונות גבואה) ולא נוכל להכליל למודדים נוספים ובכך לחזות.
אם כך, אנחנו רוצים למצוא את נקודת האיזון האידיאלית, בין השונות ובין הטעיה:

1.1 פירוק שגיאת הכללה (Generalization Error Decomposition)

נזכיר כי כבר "פרקנו" שגיאת הכללה בעבר - פירקנו את MSE לרכיבי שונות ותוחלת. אנחנו במודד של להכליל דברים, אז ננסה להכליל זאת לפונקציית הפסד כללית.
תהי $h_S = \arg \min_{h \in \mathcal{H}} L_D(h)$ הייצוא של אלגוריתם הלמידה שלנו. נוכל לפרק את שגיאת הכללה של ההיפותזה שחוזרת מה"לומד" בצורה הבאה:

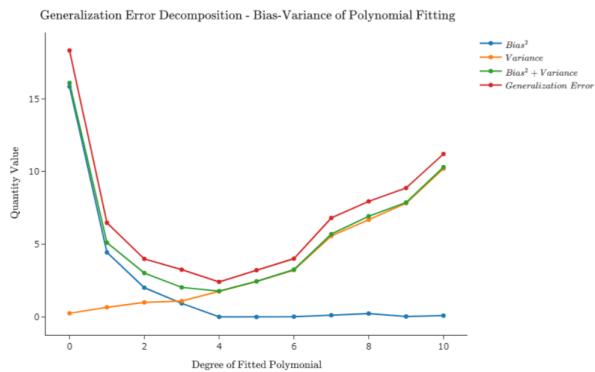
$$L_D(h_S) = \underbrace{L_D(h^*)}_{\varepsilon_{\text{approximation}}} + \underbrace{L_D(h_S) - L_D(h^*)}_{\varepsilon_{\text{estimation}}}$$

כעת, ניגע בכל דבר לנו:

◻ שגיאת הקירוב ($\varepsilon_{\text{approximation}}$) הינה $L_D(h^*)$. מדובר בשגיאת המזערית ביותר שמשיגה היפותזה \mathcal{H} כולה. כלומר, זה לא תלוי במדגם שלנו כלל, ולמעשה אם נגדיל את מחלוקת ההיפותזות יכול להיות שנמצא היפותזה שימושה תוצאה יותר! מעבר לכך, במקרים מסוימים שגיאת זו היא למעשה הטעיה (bias).

◻ שגיאת ההערכתה ($\varepsilon_{\text{estimation}}$) הינה $L_D(h_S) - L_D(h^*)$. כלומר, מדובר על ההפרש בין השגיאת שהושגה במדגם הנוכחי ובין השגיאת הטובה ביותר. זה פן תלוי במדגם הספציפי ובגודלו. שגיאת זו הינה למעשה השונות (variance).

ניתן לראות דבר זה למעשה בגרף הבא, במקרה הספציפי של התאמת פולינומית (polynomial fitting):



2 ועדות ואנסבל (Ensemble/Committee Methods)

לפנינו שוניגש להבין את שיטת האנסטבל, ניגע בדוגמה שתעורר לנו להבין זאת טוב יותר.
נניח שקיימת לנו ועדה עם T חברים, כאשר לכל חבר i יש סיכוי p_i שהוא בוחר את החלטה הנכונה, ו- $1 - p_i$ שהוא לא בוחר את ההחלטה הנכונה. במקרה שלנו נניח כי לכל החברים סיכוי שווה לבחור את ההחלטה הנכונה. ההחלטה שמתקובלת בסופו של דבר הינה לפי כמות הצביעת הדולה ביותר.

דבר זה מוביל אותנו למספר שאלות: מה ההסתברות שנבחרה ההחלטה בסיומו של דבר? מה נחשבת ההחלטה אופיינית (כלומר, מה התוחלת)? כיצד מספר החברים בועדה משפיע על ההחלטה? מה קורה כאשר ההחלטה של כל אחד מהחברים אינה בלתי תליה?

על מנת שנוכל להוכיח שאלה זו, נציג מספר טענות.

למה

יהיו $X_1, \dots, X_T \stackrel{\text{iid}}{\sim} \text{ber}(p)$ שמקבלים ערכים בתחום $\{\pm 1\}$ ונסמן $X = \sum_{i=1}^T x_i$. ההסתברות שהועדה עשויה את ההחלטה הנכונה הינה $\mathbb{P}(X > 0)$.

הוכחה

כיוון שההחלטה של הוועדה תליה למעשה בדעת הרוב, אז علينا לחשב למעשה את הסיכוי שיש יותר חברים שצדוקים מאשר חברים שטוענים.

נניח בה"כ כי התשובה הנכונה הינה +1. כיוון שהתחום הינו $\{-1, 1\}$, נוכל לסמן למעשה את ההחלטה הכלולת בתווך $X = \text{sign}\left(\sum_{i=1}^T X_i\right)$. אם הוועדה צודקת, אז יש יותר חברים שצדוקים ולכון $0 < X < 1$ - ככלומר

אם ההפך, אז כאמור $-1 < X \leq 0$ ולכון מתקבל $0 < X \leq 1$. בקיצור נרצה כי יתקיים כי $\mathbb{P}(X > 0) \geq \frac{1}{2}$, כפי שרצינו.

למה

יהיו $X_1, \dots, X_T \stackrel{\text{iid}}{\sim} \text{ber}(p)$ שמקבלים ערכים בתחום $\{\pm 1\}$ בסיכוי $p > 0.5$, ונסמן $X = \sum_{i=1}^T X_i$. ההסתברות שהועדה החלטה נכונה נקבעה ממלמטה על ידי:

הוכחה

בה"כ, נניח כמו קודם כי התשובה הנכונה הינה +1. אנחנו מעוניינים אם כך (מהלמה הקודמות) לחסום את $\mathbb{P}(X > 0) < \mathbb{P}(X \leq 0)$. נבחן כי עבור $a > 0$ כלשהו מתקאים:

$$\mathbb{P}(X \leq 0) = \mathbb{P}(-aX \geq 0) = \mathbb{P}(e^{-aX} \geq e^0)$$

בשימוש במרקוב²⁹ נקבל:

$$\mathbb{P}(X \leq 0) \leq \mathbb{E}[e^{-aX}] = \mathbb{E}\left[e^{-a\sum_{i=1}^T X_i}\right] \stackrel{\text{iid}}{=} \mathbb{E}[e^{-aX_1}]^T$$

נבחן כי $\mathbb{E}[e^{-aX_1}] \leq 1$ ועל כן נקבל:

$$\mathbb{E}[e^{-aX_1}] = pe^{-a} + (1-p)e^a \leq e^{a-p+pe^{-2a}}$$

²⁹שימוש לב שחרישוב כאן דומה מאוד לחישובים שעשינו בהסתברות לגביה פונקציה יוצרת מומנטים וכו'.

כאשר המעבר הראשון נובע מחישובים של פונקציה יוצרת מומנטים, והשני נובע מא"ש $.1 + x < e^x$ בukt, נשתמש בא"ש הידוע³⁰ עבור בחירה של $a = \frac{1}{2} \ln(x) \geq \frac{x^2}{2} - \frac{1}{2}$ שחויבי עבור $p > 0.5$ ונקבל:

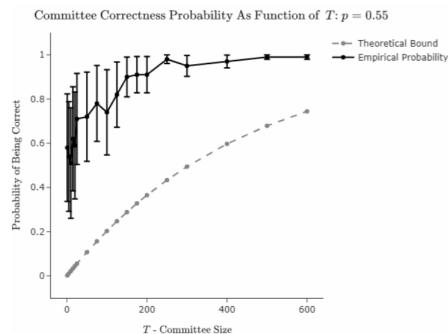
$$\begin{aligned}\mathbb{P}(X \leq 0) &\leq \mathbb{E}[e^{-aX_1}]^T \leq \exp(T(a - p + pe^{-2a})) = e^{T(\frac{1}{2} \ln(2p) - p + \frac{1}{2})} = e^{Tp(-\frac{1}{2p} \ln(\frac{1}{2p}) - 1 + \frac{1}{2p})} \\ &\leq e^{Tp(\frac{1}{2} - \frac{1}{2(2p)^2} - 1 + \frac{1}{2p})} = e^{-\frac{Tp}{2}(\frac{1}{4p^2} - \frac{1}{p} + 1)} = e^{-\frac{Tp}{2}(\frac{1}{2p} - 1)^2} = e^{-\frac{T}{2p}(p - \frac{1}{2})^2}\end{aligned}$$

סך הכל הגיעו איכשהו לביטוי שהוא לכאורה נחמד. ואם נרצה לסכם נקבל:

$$\mathbb{P}(X > 0) = 1 - \mathbb{P}(X \leq 0) \geq 1 - \exp\left(-\frac{T}{2p}\left(p - \frac{1}{2}\right)^2\right)$$

מכאן עולה כי אם הסיכוי שהועדה תבחר נכון גודל מכל סיכוי שמשיחו ספציפי יבחר נכון. מעבר לכך, עולה כי הסיכויים שהועדה תטעהקטנים באופן אספוננציאלי, וביחס לאלגוריתם הלמידה שלנו, זה אומר כי בביטחון של $\delta - 1$ החיזוי שלנו גדול אקספוננציאלי (כלומר מידת הוודאות שלו).

בהתבסס על טענות אלו ניתן להסיק גם כי ככל ש- T גדול (מספר המשתתפים בוועדה), ההסתברות שהועדה תבחר נכון גדלה. כמו כן, ככל שההסתברות של כל אחד מהמשתתפים לצדוק, כך יש פחות 'צורך' לצדוק בהסתברות של 1. ניתן לראות המכחשה לכך בדוגמה הבאה, שמתרארת את ההסתברות האמפירית:



2.1 מנבאים בלתי מתואמים (Uncorrelated Predictors)

תרגיל

יהיו $X_1, \dots, X_T \stackrel{\text{iid}}{\sim} \text{ber}(p)$ משתנים מקרים שמקבלים ערכים על הקבוצה $\{\pm 1\}$ עם $p > 0.5$. מה התוחלת והשונות של $X = \sum_{i=1}^T X_i$

הוכחה

תחילה נחשב את התוחלת והשונות של כל משתנה בודד:

$$\begin{aligned}\mathbb{E}[X_i] &= 1 \cdot \mathbb{P}(X_i = 1) + (-1) \cdot \mathbb{P}(X_i = -1) \\ &= 2p - 1\end{aligned}$$

³⁰רגינורזיות? הוכיחו.

והשונות:

$$\begin{aligned}\text{Var}(X_i) &= \mathbb{E} \left[(X_i - \mathbb{E}[X_i])^2 \right] \\ &= p(1 - (2p - 1))^2 + (1 - p)(-1 - (2p - 1))^2 \\ &= 4p(1 - p)^2 + 4p^2(1 - p) \\ &= 4p(1 - p)\end{aligned}$$

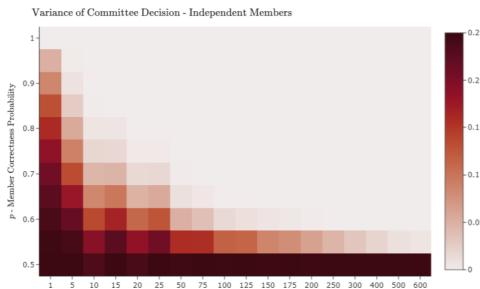
וכעת קיבל עבור X :

$$\mathbb{E}[X] = \frac{1}{T} \sum_{i=1}^T \mathbb{E}[X_i] = 2p - 1$$

והשונות:

$$\text{Var}(X) = \frac{1}{T^2} \text{Var} \left(\sum_{i=1}^T X_i \right) \stackrel{\text{iid}}{=} \frac{1}{T^2} \text{Var}(X_i) = \frac{4}{T} p(1-p)$$

אם כך, התוחלת נשארת אותו הדבר ואילו השונות יורדת בקצב של $(\frac{1}{T})^O$ - נוכל לשמור על אותו דיקן, אך געלה את הביטחון (באמצעות הקטנת השונות):



2.2 מנבאים מתואימים (Correlated Predictor)

מהו לשות ששוב המציגות מכיה על ראשנו - לרוב משתנים מקרים אינם תלויים. נניח כי כל שני משתנים מקרים הינם מתואמים - כלומר בעלי קורלציה של $[0, 1]$.

למה

יהי X_1, \dots, X_T סדרה של משתנים מקרים שווים התפלגות כ- $\text{ש-}\sigma^2$.
i. השונות של החלטות הוועדה הינה $\sigma^2 + \frac{1}{T} (1-p)$.

הוֹכָתָה

נשותמש בנוסחה שראינו בהסתברות לחישוב שונות של משתנים מקרים שאין בלתי מתואמים:

$$\text{Var}(X) = \text{Var}\left(\frac{1}{T} \sum_i X_i\right) = \frac{1}{T^2} \left[\sum_i \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j) \right]$$

נזכר כי הקורלציה (מתאם פירסום) נתונה על ידי:

$$\text{corr}(A, B) = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A) \cdot \text{Var}(B)}}$$

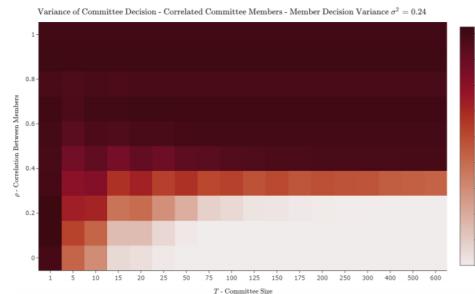
ולכן נקבל כי:

$$\text{Cov}(X_i, X_j) = \text{corr}(X_i, X_j) \sqrt{\text{Var}(X_i) \text{Var}(X_j)} = \rho \sigma^2$$

אם נחבר את הביטויים (נשים לב כי יש $\binom{T}{2}$ אפשרויות לבחירת כל שתי משתנים מקרים):

$$\text{Var}(X) = \frac{1}{T^2} \left[T\sigma^2 + 2 \binom{T}{2} \rho \sigma^2 \right] = \frac{\sigma^2}{T} + \left(1 - \frac{1}{T}\right) \rho \sigma^2 = \rho \sigma^2 + \frac{1}{T}(1 - \rho)\sigma^2$$

מבחן אינטואיציה, זה נראה כך:



קיצור, קיבלנו כי אם $0.5 > \rho > p$ ההסתברות גדלה ככל שמספר האנשים עולה, ומעבר לכך, שההחלטה תהיה הרבה יותר קונסיסטנטית (פחות אנשים שטוענים). אם גם $0 > \rho$ אז ההסתברות עולה עד גבול מסוים.

2.3 שיטות ועדה במערכות לומדות (Committee Methods In Machine Learning)

ונסה לישם את הרעיון שראינו קודם לכן, שאפשר לנו להוריד את ההטיה ואת השונות למדידה. אם נניח כי נסמן כל מדגם S_i בתור X_i , אז למעשה אם נעשה אס T דוגמאות נקבל $X_1, \dots, X_T \stackrel{iid}{\sim} \mathcal{D}^m$, וכיוון שמדובר במשתנים מקרים, גם כלל ההחלטה שיתקבל לכל אחת מהדוגמאות הוא למעשה משתנה מקרי בעל התפלגות כלשהיא. כפי שראינו קודם, ככל שנגדיל את T , אז שגיאת ההחלטה תשאף לאפס (generalization loss). אולם, אנחנו בלמידת אוצה (batch learning) ואין לנו T דוגמאות אלא אחת. אנחנו גם לא יכולים להריץ את \mathcal{A}

על S פעמים, כי תכל"ס נקבל את אותו כלל חיזוי. אז אנחנו קצת מרים וכביכול מייצרים בכל פעם דגימות חדשות מהمدגם שיש לנו. תכף נראה כיצד.

הגדירה

יהי \mathcal{A} אלגוריתם כלשהו שחוזה $\{\pm 1\}$ שיטת ועדת (committee method) מעל \mathcal{A} היא הפונקציה:

$$h(x) = \text{sign} \left(\sum_{t=1}^T t_t(x) \right)$$

הרעיון: לחתה לומד "בסיסי" ולהפעיל אותו על סדרה של T דגימות אימון. מכאן והלאה מתמקד בשני רעיונות שונים עבור בנייתו משתני הוועדה האלו.

3 צבירת אתחול (Bagging)

3.1 אתחול-אווני-נעליים (The Bootstrap)

המודל הראשון אותו נציג שאל מתחום הסטטיסטיקה. מדובר באחד הרעיון החדשניים של הסטטיסטיקה במאה ה-20 - ליצור דגימות מלאכותיות מתוך דוגמה אחת שיש לנו.

בהינתן מרחב נתונים $S = \{(x_i, y_i)\}_{i=1}^m$, נרצה לבנות מרחב מדגם חדש, נקרא לו מדגם bootstrap שננסמו בתור S^{*1} . אנחנו דוגמים m פעמים מתוך S עם חוזרות, ויוצרים למעשה מדגם חדש:

$$S^{*1} = \{(\mathbf{x}_i^{*1}, y_i^{*1})\}_{i=1}^m$$

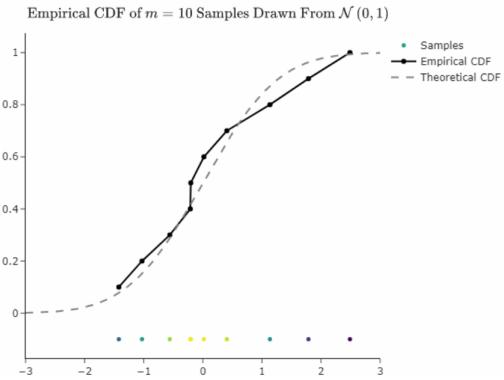
ברור כי יתכן שנקבל חלק מהדges פומים, בשונה מהמדגם המקורי. נחזר על התהליך B פעמים, כשבכל פעם ניקח דגימות באורך m - S^{*1}, \dots, S^{*B} .
בuit, נוכל למעשה לישם את הרעיון שראינו קודם לכן ולהריץ את האלגוריתם שלנו על כל אחת מהדges הללו.
אבל איך תכל"ס זה עובד? הדges נוצרות באופן בלתי תלוי ושוואה התפלגות מפונקציית הסתברות \mathcal{D} מעל $\mathcal{X} \times \mathcal{Y}$. אנחנו מכוונים שאם נבצע bootstrap מעל S זה יתנהג כמו פונקציית הסתברות אמיתית.
בהינתן מדגם S , נוכל להגדיר את ההתפלגות האמפירית $\hat{\mathcal{D}}_S$ שמתකבת מ- S מעל $\mathcal{X} \times \mathcal{Y}$ על ידי:

$$\hat{\mathcal{D}}_S((X, Y) = (x, y)) := \begin{cases} \frac{1}{m} & (x, y) \in S \\ 0 & (x, y) \notin S \end{cases}$$

בצורה דומה, עבור כל $C \subset \mathcal{X} \times \mathcal{Y}$

$$\hat{\mathcal{D}}_S(C) := \frac{|C \cap S|}{m}$$

(אנחנו מניחים כי כל הדגימות ב- S הינן ייחודיות).
 נבהיר כי ככל ש- m גדול יותר ההתפלגות האמפירית \hat{D}_S מתכנסה להתפלגות ל- \tilde{D} . אם כך, כיוון שההבדל בין D - S - \hat{D}_S איננו כה גדול, יוכל באמצעות m דגימות שנוצרות מ- \hat{D}_S ליצור קירוב טוב ל- m דגימות שנוצרות מ- D . ניתן לראות כיצד התפלגות האמפירית מתכנסת להתפלגות התיאורטית ככל ש- m גדול.



3.2 צבירת אתחול - Bagging

הרענון של Bagging הוא למעשה שום קוד לכל מקורה בו נרצה להרחיב דגימות מדגימה נתונה. למעשה, כשהשתמשנו ב-bootstrap השתמשנו למעשה ב-`bagging`.

התחלנו באלגוריתם למידה בסיסי \mathcal{A} ומודם אימון S . לאחר מכן פיתחנו T דגימות bootstrap שמוגדרת על ידי S^{*1}, \dots, S^{*T} , כל אחת מגודל m .
 אנחנו מאמנים את הלומד שלנו על כל אחד מ- T דגימות האימון מסוג ה-bootstrap. לאחר מכן, ניצר את ה"יועדה" (committee) שמוגדרת על ידי $h_{S^{*1}}, \dots, h_{S^{*T}}$ - כלל החלטה שהתקבלו מכל אחת מהדגימות.
 אם נרצה לסוג דגימת מבחן $\mathcal{X} \in \mathcal{X}$, נרים את x על כל אחת מכללי ההחלטה ונסוג בשימוש ב"הצבעת הוועדה":

$$h_{\text{bag}}(x) = \text{sign} \left(\sum_{t=1}^T h_{S^{*t}}(x) \right)$$

אם נרים את ה- bagging הזה על מסווג עצ החלטה, נקבל למעשה:

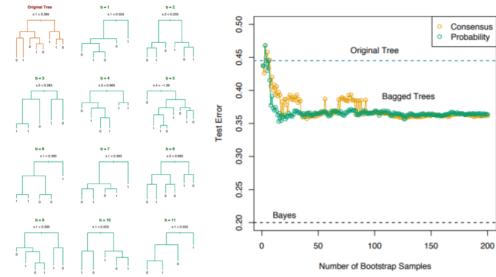


Figure 5.7: Collection of Bagged Decision Trees. (Source: ESL)

נבחן כי האלגוריתם שלנו חייב להיות מסוגל להתמודד עם דוגמאות חזרות. זה יכול לקרות גם ב- S אבל ב-bootstrap זה בטוח יקרה. חלק מהאלגוריתמים מסתובבים עם זה (למשל, רגסיה לוגיסטיבית וליניארית) וחלק ממש סבבה עם זה (למשל עצי החלטה ו-NN- k).

3.3 הפחיתה השונות על ידי bagging

כיוון שאחנו משתמשים ב"עקרון הוועדה", ככלורן דומה למחרינו בוועדה, נצפה לתוצאות דומות - ככלומר להפחיתה השונות ככל ש- T גדול, אך בצורה מסוימת, כתלות בקורסציה שיש בין כלל ההחלטה שיצרנו.

3.4 יערות רנדומליים - Random Forests - ביטול הקורולציה בעצי החלטה ו- Bagging

אנו ממעוניינים אם כך בדרכם להפוך את חברי הוועדה להיות בלתי מתואמים - שההתוצאות יהיו פחות מתואמות. ככל ליקוט כי אם נגביל כל אחד מהלומדים בצורה רנדומלית זה יגרום לאית התאמת ולביצועים טובים יותר. העקרון המוכר ביותר בתחום זה נקרא יערות רנדומליים, או בלאי Random Forests.

הreasון של "יערות רנדומליים" הינו כזה. אנחנו מבצעים bagging על כל אחד מעצי ההחלטה, אך עם זאת חשוב שמבצע את אי ההתאמה: לאלגוריתם ישנו פרמטר $d \leq k$ זה גורם הממד שאחנו עובדים על גבי) כאשר אנחנו יוצרים עץ החלטה, אנחנו בוחרים בצורה איחידה תת קבוצה k מתוך d הפיצרים. אנחנו מבצעים את הפרדה על k הפיצרים בלבד:

אלגוריתם 4 Random Forests

1. תייצר Forest Random בדרכם המוכרת (עומק R , מספר דוגמאות (m_{\min})

2. לכל $1 \leq t \leq T$:

(א) תייצר דוגמאות S^{*t} bootstrap מ-

(ב) תאמן עץ החלטה $h_{S^{*t}}$ על המדגם S^{*t} שקיבלנו:

(ג) אם לא הגיעו לעומק המקסימלי או למספר המינימלי של דוגמאות m_{\min} :

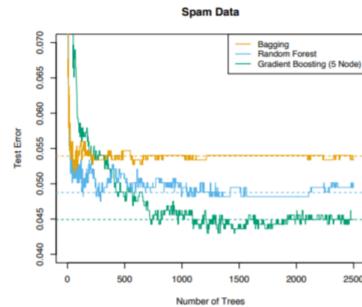
i. תבחר בצורה איחידה k מתוך $[d]$.

ii. תבחר את הפיצ'ר הטוב ביותר מתוך k האיברים.

iii. תבצע את ההפרדה על סמך הפיצ'ר שנבחר.

$$h_S(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^T w_t h_t(\mathbf{x}) \right) \quad 3.$$

זה מגניב, וזה עובד:



כל האציג בבחירה k הוא לבחור $\sqrt{d} = k$. מספיק שנבחר T שאינו גדול מדי, כדי שלא תהיה בעיה חישובית. יש לנו כמה שאלות שמתעוררות מימוש ב-boosting:

◻ האם הוא יכול לפגוע לנו באlgorigths? ובכן, אם הלומד עצמו לא מדהים, נקבל יותר 'טפסים' זה בטוח לא יעוז לנו.

◻ מה החסרונות של boost? כמובן שהוא לוקח הרבה זמן ריצה - אימון של T מוגמים ושמירת מקום של T מוגדים. בנוסף, קשה להשיג מהמודל מידע (interpretability) בגלל הרעיון של "החלטת הוועדה". עליינו לבדוק כל אחת מההחלטה שהתקבלו כביכול.

◻ אנחנו יכולים ליצור את T הדגימות ב-boost בצורה מקבילה - הם לא צריכים אחד את השני בשיבול לעבוד יחד.

◻ אנחנו לא יכולים להשתמש ביחסיות מספר חברי הוועדה שהצביעו כהסתברות, כיון שדבר זה מעיריך לו את $\mathbb{P}\{Y = +1 | X = x\} - \text{ולא את } \mathbb{P}\{h_S(x) = +1\}$.

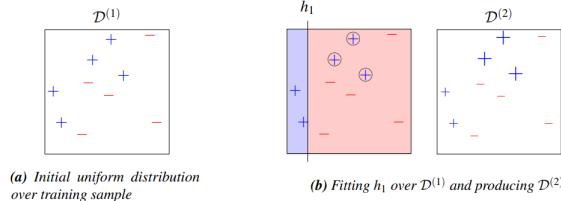
4 הגראה - Boosting

קודם לכן רימינו ולקחנו דוגמאות מתוך הדגימות המקוריות. הרעיוןicut הוא שונה. אנחנו מוגברים (notifiers) אלגוריתם קיים. איך אנחנו עושים זאת? במקרה של המציג דוגמאות מתוך המודם המקוריים, אנחנו מעמידים פנים שיש לנו T הסתברויות שונות שמתוכם מרחב המודם נוצר.

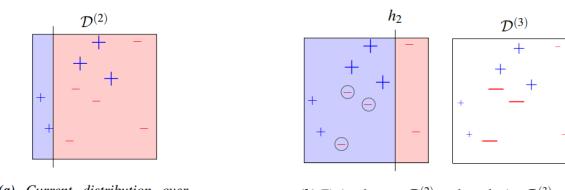
כbicool, כל h_t - כלל החלטה - הוא תוצאה של הרצת האלגוריתם על מוגם אימון S_t כך שככל מוגם נוצר מהסתברות \mathcal{D}^t שונה כbicool. במקרה זה 'חברי הוועדה' מאומנים ברצף - יש בזה יתרון. כיצד בדיקע עושים זאת? תקף נראה. אבל היתרונו הוא שאחרי כל סיום אימון של h_t , נוכל לעדכן את הסתברות במקומות בהם h_t טעה. ככלומר, h_{t+1} יזהה טוב יותר.

בגדייל, קצת עבדנו עליום כשהשאנו לוקח מודל אחר מ-*bootstrap*. מה הכוונה? למעשה ב-boosting אנחנו משתמשים ב-*bootstrap* ממושך. נוכל ליצור את המשקל באמצעות ERM למשל.

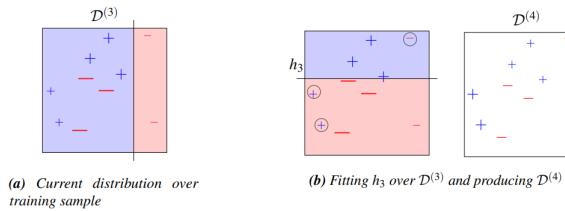
חשוב להבין את הרעיון קודם כל אינטואטיבית. נתבונן בדוגמה הבאה. בתיאלה, כל הדגימות שוות הסתברות - ישנה הסתברות איחודית על מנת לקחת m דוגמאות מתוך m המוגם המקורי. נניח שיצאנו כל החלטה, וקיים 3 דוגמאות שהין שגויות, באמצעות ה-ERM למשל נגלה זאת. 'עדכן' את הסתברות ונגדיל את המשקל של הדגימות הגרועות, כך שהן מתקבלנה בהסתברות גבוהה יותר:



icut, יש לנו הסתברות חדשה $\mathcal{D}^{(2)}$ וכעת אנחנו מיצרים כלל החלטה h_2 . שוב אנחנו מבצעים הגדלה של הסתברות של הדגימות שהתקבלו:



ובפעם הבאה:



אם כך, אם נשתמש בכל אחד מכללי החלטה נוכל לקבל סיווג מורכב הרבה יותר ממושог בודד:

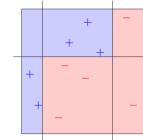


Figure 5.12: Boosting illustration: Decision boundaries of the ensemble of classifiers h_1, h_2, h_3 .

למרות שלכל אחד ממהמסוגים יש כלל החלטה פשוט, האנסטבל מאפשר לנו לתאר מקרים בהם המידע המידיע מורכב הרבה יותר. הסיווג מתבצע באמצעות:

$$h_{\text{boost}}(\mathbf{x}) := \text{sign} \left(\sum w_i h_i(\mathbf{x}) \right)$$

כאשר למעשה w_i מתייחס למשקל של כל מסווג, שמשקף כמה מוצלח הוא היה.

4.1 אלגוריתם adaBoost

האלגוריתם המקורי של boosting ידוע בתואר Adaptive Boosting. נסתכל כיצד זה הולך:

אלגוריתם 5 adaBoost

1. תהפוך את כל הדגימות להיות איחודות - $\mathcal{D}^{(1)} \leftarrow (\frac{1}{m}, \dots, \frac{1}{m})$

2. עבור $:1 \leq t \leq T$

(א) קח את הלומד הפשוט $h_t = \mathcal{A}(\mathcal{D}^{(t)}, S)$

(ב) מחשב את $\varepsilon_t = \sum_{i=1}^m \mathcal{D}_i^{(t)} \mathbb{1}[y_i \neq h_t(\mathbf{x}_i)]$

(ג) تعدכו $w_t = \frac{1}{2} \left(\ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right) \right)$

(ד) تعدכן את ההסתברות של כל אחד מהדגימות להיות $\mathcal{D}_i^{(t+1)} = \mathcal{D}_i^{(t)} \exp(-y_i \cdot w_t h_t(\mathbf{x}_i))$

(ה) תנורמל את המשקלים להיות $\mathcal{D}_i^{(t+1)} = \frac{\mathcal{D}_i^{(t+1)}}{\sum_{j=1}^m \mathcal{D}_j^{(t+1)}}$

3. תחזיר את $h_s(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^T w_i h_t(\mathbf{x}) \right)$

הרעין הינו לכך: אנחנו מורידים את המשקלים של הערכים שסווגו נכונה. אנחנו מעוניינים להפוך את בעיית הסיווג ל-'סופר קשה', במובן שהסיכון האמפירי הממושך של h_t ביחס להסתברות \mathcal{D}^{t+1} יהיה הגרוע שאפשר - $\frac{1}{2}$ (כלומר, יקרה המונע פעמיים).

טענה

עבור פקטור המשקל $\varepsilon_t = \sum_{i=1}^m \mathcal{D}_i^t \mathbb{1}[y_i = h_t(\mathbf{x}_i)]$ ו- $w_t = \frac{1}{2} \ln(e_t^{-1} - 1)$ ביחס h_t ביחס \mathcal{D}^{t+1} , יהיה $\sum_{i=1}^m \mathcal{D}_i^{t+1} \mathbb{1}[y_i \neq h_t(\mathbf{x}_i)] = \frac{1}{2}$

$$\sum_{i=1}^m \mathcal{D}_i^{t+1} \cdot \mathbb{1}[y_i \neq h_t(\mathbf{x}_i)] = \frac{1}{2}$$

הוכחה

נחשב זאת באופן ישיר:³¹

³¹המעברים לא ברורים כיון שישנן מספר דילוגים בהצבות. כיוון שהוא מסתדר בסוף, תאמינו שהוא נכון ונכו שבחור את ה- ε זהה ידע מה הוא עושה.

$$\begin{aligned}
 & \sum_{i=1}^m \mathcal{D}_i^{t+1} \cdot \mathbb{1}[y_i \neq h(\mathbf{x}_i)] = \\
 & \frac{\sum_{i=1}^m \mathcal{D}_i^{(t)} \exp(-y_i \cdot w_t h_t(\mathbf{x}_i)) \mathbb{1}[y_i \neq h(\mathbf{x}_i)]}{\sum_{j=1}^m \mathcal{D}_j^{(t+1)} \exp(-y_j \cdot w_t h_t(\mathbf{x}_j))} = \\
 & \frac{\exp(w_t \varepsilon_t)}{\exp(w_t \varepsilon_t) + \exp(w_t \varepsilon_t)(1 - \varepsilon_t)} = \\
 & \frac{\varepsilon_t}{\varepsilon_t + \exp(-2w_t)(1 - \varepsilon)} = \\
 & \frac{\varepsilon_t}{\varepsilon_t + \frac{\varepsilon_t}{1 - \varepsilon_t}(1 - \varepsilon)} = \frac{1}{2}
 \end{aligned}$$

4.2 למידת PAC, למידה חלשה (Weak Learnability)

למעשה, מבחינה היסטורית, boosting נועד על מנת לתת תשובה ל"לומד חלש".

הגדרה

אלגוריתם לומד \mathcal{A} הוא לומד γ חלש (γ -weak-learner) עבור מחלקה היפואתית \mathcal{H} , אם ישנה פונקציה $\mathbb{N} \rightarrow m_{\mathcal{H}}(0, 1)$ כך ש:

□ לכל $\delta \in (0, 1)$.

□ לכל הסטברות \mathcal{D} מעל מרחב המודג \mathcal{X} .

□ לכל פונקציות תיוג $f : \mathcal{X} \rightarrow \{\pm 1\}$.

אם מתקיימת הנחת הריאלבליות ביחס $\mathcal{L}, \mathcal{D}, f, \mathcal{H}$, כאשר נריץ את \mathcal{A} על $m \geq m_{\mathcal{H}}(0, 1)$ דוגמאות אימון שווות התפלגות ובלתי תלויות שנוצרות מ- \mathcal{D} ומתיוגת על ידי f , האלגוריתם יחזיר היפואת $(S, h_S = \mathcal{A}(S))$, כך שבסתברות של לפחות $\delta - 1$ נקבע כי $L_{\mathcal{D}, f}(h_S) \leq \frac{1}{2}$.

הגדרה

מחלקה היפואתית \mathcal{H} הינה ניתנת ללמידה γ חלשה (γ -weak-learnable) אם ישנו לומד γ חלש עבור \mathcal{H} .

ההבדל בין זה ובין למידת PAC הוא כי שם אנחנו יכולים לבחור איזה ε שבא לנו. כאן זה רק עבור γ לא בטוח שמצא לומד טוב יותר.

אם כך, מה הקשר בין הלומד החלש ובין boosting? נניח כי יש לנו \mathcal{H} שהוא ניתן ללמידה PAC. ראיינו כבר כי אם השתמש ב- $\text{ERM}_{\mathcal{H}}$, נצליח לסwoג היטב. אבל מה אם זה מאד מורכב חישובי? נוכל לקחת מחלקה היפואתית פשוטה $\mathcal{H}_{\text{base}}$ שקל לחשב את ה- $\text{ERM}_{\mathcal{H}_{\text{base}}}$ שלה, ושיינו לומד γ חלש (כל לנו חשיבות ללמידה עם דיוק של $\gamma - \frac{1}{2}$), ונרצה לחפש דרך להגבר את ה- $\text{ERM}_{\mathcal{H}_{\text{base}}}$ וליצור לומד \mathcal{A} שדומה ל- $\text{ERM}_{\mathcal{H}}$ מעל \mathcal{H} .

אם נתבונן למשל בעצם החלטה. ראיינו כי ERM לא כימי לחישוב. אבל עם עז פשוט, זה יותר קל. וכך נוכל להשתמש ב-adaBoost.

טענה

יהי S מודגם אימון. נניח כי בכל הריצה של adaBoost, הלומד הבסיסי מוחזיר כלל חיזוי, שהסיכון האמפירי הממושקל שלנו מקיים:

$$\sum_{i=1}^m \mathcal{D}_i^{t+1} \cdot \mathbb{1}[y_i \neq h(\mathbf{x}_i)] \leq \frac{1}{2} - \gamma$$

אזי הסיכון האמפירי של כלל החיזוי של adaBoost, h_{boost} , מקיים:

$$L_S(h_{\text{boost}}) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}[y_i \neq h_{\text{boost}}(\mathbf{x}_i)] \leq \exp(-2\gamma^2 T)$$

4.3 הטייה וסוגות ב-boosting

אנחנו מעריכים כי הסיכון האמפירי הנמוך ייתן לנו שגיאות הכללה נמוכה. האם זה באמת קורה? למעשה, הריצה של T איטרציות של adaBoost תחזיר לנו פונקציה מחלוקת ההיפותזה:

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

מדובר בחלוקת ההיפותזה של צירופים קמורים של פונקציות מ- $\mathcal{H}_{\text{base}}$. אם כך, ברור שככל ש- T גדול, גם \mathcal{H}_T גדול (④) אבל! לא כל כך מהר (⑤). אם הייתה לנו דרך קנוונית להציג את "הגודל" של \mathcal{H} , אזי היינו יכולים להגיד כי \mathcal{H}_T היו בהערכת גסה ($\mathcal{H}_{\text{base}} \cdot \text{VCdim}(\mathcal{H}_{\text{base}})$). כמובן, אכן השונות גדולה, אבל לא בצורה מאוד דרמטית כמו שחששנו (אנחנו לא אוהבים אקספוננציאלי).

מайдעך, ברור לנו כי ההטייה יורדת! ניתן לראות זאת מכך שהסיכון האמפירי יורדת שכפי שראינו קשור להטיה. מעבר לכך, העובדה שהסיכון האמפירי יורד בצורה אקספוננציאלית (פתרונות זה טוב לנו, אה?) אנחנו רואים שהוא יורד גם די מהר. אם כך, רואים כי ההטייה יורדת מהר יותר מהשונות, כלומר שגיאות הכללה משתפרת די דרמטית. מה שנשאר לנו לוודא זה האם בשימוש ב- T גדול מדי יתרבצע overfit (ההתאהבות שלנו במדגם, הכנישה לפיקסלים, כל זה).

בגדר, אפשר לראות כי ככל שההיפותזה הבסיסית יותר פשוטה, ההגברת שלה (boosting) משיגת תוצאות טובות יותר.

חלק VI

רגולרייזציה ובחירה מודל

1 רגולרייזציה

נבחן כי בפרקimos הקודמים התעסקנו רבות בבחירה מודל $\mathcal{H} \in \mathcal{H}_s$. ברוב המקרים התבססו על מזעור פונקציית משקל \mathcal{F}_S כלשהי מעל $\mathcal{H} \in \mathcal{H}$. ככלומר למשה $h_S = \mathcal{A}_0(S)$ נתונה על ידי:

$$h_S := \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{F}_S(h)$$

למעשה, פונקציה זו מאפשרת לנו למדוד כמה h תואם את המודגם שלנו. נוכל לקרוא לפונקציה זו בתור **מדד המהימנות** שלנו (fidelity term). ראיינו כמה פונקציות כאלה בעבר: למשל, RSS ברגression ליניארית, ברגression לוגיסטיבית הנראות (שנרצה למקסם) וב-SVM השתמשנו בשול של העל מישור.

כמובן שהדרך הכי קלה למאער היא באמצעות ERM. בכל לומד שmobוס על ERM אנחנו מגדירים פונקציית הפסד $\ell(\cdot, \cdot)$ שמדירה את הסיכון האמפירי שמתתקבל על ידי ℓ , ביחס למרחב המודגם S :

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h(\mathbf{x}_i), y_i)$$

כלומר, במקרה זה **מדד המהימנות** שלנו הוא $L_S(h)$.

כמו כן, דיברנו בעבר על העשירות והאסתטיקות של מחלקות היפותזה, שימושיות על הטורייד אוף המפורסם. ראיינו כי לא בא לנו שהמודל יתאים במיליון אחוז למודגם האימון שלנו (במקרה שהוא מסתובך) - כי אנחנו חוששים שלא נצליח לחזות.

אחד הדריכים (הדי מבאסוט יש לציין) שהשתמשו בהן על מנת להתמודד עם בעיה זו, הייתה להקטין את גודל \mathcal{H} - אבל אנחנו מגבילים את עצמנו. היינו רוצים למצוא דרך חכמה שבה נעדייף אמן היפותזה פשוטות, אבל במקרה בו יש היפותזה מורכבת שווה אותה זה, ניקח אותה. כיצד נעשה זאת? נציב תנאי נוסף לבעה. נבחר $0 \geq \lambda$ ונגידיר לומד $\mathcal{H} : S \rightarrow \mathcal{A}_\lambda$:

$$h_S := \left(\operatorname{argmin}_{h \in \mathcal{H}} \mathcal{F}_S(h) + \lambda \mathcal{R}(h) \right)$$

הביטוי \mathcal{R} נקרא "מקדם הרגולרייזציה" regularization term. אם נבחר את \mathcal{R} בצורה חכמה, אז נוכל למצוא באמצעותו את ה'סיבוכיות' של היפותזה כלשהי: ככל שהיא היפותזה תהיה מורכבת יותר, אז $\mathcal{R}(h)$ יהיה גדול. אם כך, אם מזעיר את הביטוי נקבל מצד אחד ביטוי טוב יותר למודגם-היפותזה מורכבת יותר- (h) קטן, \mathcal{F}_S

ומאידך (h) יגדל. כך גם באופן הפוך (ומקביל).³²
אם כך, λ שולטות למעטה בטרוייד-אורוף זה:

□ אם $0 = \lambda$, אין רגוליזציה כלל, ונשאר כמו קודם.

□ אם $\infty \rightarrow \lambda$, אז נרצה למצוא פשטוט את הפונקציה המכונה פשוטה $h \in \mathcal{H}$.

□ כל $(\infty, 0) \in \lambda$ למעטה מוגדר trade-off ספציפי, בין הצורך במינימנות h ובין הצורך בפשטות h .

2 רגוליזציה עצי החלטה

כשעסקנו בעצי החלטה, רأינו כיצד לבנות עץ החלטה עמוק של מקסימום k באמצעות עקרון ERM. רأינו גם כי יש חשש של אוברפיט של המידע אם מגדילים ממש את k . הנה וنمמש רעיון זה כתע, בשביל להימנע מאוברפיט.

עצי רגרסיה

תחילה, נציג את הרעיון של האלגוריתם לעצי רגרסיה. למעשה, כל סיווג מבוסס על חלוקה זרה B_j של \mathbb{R}^d ל- N תיבות ציריים מקבילים (axis-parallel boxes) $[C] \rightarrow h \in \mathcal{H}_{ct}$ כאשר כל עץ h הוא פונקציה $: \mathbb{R}^d \rightarrow [C]$ שמוגדרת על ידי

$$h(\mathbf{x}) = \sum_{j=1}^N c_j \mathbb{1}[\mathbf{x} \in B_j]$$

כאשר $c_j \in [C]$ הוא התגית של כל תיבה j . במקרה של עצי הרגרסיה c_j לוקח ערכים ב- \mathbb{R} במקום ב-{ $-$ }.
נבוד ממש בדומה לעצי סיווג ונרצה - בהינתן מודם S - למצוא מחלקה היפותזה \mathcal{H}_{RM} שמצוירת את הסיכון האמפירי. זה מרכיב חשוב, שכן נסתמך על דרך שבוססת על מודם האימון.

בהינתן B_j כמו קודם, הסיכון האמפירי ביחס להפסד הריבועי יהיה

$$c_j := \underset{c \in \mathbb{R}}{\operatorname{argmin}} \sum_{i|\mathbf{x}_i \in B_j} (y_i - c)^2 = \frac{1}{|B_j|} \sum_{i|\mathbf{x}_i \in B_j} y_i$$

דבר זה נובע מכך שאם נבחרו את c המינימלי ונשווה לו נקבל:

$$\begin{aligned} f'(c) &= \sum_{i=1}^{|B_j|} 2(y_i - c) = 2 \sum_{i=1}^{|B_j|} (y_i - c) = 0 \rightarrow \\ &\rightarrow |B_j| y_i - \sum_{i=1}^{|B_j|} c = 0 \rightarrow |B_j| y_i = \sum_{i=1}^{|B_j|} c \rightarrow \\ &y_i = \frac{1}{|B_j|} \cdot \sum_{i=1}^{|B_j|} c \end{aligned}$$

כעת, אם נתאים את אלגוריתם CART באמצעות החלפת שגיאה הסיווג בהפסד הריבועי, נקבל כתוצאה מהביטוי הקומס:

³²למעטה כבר רأינו זאת ב-SVM - לא ידעו שקוראים אלה לכך. אבל $\frac{1}{m} \sum_{i=1}^m \xi_i$ היה מקדם הרגוליזציה. ככל שהביטוי הזה היה קטן יותר, אז ההיפותזה פשוטה יותר ו'מאפשרת' פחות הפרות של השול.

$$\hat{y}_S = \frac{1}{|B|} \sum_{x_i \in B} y_i$$

גיזום עצי החלטת CART

האלגוריתם CART מסתים בעץ מיאוזן. כאשר אנחנו יוצרים את העץ, בכל פעם שאנחנו חותכים את העץ, אנחנו מורידים את ההטייה אבל גם מגבירים את השונות. במהלך הבדיקה של האלגוריתם עליינו 'לגוזם' את העץ. הכוונה היא להוריד את גודל העץ באמצעות איחוד מספר קופסאות יחד. כך נתמודד עם הטרידי-אוף הידעו לשם זה.

נניח ויש לנו מדגם S ועץ וגרסיה T שנוצר בצורה דומה עם חלוקה עם B_j כמו קודם. הסיכון האמפירי של T נתון על ידי:

$$L_S(T) = \sum_{j=1}^N \sum_{i|x_i \in B_j} (y_i - \hat{y}_S(B_j))^2$$

זהו למעשה מועד המהימנות (T). מועד הרגולרייזציה (T) יהיה פשוט $\mathcal{R}(T) = |T|$ - מספר העלים ב- T . בנוסף, עברו T_0 - עץ מלא שנוצר באמצעות CART, נסמן $T \subset T_0$ אם T הוא תת עץ של T_0 ואפשר להשיגו באמצעות איחוד קופסאות ב- T_0 . אז, בעיית האופטימיזציה לאחר רגולרייזציה תהיה:

$$T^* = \underset{T \subset T_0}{\operatorname{argmin}} L_S(T) + \lambda |T|$$

כלומר, מבין כל תת עץ הקימיים, אנחנו מוחפשים אחרי האופטימלית. שימוש לב כי אנחנו לא מבצעים אופטימיזציה על כל מחלקת ההיפותזה - דבר שיכול להיות לא פיזבלי, אלא רק על העצים שנוצרו באמצעות CART.

3 רגולרייזציה וגרסיה

3.1 בחירת תת קבוצות

נזכר ברגסיה ליניארית והאומדן LS . הגדרנו את מחלוקת ההיפותזות בטור:

$$\mathcal{H}_{reg} := \left\{ h(x_1, \dots, x_d) = w_0 + \sum_{i=1}^d x_i w_i \mid w_0, w_1, \dots, w_d \in \mathbb{R} \right\}$$

כעת, בהינתן y, x , נבחר $h_S \in \mathcal{H}_{reg}$ באמצעות מציאת האומדן LS - אותו קיבלנו באמצעות ERM והחפסד הריבועי או עקרון הנראות המירבית. נזכר כי דרשו שייהיו יותר דוגמאות מפיצרים.

אמנם, לעיתים מספר הפיצ'רים d יכול להיות מאד גדול, במיוחד כשוייצרים אותם בצורה אוטומטית. במקרה זה, ישנו פיצ'רים שיש בינם קורלציה. בעת הגרסיה הליניארית לא תעבור כי יש יותר מדי פיצ'רים - מרבב הפתרונות גדול מ-1.

אם כך, נרצה למצוא דרך בה נשמר רק חלק מהפיצ'רים. יהיו $k \leq d$ מספר הפיצ'רים הרצויים. לכן נדרש לפתור את בעיית האופטימיזציה הבאה:

$$\begin{array}{ll} \text{minimize}_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} & \|w_0 \mathbf{1} + \mathbf{X}w - \mathbf{y}\|^2 \\ \text{to subject} & \|w\|_0 \leq k \end{array}$$

כאשר $[w_i \neq 0] = \sum_i \mathbf{1}[w_i \neq 0]$. כמובן, נרצה למצוא וקטור מקדים w ששמזעיף את RSS על k פיצ'רים. זה מה שקרה למה באלגוריתם זה.

אלגוריתם 6 בחירת תת הקבוצה הטובה ביותר

1. לכל $[d] \subset S$ כאשר $|S| = k$:

(א) תתאים (fit) מודל גרסיה $\hat{w}^{(S)}$ ונחשב את $RSS(\hat{w}^{(S)})$

2. תחזר את $(S^*, \hat{w}^{(S^*)})$ כאשר S^* הוא $RSS(\hat{w}^{(S)})$ המינימלי.

כיוון שהאלגוריתם נותן מעט מידע על סיבוכיות ההיפותזה, נוכל לחשב על משפחת לומדים $\{\mathcal{A}_k^{\text{best-subset}} \mid 0 \leq k \leq d\}$ כאשר k שולט למעשה על הטריד-אוף המפרנס. אנחנו רוצים למצוא את ה- k האידיאלי. natürlich, אפשר בשביל ערך בודד, נctrיך לעבור על $\binom{d}{k}$ אפשרויות לפיצ'רים - בעיית NP קשה. אם נשנה את הנוסחה לעיליה לבעה קמורה, זה יאפשר לנו למצוא קירובים טובים לבעה זו. הבעיה הבאה קשורה לבעה הנוכחית אבל לא שcolaה אליה:

$$\hat{w}_\lambda^{\text{subset}} := \underset{w_0 \in \mathbb{R}, w \in \mathbb{R}^d}{\operatorname{argmin}} \|w_0 \mathbf{1} + \mathbf{X}w - \mathbf{y}\|^2 + \lambda \|w\|_0$$

כאשר $0 \geq \lambda$. בשתי הביעות הללו לא כללו את intercept. (הוא אינו חלק מהפיצ'רים).

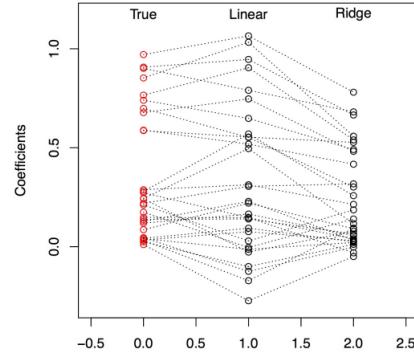
3.2 Ridge Regression

נוסף לביעות החישוביות שראינו בבחירה תת הקבוצה הטובה ביותר, הוא עלול לסייע במקרים גבואה, תלוי בפיצ'רים שבחרנו או לא בחרנו במדגם הנוכחי. על מנת להתמודד עם שתי בעיות אלו, השתמש בשתי דרכי ציוק, שיאפשרו לנו להגביל את הערכיהם ולפעמים לצמצם אותם לכיוון 0.

דרך אחת האפשרית הינה גרסיה Ridge regression, באמצעות כפיטת נורמה על וקטור המקדים:

$$\hat{\mathbf{w}}_{\lambda}^{\text{ridge}} := \underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|w_0 \mathbf{1} + \mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_2^2 \quad \lambda \geq 0$$

למעשה, הפרמטר $\lambda \geq 0$ הוא פרמטר הרגולרייזציה ששולט ב'כמויות הצימוק'. עבור $\lambda = 0$ קיבל את הפתרון של 'הרביעים הפלות' (least squares). כאשר $\lambda \rightarrow \infty$ → אנחנו 'פוגעים' יותר ויותר בגודל וקטור המקדמים, ברמה שגורמת לכך ש- $\hat{\mathbf{w}}_{\lambda}^{\text{ridge}} \rightarrow 0$.



ניתן לראות את ההבדלים בין 'הצגה הליניארית' ובין הצגת 'הרכס'.
משמעותו לב כי בעיית ridge היא בעיית אופטימיזציה קמורה שניתן לפתור את עם תכנון ריבועי. אמנם, במקרה שלנו אין צורך כיון שאנו יכולים להשיג 'נוסחה סגורה' עבור הממערך.

טענה

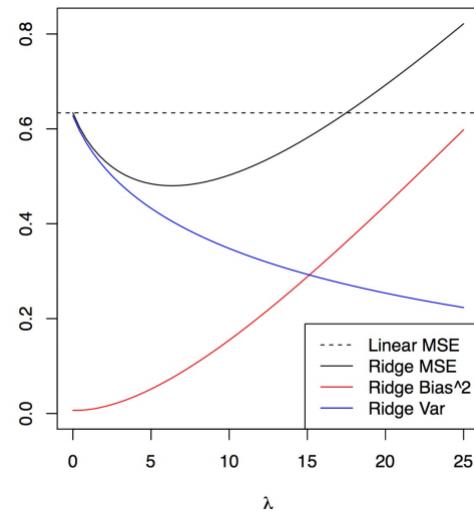
יהו \mathbf{y} , \mathbf{X} , בעיית גרסיה, $\lambda \geq 0$. יהי $\mathbf{x} = U\Sigma V^T$ של \mathbf{X} .
אומדן הרכס (Ridge estimator) נתון על ידי:

$$\hat{\mathbf{w}}_{\lambda}^{\text{ridge}} = U\Sigma_{\lambda}V^T\mathbf{y}, \quad [\Sigma_{\lambda}]_{ii} = \frac{\sigma_i}{\sigma_i^2 + \lambda}$$

תחליה, נחליף את \mathbf{X} ב-SVD :

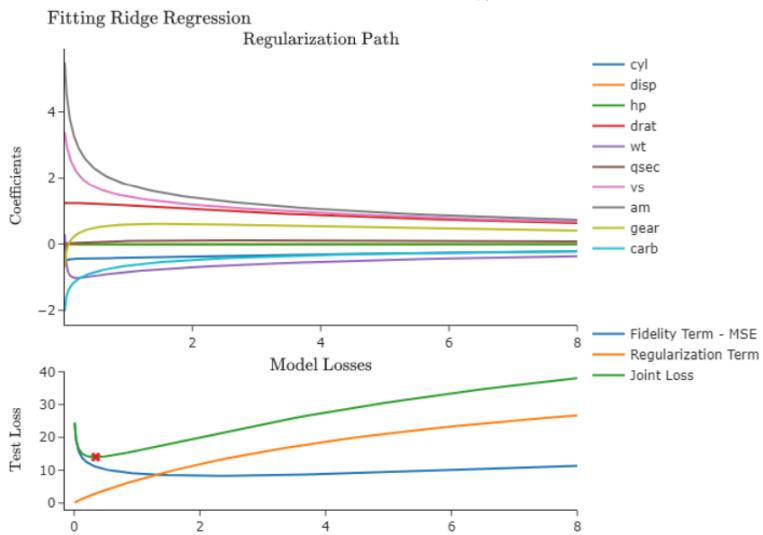
$$\begin{aligned} \hat{\mathbf{w}}_{\lambda}^{\text{ridge}} &= (\mathbf{X}^T \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{V} \Sigma^T \mathbf{U}^T \cdot \mathbf{U} \Sigma \mathbf{V}^T + \lambda I_d)^{-1} \mathbf{X}^T \mathbf{y} \\ &\square = (\mathbf{V} \Sigma^T \Sigma \mathbf{V}^T + \lambda I_d)^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{V} (\Sigma^T \Sigma + \lambda I) \mathbf{V}^T)^{-1} \mathbf{X}^T \mathbf{y} \\ &\square = \mathbf{V} (\Sigma^T \Sigma + \lambda I)^{-1} \mathbf{V}^T \cdot \mathbf{X}^T \mathbf{y} = \mathbf{V} (\Sigma^T \Sigma + \lambda I)^{-1} \mathbf{V}^T \cdot \mathbf{V} \Sigma^T \mathbf{U}^T \mathbf{y} \\ &\square = \mathbf{V} (\Sigma^T \Sigma + \lambda I)^{-1} \Sigma^T \mathbf{U}^T \mathbf{y} = \mathbf{V} \Sigma_{\lambda} \mathbf{U}^T \mathbf{y} \end{aligned}$$

אם נבחר λ חיובי בהחלט, אז נוכל להראות כי הפתרון הינו מהצורה של $\mathbf{y}^T \mathbf{X}^T \mathbf{X} + \lambda I_d$, כאשר המטריצה $\lambda I_d + \mathbf{X}^T \mathbf{X}$ בהכרח הפיכה. בשונה מהאומדן לריביעים הפלות (LSE) שראינו שאנו מוטה, אומדן זה הינה מוטה. אמנם הוא מוריד את השונות מסווגים כך שטח הכל ה-MSE נמוך ביחס לריביעים הפלות.

**דרך הרגולרייזציה**

משמעותו לעקבות אחרי כל אחד מהמשקלים \hat{w}_i כאשר λ משתנה. בתמונה הבאה נראה כיצד כל אחד מהמערכים משתנה לפי ערכיו λ .

שימוש לב כיצד המשקל מתחילה ממישקל הרגסיה הליניארית (כלומר, כאשר $0 = \lambda$), ומצטמכים והולכים להם לכיוון אפס, בערך בסדר גודל של $\frac{1}{\lambda}$, ככל ש- λ גדל:

**3.3 רגולרייזית לאסו (Lasso) - נורמת ℓ_1**

למרות שהצלחנו לעשות כמה דברים יפים עם Ridge, מה לשות, הוא לא בדיק עושה את מה שבחריתת תת הקבוצות הטובה ביותר עשו. ככלומר, הוא לא יכול לבחור לנו את תת הקבוצה הטובה ביותר ביותר לרוגסיה.

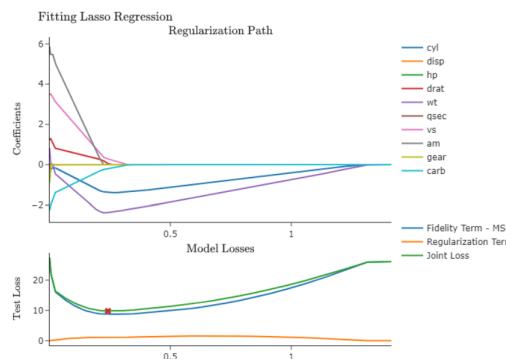
הreasון של Lasso או בקיצור Least Absolute Shrinkage and Selection Operator מנסה להשיג בדיק את זה.

רגולרייזציה זאת משתמשתBNORMATA ℓ_1 במקומBNORMATA ℓ_2 :

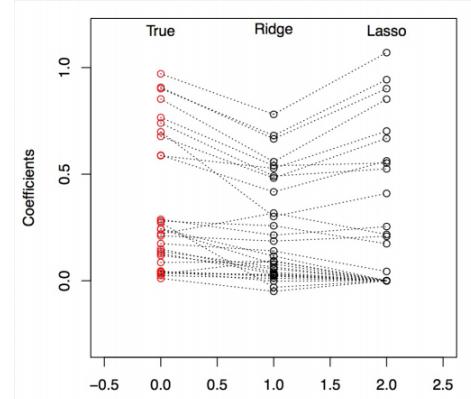
$$\widehat{\mathbf{w}}_{\lambda}^{\text{lasso}} := \underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|w_0 \mathbf{1} + \mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|_1 \quad \lambda \geq 0$$

מה שונה כאן מה Ridge הוא כי דרך הצימוק שונה מאוד בין שתי הביעות. הפיתרון בו משתמש לאסו מאוד דليل (sparse). כמובן, יש גם אפסים בוקטור שמתකבל ממנה. למעשה, ככל ש- λ גדול³⁴, אין יהיו יותר אפסים בפיתרון ועל ידי כך ניתן להחליט איזה פיצ'ר יהיה במודל ואיזה לא - אם יש פיתרון שמקבל 0, הוא לא יהיה במודל. אם כך, יש למודל זה יתרון משמעותי בפרשנות של הפיצ'רים ובהחלטה לגבים.

כמו כן, בעיית האופטימיזציה של לאסו, היא בעיה קמורה ובפרט ניתן לפתור אותה עם תכנון ריבועי. יש גם דרכים אחרים לפתור את הבעיה, בדומה ל-Ridge. בכלל מקרה, כך נראה התמונה של המודל:

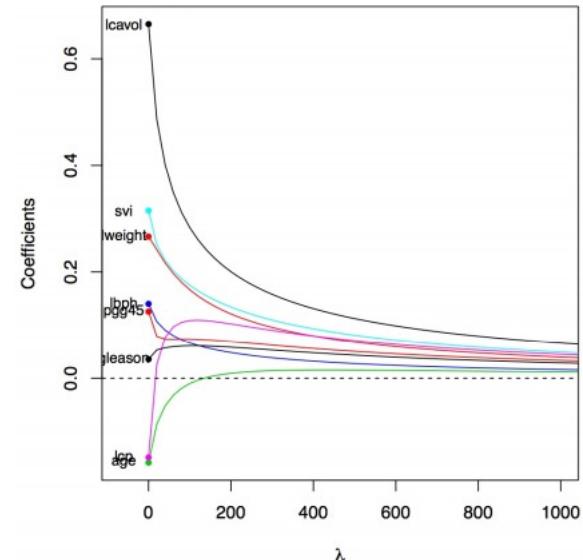


ותמונה נוספת, ליחסיות מול Ridge:

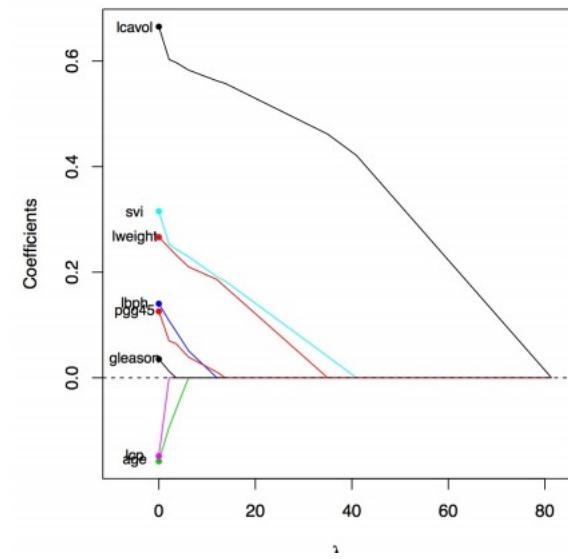


או כך ברידג':

³³ככלל, המבנה של נורמת p הינו מהצורה $(\sum x^p)^{\frac{1}{p}}$.
³⁴האינטואיציה והקשר בין שתי הביעות נתונה בהמשך. בכלל, היא מותקנרת לכדו היחידה של הנורמות הללו.



לעומת הלאסו:

**קמיירות מול צלילות (Convexity vs. Sparsity)**

ראינו למעשה שלושה דרכי רגולרייזציה שמתאימות לנורמות $\|\cdot\|_2$, $\|\cdot\|_1$, $\|\cdot\|_0$. ראיינו כי ב諾רמות 1 ו-2 מדובר בעיית אופטימיזיה קמורה, ובאילו בנו רמת 0 היא לא קמורה. כמו כן, ראיינו כי באמצעות נורמת 1 נוכל להוציא 'צלילות' לפיתרון. נוכל להרחיב דבר זה למשפחת L_q של נורמות:

$$\text{For } 0 < q \in \mathbb{R}, x \in \mathbb{R}^d \quad \|x\|_q := \left(\sum_{i=1}^d (x_i)^q \right)^{\frac{1}{q}}$$

הבה ונתבונן בכל אחד מה諾רמות ובמרכיביה:

- דليلות - לכל $1 \leq q$ נשיג פתרונות דليلים ולמעשה מספר הפיצרים קטן מ- d .
- קמיירות - לכל $1 \geq q$ מדובר בבעיית אופטימיזציה קמורה, וכך ניתן לפתור אותה ביעילות.
- אם כך $1 = q$ מושג את שני הדברים. על מנת להבין מדוע נוכל לקבל את ה'دليلות' בנורמות קטנות מ-1, נתבונן בהגדלה הבאה, של כדור היחידה.

הגדלה

עבור נורמה $\|\cdot\|_q$ ב- \mathbb{R}^d , כדור ברדיוס ρ הוא הקבוצה $\{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\| \leq \rho\}$. כדור היחידה הוא עם רדיוס $\rho = 1$.

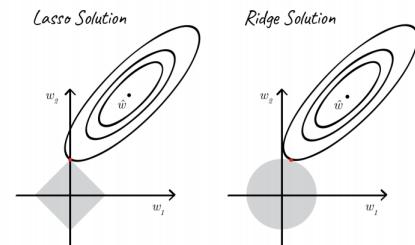
במקרה בו הנורמה גדולה מ-1, כלומר $q > 1$, לכדור היחידה אין פינות. אינטואיטיבית, אם אנחנו על 'פינה', אז זה אומר שיש הרבה ערכים שהינם 0-ים. ככל שאחננו קטנים מ-1, מספר הפינות גדל, במובן שיש יותר ערכים שקרובים ל-0:³⁵:



אם נרצה יותר נספת לאינטואיציה על כך, נתבונן בשני בעיות האופטימיזציה שראינו לגבי כל אחת מהנורמות:

$$\begin{array}{ll} \text{minimize } w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d & \|w_0 \mathbf{1} + \mathbf{Xw} - y\|^2 \\ \text{to subject} & \|\mathbf{w}\|_2^2 \leq \rho \end{array} \quad \left| \quad \begin{array}{ll} \text{minimize } \mathbf{w}_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d & \|w_0 \mathbf{1} + \mathbf{Xw} - y\|^2 \\ \text{to subject} & \|\mathbf{w}\|_1 \leq \rho \end{array} \right.$$

אם אנחנו מגדילים את ρ , אז תכל'ס אנחנו מתקדמים לפיתרון של ה"ריבועים הפחותים", כי תכל'ס אין משמעות להגבשות. ככל שנגביל יותר ויוטר את ρ , נגביל את הפיתרון להיות בתחום הcéדור הספציפי. הפיתרון יתקבל בחיתוכים עם הcéדור ρ ³⁶ וכן במקרה של נורמה מ-1, החיתוכים יתקיימו באחת מהפינות.

**3.4 המקרה האורתוגונלי**

ונסה להבין את האינטואיציה מתחום דليلות הלאסו בקורס נספת. נניח כי כל הפיצרים אורתוגונליים אחד לשני. ככלומר בפרט $\mathbf{X}^\top \mathbf{X} = I_d$. במקרה זה, הכל הינו ניתן לכתוב את y בתור צירוף ליניארי של קטוריים אורתוגונרמליים. במקרה זה, קיבלנו נוסחה סגורה ל-LASSO- $\hat{\mathbf{w}}_\lambda^{\text{subset}}$, $\hat{\mathbf{w}}_\lambda^{\text{ridge}}$, $\hat{\mathbf{w}}_\lambda^{\text{lasso}}$ שmbוססת על \mathbf{L}^S .

³⁵ גם אם זה לא למורי אינטואיציה מדויקת, חרטוט גמור זה לא.
³⁶ בספר לא מצוין בדיקות "למה". מניות שזאת ההגדלה.

נגידר תחילה שתי פונקציות סף (thresholding functions)

הגדרה

פונקציות סף-קשה וסף-מרוכך יקרו הfonקציות $\eta_\lambda^{\text{hard}}, \eta_\lambda^{\text{soft}} : \mathbb{R} \rightarrow \mathbb{R}$ שמודדרת על ידי :

$$\eta_\lambda^{\text{hard}} := \mathbf{1}[|x| \geq \lambda] \cdot x \quad \eta_\lambda^{\text{soft}} := \text{sign}(x)[|x| - \lambda]_+ = \begin{cases} x - \lambda & x \geq \lambda \\ 0 & |x| < \lambda \\ x + \lambda & x \leq \lambda \end{cases}$$

פונקציות אלו מייצגות התכונות של x לכיוון 0, כתלות ב- λ . פונקציית הסף הקשה מופסת קלט בטוחה $(-\lambda, \lambda)$ ומיאירה את כל השאר ללא שינוי. פונקציית הערך המרוכך מכוצת ערכים בתחום $(\lambda, -\lambda)$ ומופסת את כל הערכים שלא בטוחה.

טענה

יהיו y, X מטריצה אורתוגונלית וקטור תגובה בהתאם. נסמן את \hat{w} בטור הפתרון של OLS. הרגולרייזציה של האופטימיזציה באמצעות לאסו, הינה מהצורה של $\hat{w}^{\text{lasso}}(\lambda) = \eta_\lambda^{\text{hard}}(\hat{w})$.

הוכחה

זכור כי הפתרון של הריבועים הפחותים (least squares) הינו $Xy = \hat{w}$. נוכל לכתוב את פונקציית היעד בטור:

$$\begin{aligned} f_{\ell_1}(\mathbf{w}) &= \underbrace{\frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2}_{\text{פתיחת}} + \lambda \|\mathbf{w}\|_1 = \\ &\downarrow \\ &= \underbrace{\frac{1}{2} (\|\mathbf{y}\|^2 - 2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w})}_{\text{הצבה}} + \lambda \|\mathbf{w}\|_1 = \\ &\downarrow \\ &= \underbrace{\frac{1}{2} (\|\mathbf{y}\|^2 + (\mathbf{w}^\top - 2\hat{\mathbf{w}}^\top) \mathbf{w})}_{\text{שינויי הגדרה}} + \lambda \|\mathbf{w}\|_1 = \\ &\downarrow \\ &= \underbrace{\frac{1}{2} \|\mathbf{y}\|^2 + \sum_{j=1}^d \left(\frac{1}{2} w_j^2 - \hat{w}_j w_j + \lambda |w_j| \right)}_{\text{שינויי לפונקציית סימן}} = \\ &= \frac{1}{2} \|\mathbf{y}\|^2 + \sum_{j=1}^d \left(\frac{1}{2} w_j - \hat{w}_j + \lambda \text{sign}(w_j) \right) w_j \end{aligned}$$

כעת, נרצה למזער כל ערך של w_j כתלות ב- λ . נקבל:

$$\frac{\partial \frac{1}{2} \|y - \mathbf{xw}\| + \lambda \|w\|_1}{\partial w_j} = \frac{\partial (\frac{1}{2} w_j - \hat{w}_j + \lambda \text{sign}(w_j)) w_j}{\partial w_j} = w_j - \hat{w}_j + \lambda \text{sign}(w_j) = 0$$

$$\downarrow \downarrow$$

$$w_j = \hat{w}_j - \lambda \text{sign}(w_j)$$

כעת, נחלק למקרים:

□ אם $|\hat{w}_j| < \lambda$. למה?

- אם $\hat{w}_j > 0$ אז $w_j = \hat{w}_j + \lambda < 0$.

- ואם $\hat{w}_j < 0$ אז $w_j = \hat{w}_j - \lambda > 0$.

□ אם $|\hat{w}_j| \geq \lambda$ מה שנכוון רק $l-0$ $w_j = \hat{w}_j - \lambda \text{sign}(w_j)$ ולכן

. $w_j = \hat{w}_j + \lambda$ אז $w_j = \hat{w}_j - \lambda \text{sign}(w_j) \leq -\lambda$

□ אם $|\hat{w}_j| < \lambda$ מה שנכוון רק $l-0$ $w_j = \hat{w}_j - \lambda \text{sign}(w_j)$ ולכן

אם לחבר את כל הדברים הללו יחדיו, נקבל כי:

$$w_j(\hat{w}_j, \lambda) = \begin{cases} \hat{w} - \lambda & \hat{w} \geq \lambda \\ 0 & |\hat{w}| < \lambda \\ \hat{w} + \lambda & \hat{w} \leq \lambda \end{cases} \Rightarrow \hat{w}_j^{\text{lasso}}(\lambda) = \eta_{\lambda}^{\text{soft}}(\hat{w}_j)$$

בצורה דומה, ניתן להראות כי תת הקבוצה הטובה ביותר ופתרון ridge הינה:

$$\begin{aligned} \hat{w}_{\lambda}^{\text{subset}} &:= \eta_{\sqrt{\lambda}}^{\text{hard}}(\hat{w}^{\text{LS}}) \\ \hat{w}_{\lambda}^{\text{ridge}} &:= \hat{w}^{\text{LS}} / (1 + \lambda) \end{aligned}$$

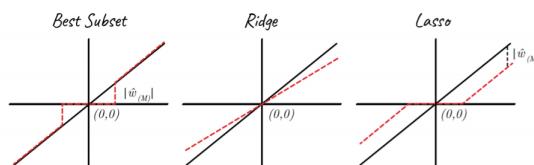
בקיצור, ההצעה באמצעות המטריצה האורטוגונלית מאפשרת לנו להציג הפתרונות השונים השונים באמצעות הפעלת פונקציית כיווץ לכל אחד מהכטיניות בוקטור \hat{w} . כלומר למעשה:

□ תת הקבוצה הטובה ביותר מ微微ירה חלק מהמקדמים $l-0$ ומושאירת את השאר כפי שהם.

□ לאסו微微ירה רק מהמקדמים $l-0$ ומכווצת את השאר לפי λ .

□ Ridge היא פשוטה מכפלה בסקלר.

נקבל:



רגולרייזציית רגרסיה לוגיסטיות

ונכל לשיים את הרגוליזציה שראינו גם לרגרסיה לוגיסטיבית. בדומה לרגרסיה ליניארית, אם הפיצ'רים גדולים ממדגימות, נקבל בעיות.
נזכיר בהגדרת הסיווג על ידי הרגרסיה הלוגיסטיבית:

$$\hat{\mathbf{w}} := \underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^m \left[y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + w_0) - \log \left(1 + e^{w_0 + \langle \mathbf{w}, \mathbf{x}_i \rangle} \right) \right]$$

ונכל להוציא את ℓ_1 ל'ביטוי המהימנות', על מנת לקבל רגוליזציה:

$$\hat{\mathbf{w}} := \underset{w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d}{\operatorname{argmax}} \underbrace{\sum_{i=1}^m \left[y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + w_0) - \log \left(1 + e^{w_0 + \langle \mathbf{w}, \mathbf{x}_i \rangle} \right) \right]}_{\mathcal{F}_S(\mathbf{w})} + \lambda \underbrace{\|\mathbf{w}\|_1}_{\mathcal{R}(\mathbf{w})}$$

מדובר עדיין בבעיית אופטימיזציה קמורה שיש לה פתרון. הרגרסיה הלוגיסטיבית שעבירה רגוליזציה ℓ_1 היא מסובב חזק ביותר מעל $\mathcal{X} = \mathbb{R}^d$ - יש לה שונות נמוכה, אפשר לשЛОט בטריידאוף המפורסם והוא לגמרי ניתנת לפרשנות.

4 בחרת מודל והערכתה (Model Selection and -Evaluation)

אם נסכים, נוכל לומר כי עד כה למדנו מספר אלגוריתמים למידה וריעונות למידה כלליים. כעת ננסה לענות על השאלה הבאות:

◻ כיצד נוכל לבחור מודל? חלק מהמודלים שראינו הם למעשה משפחות של מודל שתלוים בפרמטר מסוים (למשל k בשכנים, עומק העץ, או λ -ב-SVM).

אם נרצה לעשות boosting או נרצה לבחור כמה איטרציות להפעיל אותם. אם נפעיל bagging חסר קורלציה, נצורך פרמטרים נוספים.

◻ כיצד נוכל להעריך את הביצועים של מודל שבחרנו? נוכל להעריך זאת על קלט מסוים ואז לבדוק האם הוא מתאים. לעיתים נרצה גם לדעת את יכולת לפני הפעלה.

היכולת שנרצה להעריך שימושה גם על בחרת המודל הינה למעשה שגיאת הכללה (generalization error). נניח כי קיימת פונקציית loss כלשהי שמוגדרת על ידי (\cdot, \cdot, ℓ) . נגידר את הסיכון האמפירי של היפותזה $\mathcal{Y} \rightarrow \mathcal{X}$: h על ידי ממוצע ההפסד על מרחב המודדים:

$$L_S(h) := \frac{1}{m} \sum_{i=1}^m \ell(h(\mathbf{x}_i), y_i) \quad S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$$

שגיאת הכללה (לפעמים נקרא לה test error) היא למעשה השגיאה החזואה על דוגמיה בלתי תלויות למודל שלנו:

◻ אם אנחנו מניחים כי הדוגמיה שלנו לא נוצרה בצורה מובנת כלשהי, נגידר את שגיאת הכללה בתורו:

$$L(h) := \sum_{i=1}^{|T|} \ell(h(\mathbf{x}_i), y_i)$$

□ יש פונקציית הסתברות מעל \mathcal{X} ופונקציית תיוג לא ידועה $\gamma \rightarrow \mathcal{X}$: כאשר נניח כי מדובר במודל PAC, שגיאת ההכללה היא התוחלת של השגיאה מעלה מרחב המדגמים:

$$\mathbb{E}_{x \sim \mathcal{D}} [\ell(h(\mathbf{x}), f(\mathbf{x}))]$$

□ ישנה פונקציית הסתברות מעל $\gamma \times \mathcal{X}$: כאשר נניח כי מדובר במודל PAC אגנוסטי, שגיאת ההכללה היא התוחלת מעלה זוג מההתפלגות:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\ell(h(\mathbf{x}), y)]$$

אחרי שאנו מבינים מה שגיאת ההכללה הנכונה, הצעד הבא הוא להעריך זאת באופן נכון. נרצה בהינתן מוגם כלשהו ליציר 'בחירה' מודל ו'הערכת' מעלה משפחת אלגוריטמי הלמידה $\{A_\alpha\}$, על ידי S :

אלגוריתם 7 בחירת מודל והערכתה

1. תאמן כל מודל $A \in A_\alpha$ על S וייצר את הקבוצה $\{h_\alpha := A_\alpha(S)\}$.
 2. תבחר את המודל הטוב ביותר באמצעות $. \alpha^* = \operatorname{argmin}_\alpha L_S(h_\alpha)$
 3. תגדיר את שגיאת ההכללה של h_{α^*} מעלה המוגם $. L_{\mathcal{D}(h_{\alpha^*})} := L_S(h_{\alpha^*})$
 4. תחזיר את $. \alpha^*, h_{\alpha^*}, L_S(h_{\alpha^*})$
-

נבחין כי למעשה באלגוריתם זה אנחנו מוצאים את השגיאה האמפירית הטובה ביותר. אבל זה לא ממשו, כי כפי שאנו מבינים, נתאים את עצמנו בזרה הטובה ביותר למוגם שיש לנו. אבל למעשה בעקבות כך נסובל משונות גבואה - ככלומר שגיאת ההכללה לא תהיה מדהימה. לכן נסה לחפש דרך אחרת להערכת מודל.

4.1 סכימת אימון-תוקף-מבחן (Train-Validation-Test Scheme)

הבעיה במה שעשינו היא שהשתמשנו בקבוצה סופית וניסינו באמצעות האימון עליה להעריך כלפי העתיד. נוכל להשתמש בשלוש קבוצות - קבוצת האימון S (training), קבוצת התוקף V (validation) וקבוצת המבחן T (testing). ננסה לעשות זאת באמצעות האלגוריתם הבא:

אלגוריתם 8 בחרת מודל והערכתה - מחודש

1. תאמן כל מודל $A \in A_\alpha$ על S וייצר את הקבוצה $\{h_\alpha := A_\alpha(S)\}$.
2. תבחר את המודל הטוב ביותר מעל קבוצת התוקף ($\alpha^* = \operatorname{argmin}_\alpha L_V(h_\alpha)$).
3. תעריך את שגיאת ההכללה של h_{α^*} מעל מדגם המבחן ($L_{\mathcal{D}(h_{\alpha^*})} := L_T(h_{\alpha^*})$).
4. תחזיר את $(\alpha^*, h_{\alpha^*}, L_S(h_{\alpha^*}))$.

באמצעות הקבוצה החדשה שיצרנו, אנחנו למשה מפרטים את התפקידים של קבוצת האימון. באמצעות דבר זה נוכל להבטיח כי מדובר באומדן לא מוטה.

טענה

תהי $h_S = \mathcal{A}(S)$ מחלקת ההיפتواזה שחווארת באמצעות "לומד" מדגם אימון S ותהי V קבוצה חדשה של דוגמאות שנוצרות בצורה בלתי תלויות ו שוות התפלגות מעל \mathcal{D} . הסיכון האמפירי של h_S על V הוא אומדן בלתי מוטה של שגיאת ההכללה של h_S .

הוכחה

נסמן את $|V| = m_V$. כיוון שהדגימות נוצרות בצורה שווה התפלגות, נקבל:

$$\mathbb{E}_{V \sim \mathcal{D}^{m_V}} [L_V(h_S)] = \frac{1}{m_V} \sum_{i=1}^{m_V} \mathbb{E}_{(\mathbf{x}_i, y_i) \sim \mathcal{D}} [\ell(h_S, (\mathbf{x}_i, y_i))] = L_{\mathcal{D}}(h_S)$$

נניח כמובן כי פונקציית ההפסד חסומה, אחרת שגיאת ההכללה גם היא אינה חסומה. נניח כי פונקציית ההפסד חסומה על ידי 1. מכאן נגע למסקנה הבאה:

מסקנה

שגיאת ההכללה של h_S חסומה על ידי:

$$\mathbb{P} \left[|L_V(h_S) - L_{\mathcal{D}}(h_S)| \leq \sqrt{\frac{\log(2/\delta)}{2m_V}} \right] \geq 1 - \delta$$

הוכחה

נשתמש dabei שוויון הופding³⁷.

כיוון ש- V היא קבוצה של דוגמאות שוות התפלגות ובלתי התפלגות, נסמן את $Z_i = Z_{m_V} + \dots + Z_1$ בטור המשתנה המקרי של הזוג. כיוון שהדגימות הן שוות התפלגות ובלתי תלויות, נקבל כי גם Z_1, \dots, Z_{m_V} הם בלתי תלויים ושוות התפלגות.

כעת, נסמן את $X_i := X_{m_V} + \dots + X_i$ - המשתנה המקרי של ההפסד מעל דגימת Z_i . מכאן עולה כי X_1, \dots, X_{m_V} הם משתנים מקרים בלתי תלויים ושווי התפלגות כך ש- $1 \leq i \leq m_V$ $0 \leq X_i \leq 1$.
היא קבוצה של משתנים מקרים בלתי תלויים ושווי התפלגות כך ש- $1 \leq i \leq m_V$ $0 \leq X_i \leq 1$.
אם נסמן $(X_I = L_V(h_S))$ - כיוון ש- $\bar{X} = \frac{1}{m_V} \sum X_I = L_V(h_S)$ הוא אומדן בלתי מוטה של שגיאת ההכללה, נקבל באופן ישיר כי:

³⁷השתמשנו בו לאחרונה, להזכירות כניסה לפרק של הסתברות

$$\mathbb{P}[|L_V(h_S) - L_{\mathcal{D}}(h_S)| \geq \varepsilon] \leq 2 \exp\left(2m_V\varepsilon^2\right)$$

אם נבחר $\delta = \sqrt{\frac{1}{2m_V} \ln \frac{2}{\delta}}$, נקבל כמפורט כיו:

$$\mathbb{P}\left[|L_V(h_S) - L_{\mathcal{D}}(h_S)| \leq \sqrt{\frac{\log(2/\delta)}{2m_V}}\right] \geq 1 - \delta$$

4.2 שיטת ה-Cross Validation

מציאותית, לחלק את המידע שלנו יכול להיות בעייתי (הקטנת המידע גורמת לתוצאות לא מספיק טובות). אם כך, נרצה למצוא דרך אחרת לחלק את המידע. הדרך הכי פשוטה לעשות זאת, היא באמצעות Cross-validation. במקום לחושב על S בטור קובוצה בודדת, נחשב על איחוד זר של K קובוצות. לכל קובוצה $K \leq k \leq 1$ נאמן מודל שישתמש בכל הדוגמאות של S , חוץ מהדגימה ששicityת לקובוצה ה- k . בדגימה ה- k על מנת לחשב את שגיאת החיזוי. לבסוף, נחזיר את שגיאת ההכללה מעל כל K הקובוצות. זו למעשה שיטת ה- k -fold-Cross Validation. או האלגוריתם נראה כך:

אלגוריתם 9 Cross-Validation

1. תחלק את S בצורה רנדומלית ל- k קובוצות זרות.

2. לכל $1 \leq i \leq k$:

(א) תאמן את המודל S חוץ מהקובוצה ה- i .

(ב) תחשב את ההפסד מעל S_i .

3. תחזיר את הממוצע ואת סטיית של k ההפסדים.

זו השיטה הפופולרית ביותר על מנת לבחור פרמטר כוונון (tuning parameters). לכל לומד \mathcal{A}_α , נאמן אותו k פעמים. לאחר מכן נבחר α שהממוצע שלו הוא הקטן ביותר. לאחר מכן, נאמן את המודל \mathcal{A} שקיבלו על כל המידע. נוכל להשתמש בשיטה זו גם לאמן מודל \mathcal{A} שכבר קיבלנו - וכך בדרכז או להציג הערכה של שגיאת ההכללה ומידת דיוק של הערכה זו.

בחירה מס' החלוקות k

כיצד נוכל לבחור את הערך k ? אם $1 = k$ אז ברור שאין חלוקה. אם נחלק ל-2, חלק אחד נאמן וחלק אחד נבחן - מה שగירום לכך שהשגיאה תהיה גדולה משגיאת ההכללה האמיתית... אם $m = k$, כלומר לגודל המדגמים, דבר זה נקרא one-out-CV. בקיצור, אם k גדול, נאמן על מידע קטן מדי ולכך שגיאתו עלולה להיות מוטה, ולהיפך אם k גדול מדי - כל מדגם יהיה דומה לחברו ומילא נקלט שגיאות גבוהה מאוד.

שימוש ב-Bootstrap להערכת שגיאת ההכללה

ונכל להשתמש בשיטת *the-k*-Bootstrap גם להערכת שגיאת ההכללה, באמצעות המודם היחיד שלנו, S .

ניקח את \mathcal{A}_α שהוא לומד כלשהו, ו- $N \in B$ מספר ה-*bootstrap* שנבצע. בעת נוכל:

□ ליצור מודם $S^{(b)}$ עם m דוגמאות עם החלפות.

□ ניקח קבוצה $S^{(b)} = S \setminus T^{(b)}$ - הדוגמאות שלא נבחרו בצעד זה.

□ נאמן את \mathcal{A} מעל $S^{(b)}$, נקבל h_S ונבחן זאת מעל $T^{(b)}$.

□ נחשב את שגיאת ההכללה וסתית הקן של ההכללה.

שגיאות נפוצות בבחירה מודל

ישנן שתי בעיות נפוצות בבחירה המודל בכל אחת מהדריכים שראינו.

שגיאה אחרת גורמת להערכת יתר של שגיאת ההכללה והשניה גורמת להערכת-חסר של שגיאת ההכללה.

הערכת יתר של שגיאת ההכללה

נניח שאחרי שהשתמשנו ב-cross-validation מעל \mathcal{A}_α ומיצאו α מסוים. כל מודל למעשה אומן בשימוש ב- $m \frac{(k-1)}{k}$ דוגמאות.

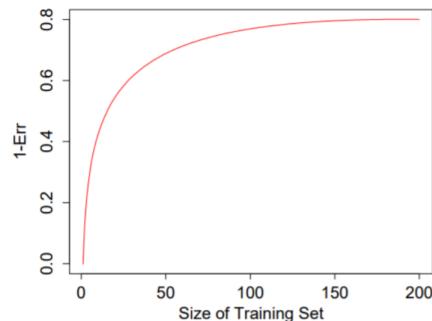
ראינו כבר ב-PAC כי במידה מוצלחת תלולה בגודל המודם. כך למשל הגדרנו את סיבוכיות המודם בהתאם להדגימות המינימלי הנדרש ללמידה היפואת מסויימת וכו'.

יתכן אם כך כי m הדוגמאות אכן מספיקות ללמידה, אבל $\frac{1}{k}m$ לא וכן זה לא בהכרח מעריך טוב את שגיאת ההכללה והיא תהיה גבוהה מדי.

על מנת שנוכל למצוא מספר טוב יותר של חלוקות היינו רוצים שיהיה לנו איזושהי עוקומה שהיא סוג של פונקציית סיבוכיות המודם:

בහינתו גודל מודם מסוים, נקבל את ההצלחה הצפואה. במצב כזה, נבדוק האם לאחר k החלוקות, נשאר באזורי נורמלי ולא נשתנה משמעותית ממוחשב המודם.

אם, לעומת זאת, העוקמה הינה תלולה למדי, אז במצב כזה נסבול מהערכת יתר של שגיאת ההכללה:



הערכת חסר של שגיאת ההכללה

בעיית אפיו נפוצה יותר - היא שאנו אופטימיים מדי ויש לנו הערכת חסר של שגיאת ההכללה.

כאשר יש לנו איזו בעיית למידה מסויימת, נוכל לשנות את המודם באמצעות מחיקת או הוספת פיצרים ולשחק עם דוגמאות בעייתיות - וכך נקבל מודם שעובד יפה.

במצב כזה, נקבל הערכת חסר של שגיאת ההכללה, בעקבות שתי שגיאות:

□ חטטנות מודל (model snooping) - כאשרנו בודקים המון לומדים ומשנים את הפרמטרים של כל אחד, אנחנו למעשה מתחילהים 'להתאהב' במידע - אנחנו מעדיפים מודל מסוים בגלל שיש לו הטיה נמוכה יותר. זה קורה אפיו אם validation-test.

- חטטנות מידע (data snooping) - כאשר נבדוק את המודל שלנו על דוגמאות חדשות - יתכן שהיא תחסן כל מיני דוגמאות בעייתיות. בדומה זאת, שיטת ה-cross-validation תבצע הערכת חסר. על מנת למנוע בעיות אלו, נדרש להתמודד עם שתי שיטות 'חטטנות' אלו. נמצמצ את חטטנות המידע והמודל לקבוצה קטנה של S 'מצוות' באופטימיות. בדומה זאת, נקבל הבנה כללית של המידע והמודלים האפשריים. רק לאחר שסיימו את שלב זה ובחרנו מודלים אפשריים, נשתמנש במודם האימון הכללי לבחירת המודל שלנו. בנוסף, נמנע מחתטנות מידע ידנית. כתוב את כל תהליך ה-pre-processing. בכל איטרציה של Bootstrap או בנוספ', נמנע מחתטנות מידע cross-validation נחזיר את כל תהליך ה-pre-processing מעל הדגימה הנוכחית, כביכול אנחנו דוגמים מעל מידע חדש.

חלק VII

למידה בלתי מונחת (Unsupervised Learning)

עד כה, דיברנו בעיקר על למידת מונחת קבוצה (supervised batch learning) - כל הסיפור עם אימון קבוצה והחיזוי על פיה. ישן בעיות למידה נוספות שלא משתמשות בתחום זה - בעיות ללא תגיות, ועם מרחב נתונים X בלבד. ישן מספר דוגמאות לכך:

- חשיפת מבנים מממדים נמוכים (Uncovering low-dimensional structures) - במקרים מסוימים תהיה לנו סיבה טובה להאמין כי משהו שמיוצג בממד גובה ניתן להציג גם בממד נמוך. למשל, אם ניקח את MNIST אותו חיצינו באחד התרגילים, על אף שהוא מושהו מממד 28×28 , חלק מהפרמטרים שם נשאים קבועים ולא תורמים לחיזוי ולכך אפשר להעיף אותם - ככלומר נוכל להוריד את הממד שלהם לממד נמוך יותר.
- איחוד (Clustering) - לעיתים נקבל מידע מסוים ונרצה לאחד בין קבוצות המידע "שדי קרובות", ביחס לשאר הדגימות. נתיחס לתתית קבוצות אלו בתור clusters. אם נסתכל שוב על MNIST, הגיוני שככל תמונה של מספר ספציפי מייצגת אלמנטים יחסית קרובים. לכן, הגיוני שננסה לחבר ביניהם. דוגמאות חדשות נשווה אותן לקודמות.
- זיהוי חריגות (Anomaly detection) - נניח שאנו רוצחים לבדוק متى מערכת מסוימת לא מתנהגת 'כרגיל'. למשל, אם ניקח תחנת כוח, נרצה לבדוק متى היא מתנהגת 'מוואר'. אנחנו לא יודעים כיצד התנהגות מווארה נראהית - ישן מספר אפשרויות. כיצד נוכל לבדוק האם ההתחנוגות חריגה? דוגמאות מסוימות מראיש, ובהינתן דוגמאות חדשות נשווה אותן לקודמות.

1 הקטנות ממדים

הקטנות ממדים היא למעשה תהליכי המיפוי של מרחב מממד גובה ל透ק מרחב מממד נמוך מממד. ככלומר, בהינתן מידע $\mathbf{x}_m \in \mathbb{R}^d$, נרצה למצוא העתקה $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$, אשר $f(\mathbf{x}_1), \dots, f(\mathbf{x}_m)$ עדין מסמלים את $\mathbf{x}_1, \dots, \mathbf{x}_m$ בצורה מסוימת. בדרך כלל מחלקים את טכניקות הקטנות הממד לטכניקות ליניאריות ולא ליניאריות, כתלות ב- f הנבחרת.

ישן מספר סיבות ללמידה ולהתאים הקטנות ממד:

- יכולת ללמידה (Learnability) - חלק מאלגוריתמי הלמידה עובדים טוב יותר עם ממדים קטנים. למשל, עבור רגרסיה ליניארית, ראיינו כי כאשר הפיצרים קטנים מהדגימות, התוצאות בעייתיות

□ **חשיבותים -** עבור הרבה מאלגוריתמי הלמידה, זמן הרצאה והמקום תלויים בגודל הממד. על ידי הורדת הממד אנחנו משתמשים בפחות משתנים.

□ **יזואלייזציה -** קל יותר להסביר ולחזור את המידע כשהוא ניתן להצגה. אם נוכל להוריד את הממד $k \leq 4$, אז נוכל להציג את המידע.

1.1. ניתוח גורמים ראשיים (Principal Component Analysis)

יהיו $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^d$, כלומר, שטחים כלשהם, שמצמצמים אותם לתת-מרחב ליניארי מגודל k (נום יותר). היינו רוצים למצוא העתקה ש'טטייל' את המידע המקורי גדול הקטן. נתמודד עם שני אטגרים במצב זה:

1. ישנו אינסוף תתי מרחבים שנוכל לבחור. כיצד נבחר את המתאים ביותר?

2. אפילו אם אנחנו יודעים מהו תת-המרחב הראשי, הטלת נקודות $\mathbf{x} \in \mathbb{R}^d$ לתוך תת-המרחב עדין לא מצמצת את המידע. היינו רוצים למצוא דרך לייצג כל דוגמה באמצעות k הקוארדינטות בתת-המרחב. דרך זו ידועה בתור הטבעה (embedding) של הדוגמאות.

במקרה של ניתוח גורמים ראשיים (PCA), בהינתן מידע $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathbb{R}^d$, אנחנו ממחפשים העתקה ליניארית כך שמניער את השגיאה הריבועית בין הדוגמאות החדשנות והדוגמאות לאחר העתקה:

$$f^* := \operatorname{argmin}_f \sum_{i=1}^m \| \mathbf{x}_i - f(\mathbf{x}_i) \|^2$$

ניתן לפתח בעיה זו במספר כיוונים:

1. למצוא את תת-המרחב האפיני הקרוב ביותר לנקודות.

2. למצוא את תת-המרחב האפיני ששומר על רוב הגוונים במידע.

3. למצוא את תת-המרחב האפיני שמאוצר את העיות (distortion) של זוגות מרחוקים בין נקודות בסביבה המקורית לתתי המרחבים הפנימיים.

4. הכללה של רגרסיה ליניארית עם רעש גאוסיני גם במידע המקורי וגם בכיוון שנוצר. דרך זאת נקראת גם בתור PCA הסתברותי.

נתיחס כאן לשתי הגישות הראשונות והモוכרות יותר:

1.1.1. תת-המרחב האפיני הקרוב ביותר

על מנת לפשט את הרעיון של מציאות תת-המרחב האפיני הקרוב ביותר נניח כי המידע שלנו ממורכב סביב הראשית, כך שבמקרים למצוא תת-מרחב אפיני, אנחנו ממחפשים תת-מרחב רגיל. כלומר, אנחנו ממחפשים העתקה ליניארית $U : \mathbb{R}^k \rightarrow \mathbb{R}^d$ והעתקה 'הופכית' $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$

$$W^*, U^* = \operatorname{argmin}_{W \in \mathbb{R}^{d \times k}, U \in \mathbb{R}^{k \times d}} \sum_{i=1}^m \| \mathbf{x}_i - UW\mathbf{x}_i \|^2$$

למה

יהיו (U, W) פתרון לביטוי הקודם. אזי העמודות של U הן אורתוגונורמליות ו- $W = U^\top$ הוכחה

יהיו W מטריצות ונניח כי קיימת העתקה $\mathbf{x} \rightarrow \mathbf{x}$. המטריצה (UW) היא מטריצה מגודל $d \times d$ עם ממד k , כלומר התמונה שלה היא תת מרחב של \mathbb{R}^d מממד k . נסמן $S := \text{Im}(UW) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} = UW\mathbf{u}, \mathbf{u} \in S\}$. נקבל כי ההטלה שמצערת את \mathbf{x} היא הטלה אורתוגונלית לתווך המרחב (כפי שראינו או דיברנו, ההטלה האורתוגונלית היא זאת שמצערת את את המרחב מהנקודה למטה המרחב). בambilים אחרות, הנקודה ב- S הקרובה ביותר ל- \mathbf{x} נתונה על ידי $\mathbf{x}' = VV^\top \mathbf{x}$ כאשר העמודות של V הם בסיס אורתוגונורמלי של S :

$$\forall \mathbf{u} \in S \quad \|\mathbf{x} - \mathbf{u}\|_2 \geq \|\mathbf{x} - VV^\top \mathbf{x}\|_2$$

כלומר, כפי שרצינו, הפתרון לביטוי הראשוני הוא U עם עמודות אורתוגונורמלית, כלומר $W = U^\top$.

בהתבסס על הלמה הקודמת נוכל לרשום בעיה שוקלה ל-PCA. נבחן כי מתקיים:

$$\begin{aligned} & \underbrace{\text{פתיחה סוגרים}}_{\|\mathbf{x} - UU^\top \mathbf{x}\|^2} \\ & \downarrow \\ & \underbrace{\text{תכונות אורתוגונליות}}_{\|\mathbf{x}\|^2 - 2\mathbf{x}^\top UU^\top \mathbf{x} + \mathbf{x}^\top UU^\top UU^\top \mathbf{x}} \\ & \downarrow \\ & \|\mathbf{x}\|^2 - \mathbf{x}^\top UU^\top \mathbf{x} = \\ & \underbrace{(\mathbf{v}^\top \mathbf{u})}_{= (\mathbf{u} \mathbf{v}^\top) \text{trace}} = \\ & \downarrow \\ & \|\mathbf{x}\|^2 - (U^\top \mathbf{x})^\top U^\top \mathbf{x} = \\ & = \|\mathbf{x}\|^2 - \text{trace}((U^\top \mathbf{x})(U^\top \mathbf{x})^\top) \\ & = \|\mathbf{x}\|^2 - \text{trace}(U^\top \mathbf{x} \mathbf{x}^\top U) \end{aligned}$$

כיון שה- trace הוא אופרטור ליניארי, נוכל לכתוב זאת כבעיה חדשה:

$$U^* = \underset{U \in \mathbb{R}^{d \times k}, U^\top U = I, i=1}{\operatorname{argmax}} = \left(\sum_{i=1}^m \text{trace}(U^\top \mathbf{x}_i \mathbf{x}_i^\top U) \right) = \underset{U \in \mathbb{R}^{d \times k}, U^\top U = I}{\operatorname{argmax}} \text{trace} \left(U^\top \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^\top U \right)$$

טענה

יהי $A = \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^\top$ והיו $\mathbf{u}_1, \dots, \mathbf{u}_k$ הוקטורים העצמיים הראשונים של A . אזי הפתרון לביעיית ה-PCA נתון על ידי $U \in \mathbb{R}^{d \times k}$, שהעמודות של הין $\mathbf{u}_1, \dots, \mathbf{u}_k$.

הכללה לתתי מרחבים אפיניים

בטענה לעליה מצאנו את תת המרחב הקרוב ביותר אבל לא את תת המרחב האפיני הקרוב ביותר. על מנת למצוא את תת המרחב האפיני הקרוב ביותר, علينا למשה להעתקה W להיות אפינית. על מנת להשיג זאת, علينا להקליל את מה שהוא מעלה באמצעות הגדרת $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$:

$$W(\mathbf{x}) := \widetilde{W}(\mathbf{x} - \mu), \quad \mu \in \mathbb{R}^d, \widetilde{W} : \mathbb{R}^d \rightarrow \mathbb{R}^k$$

דבר זה מאפשר לנו 'להסיט' את המידע לפני שנפעיל את הנטקה \widetilde{W} . כאשר נוסיף את μ לבועית האופטימיזציה לעליה, נגלה למעשה כי המוצע μ נתון על ידי $\sum_{i=1}^m \mathbf{x}_i = \frac{1}{m}$ (זהו למעשה המוצע האמפירי שמסומן לעיתים קרובות באמצעות $\bar{\mathbf{x}}$). אם כך, על מנת למצוא את תת המרחב האפיני הקרוב ביותר, אנחנו ממרכזים את המטריצה A :

$$A := \sum_{i=1}^m (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

משמעותו של שמהטריצה הזאת דומה מאוד למטריצת השינויות המשותפות ואכן נראה את הקשר בין הדברים. דבר זה מאפשר לנו להסיק את הפסאודו-קוד של אלגוריתם PCA:

הערכים העצמיים של A הינם $\lambda_1 \geq \dots \geq \lambda_n$ נחכמים בתור הערכים המרכזיים (Principcal Values) של X . הוקטורים העצמיים של A , כולם $\mathbf{u}_1, \dots, \mathbf{u}_n$ מתייחסים בתור המרכיבים המרכזיים (Principal Componenets). נבחין כי המטריצה A היא מטריצה $d \times d$ חיובית למחצה. לכן אלגוריתם PCA הוא למעשה תהליך לכISON - علينا למצוא את הערכים העצמיים והוקטורים העצמיים של מטריצה כלשהי.

אלגוריתם 10 PCA

1. תחשב את $\sum_{i=1}^m (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$.
2. תמציא את $\mathbf{u}_1, \dots, \mathbf{u}_k$ הוקטורים העצמיים המתאים ל- k הערכים העצמיים הגדולים ביותר.
3. תחזיר את $\mathbf{u}_1, \dots, \mathbf{u}_k$.

1.1.2 מקסימום שונות נשמרת (Maximum Retained Variance)

דרך אחרת לחושב על PCA היא לחשב על דרך מצומצם מדדים שמספקת את הכמות הגדולה ביותר של השינויות ב-PCA האפשרית ב-תת מרחב $k < d$ מדדי.

על מנת לפתור זאת, علينا לפתור בעיות מקסום במאזעות כופלי לגראן'.

טענה

תהי X מטריצה $d \times d$. ההטלה של X על תת מרחב ליניארי k מדדי ששמירת את מקסימום השינויות ב- X נתונה על ידי המטריצה $U \in \mathbb{R}^{d \times k}$, שכוללת את k הוקטורים העצמיים עם k הערכים העצמיים של מטריצת השינויות S .

הוכחה

נתחילה עם הטלה לתוכן מרחב מממד אחד. בלי הגבלת הכלליות, יהיו $\mathbf{v} \in \mathbb{R}^d$ וקטור שנטיל את המידע עליו. השונות של הטלה של כל \mathbf{x}_i על \mathbf{v} , שמוגדרת על ידי $\mathbf{x}_i^\top \mathbf{v}$, הינה הבאה:

$$\begin{aligned}\mathbb{E}[\mathbf{v}^\top \mathbf{x}] &= \frac{1}{m} \sum \mathbf{v}^\top \mathbf{x}_i = \mathbf{v}^\top \bar{\mathbf{x}} \\ \text{Var}(\mathbf{v}^\top \mathbf{x}) &= \mathbb{E}_{\mathbf{x}} \left[(\mathbf{v}^\top \mathbf{x}_i - \mathbb{E}_{\mathbf{x}}[\mathbf{v}^\top \mathbf{x}])^2 \right] = \frac{1}{m} \sum (\mathbf{v}^\top \mathbf{x}_i - \mathbf{v}^\top \bar{\mathbf{x}})^2 \\ &= \frac{1}{m} \sum [\mathbf{v}^\top (\mathbf{x}_i - \bar{\mathbf{x}})]^2 = \frac{1}{m} \sum [\mathbf{v}^\top (\mathbf{x}_i - \bar{\mathbf{x}})] [\mathbf{v}^\top (\mathbf{x}_i - \bar{\mathbf{x}})]^\top \\ &= \frac{1}{m} \sum \mathbf{v}^\top (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{v} = \mathbf{v}^\top \left[\frac{1}{m} \sum (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top \right] \mathbf{v} \\ &= \mathbf{v}^\top S \mathbf{v}\end{aligned}$$

כעת, עליינו למקסם את השונות המוטלת, ביחס ל- \mathbf{v} , ככלומר $\hat{\mathbf{v}} = \underset{\|\mathbf{v}\|=1}{\operatorname{argmax}} \mathbf{v}^\top S \mathbf{v}$. נשתמש בכופלי גראן, עם האילוץ $\mathbf{v}^\top \mathbf{v} = 1$. קיבל כי על מנת לפתור את בעיית האופטימיזציה הבאה, נשתמש ב-

$$\begin{aligned}\mathcal{L} &= \mathbf{v}^\top S \mathbf{v} + \lambda g(\mathbf{v}) \\ &\downarrow \\ \frac{\partial}{\partial \mathbf{v}} \mathcal{L} &= 2S\mathbf{v} - 2\lambda\mathbf{v} = 0\end{aligned}$$

אם כך, הממקסם של $\mathbf{v}^\top S \mathbf{v}$ חייב להיות וקטור עצמי של S . נבחין כי אם נכפול את הנגזרת ב- \mathbf{v}^\top נקבל את השונות עצמה שמתקובלת, שהינה:

$$\mathbf{v}^\top S \mathbf{v} = \lambda \mathbf{v}^\top \mathbf{v} \stackrel{\|\mathbf{v}\|=1}{=} \lambda$$

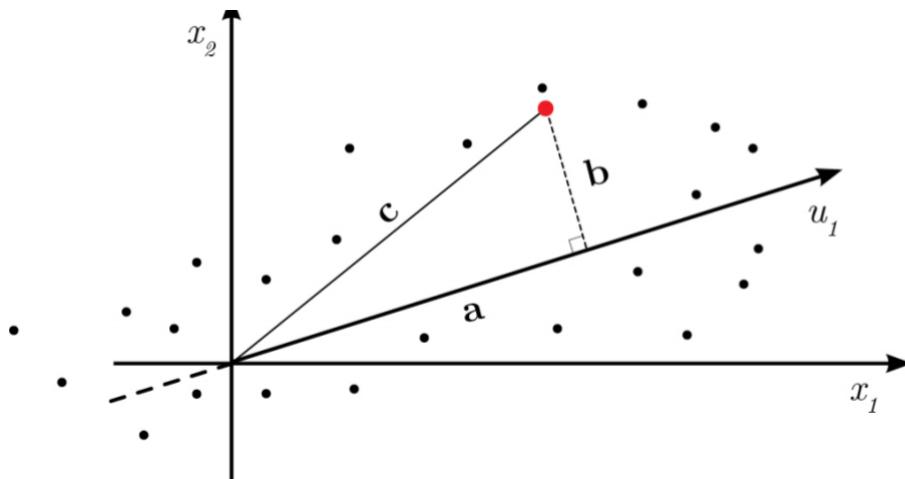
אם כך, השונות המקסימלית שמתקובלת הינה הערך העצמי המקסימלי λ_1 , שמתקובל על ידי $\mathbf{u}_1 = \mathbf{v}$. כעת, עליינו למצוא את הכיוון שמשיג את הכמות השניה בגודלה של השונות. כיוון שאנו חזו מוחפשים אחרי הטלה אורתוגונליות, נוסיף אילוץ נוספת כיוון אורתוגונלי ל- \mathbf{u}_1 , ונקבל:

$$\hat{\mathbf{v}} = \underset{\|\mathbf{v}\|=1, \mathbf{v}^\top \mathbf{u}_1=0}{\operatorname{argmax}} \mathbf{v}^\top S \mathbf{v}$$

כמו קודם, נפתרו את בעיית האילוצים עם אילוץ נוסף, ונקבל את $\mathbf{u}_2 = \mathbf{v}$ שמתאים לערך העצמי λ_2 . אם נוכיח באינדוקציה, נקבל לכל $k \leq d$, תת המרחב שמשיג את השונות המקסימלית של הטלה \mathbf{x} עליה, נתונה על ידי k הוקטוריים העצמיים של S .

1.1.3 הקשר בין תת המרחב הקרוב ביותר ומקסום השונות

ראינו למעשה שני פירושים ל-PCA קודם לכן. בתור תת המרחב הקרוב ביותר ובתור מקסום השונות. על מנת להבין את הקשר בין שני אלו, התבוננו בציור הבא ובפרט בנקודות ה- x_i :



נבחין כי באמצעות ההטלה האורתוגונלית של x_i ל- u_1 , נוצר משולש ישר זווית:

- הצלע שמסומנת בתור a הינה הadol של ההטלה x_i על u_1 . קלומר $\|x_i^\top u_1\| = \|x_i - x_i^\top u_1 u_1\|$. את זה נרצה למקסם, אם מדובר במונחים של מקסום השונות.
 - הצלע שמסומנת בתור b הינה המרחק בין הנקודה x_i וההטלה האורתוגונלית ל- u_1 . קלומר $\|x_i - x_i^\top u_1 u_1\|$. נרצה למינימיזר אותו, מובן של מציאת תת המרחב הקרוב ביותר.
 - הצלע שמסומנת בתור c הינה הadol של x_i . קלומר $\|x_i\| = c$.
- ממשפט פיתגורס, נקבל כי $c^2 = a^2 + b^2$. לכן, אם נמצא פתרון PCA שමינימזיר את b ,เราจะ למשהו פתרון שמקסם את a , וכן הפוך.

1.1.4 הטלה מול קוואדרינטות של הנקודות

כאשר אומרים כי PCA הוא כלי חשוב לעתים קרובות מתעלמים מההבדל בין הטלה ובין שיכון (embedding) של הנקודות.

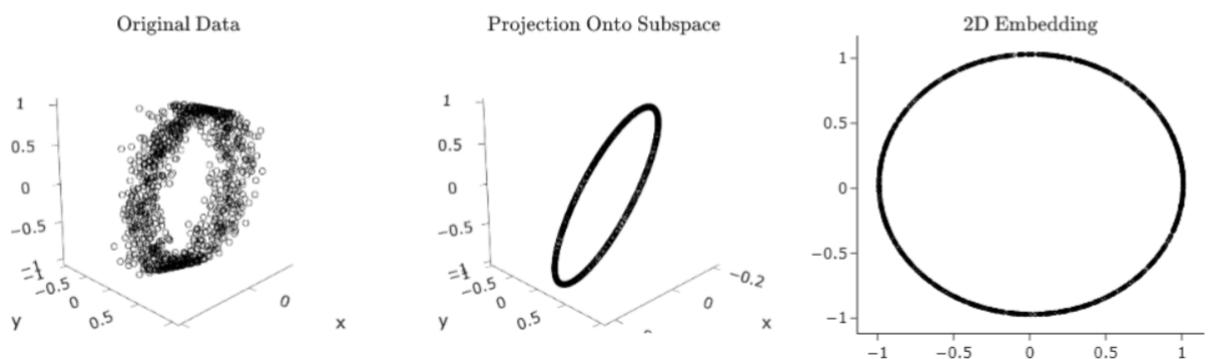
יהי $X \in \mathbb{R}^{m \times d}$ ונניח כיPCA נריצ' כדי למציאת תת מרחב ליניארי k ממדי. הפתרון האופטימלי של PCA הוא תת המרחב שמןימזר את סכום המרחקים הריבועיים בין כל נקודה x_i וההטלה האורתוגונלית למרחב. כפי שראינו קודם לכן, תת המרחב הזה נפרש על ידי k הוקטוריים העצמיים של $d \times d$ מטריצת השונות המשותפת $S = \frac{1}{m} \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^\top$.

המטריצה $U \in \mathbb{R}^{d \times k}$ מאפשרת **להטלת** של הנקודות ב- \mathbb{R}^d לתוך תת המרחב ה- k -ממדי, להפרש על ידי הוקטוריים העצמיים המוביילים הללו. יחד עם זאת, נרצה להוריד את הממד 'באמת', כלומר למציאת העתקה $W : \mathbb{R}^d \rightarrow \mathbb{R}^k$ ולעבוד עם המידע המוצמצם $(x_1, \dots, x_m)^\top W$. כפי שהוכחנו קודם, למשהו $U^\top W = U^\top$ היא העתקה שمبיאה לנו את השיכון לתוך תת המרחב הקטן יותר. הוקטור $x_i^\top U^\top$ הוא k ממדי. למעשה, קוואדרינטות של הוקטור x_i המקורי, לפי סט אורתונורמלי של k וקטורים עצמיים מוביילים.

יהיו $\mathbf{u}_d, \dots, \mathbf{u}_1$ וקטורים עצמיים של S שמושפרים בסדר יורד לפי הערכים העצמיים המתאים. כיוון שוקטוריים אלו הם מביסיס כלשהו ב- \mathbb{R}^d , נוכל להציג את הוקטור \mathbf{x}_i בתור $\mathbf{x}_i = \sum_{j=1}^d \langle \mathbf{x}_i, \mathbf{u}_j \rangle \mathbf{u}_j$.

- ההטלה של \mathbf{x}_i על תת המרחב ה- k -ממדי נתונה על ידי $\langle \mathbf{x}_i, \mathbf{u}_k \rangle \mathbf{u}_k$.

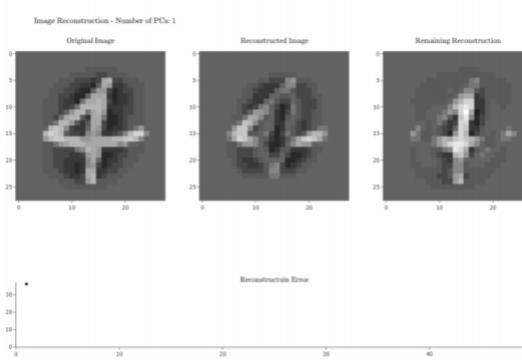
בצירור הבא נדגים את ההבדל בין הקוארדינטות ובין ההטלה של המידע. המידע נוצר בצורה הבאה: 1000 נקודות נדגו מתוך כדור היחידה (\mathbb{B}_2) ב- \mathbb{R}^2 . לעומת זאת, $\{(x_i)_1, (x_i)_2\}_{i=1}^{1000}$ כאשר $(x_i)_1^2 + (x_i)_2^2 = 1$ (x_i). באמצעות ה-PCA אנחנו מטילים את המידע לתוך \mathbb{R}^2 , אך אמם הוא עדין מייצג ב- \mathbb{R}^3 . לבסוף, באיזור השלישי, אנחנו מבצעים את השיכון לתוך \mathbb{R}^2 .



1.1.5 גורמים ראשיים כ"נקודות מידע אופייניות"

הגורמים הראשיים שנמצאו על ידי ה-PCA הם וקטורים במרחב הסביבה \mathbb{R}^d . לעומת זאת, מכך עולה כי הם חולקים את אותו ממד כמו נקודות המידע $\mathbf{x}_1, \dots, \mathbf{x}_n$. כיוון שהם וקטורים אורתונורמליים שנבחרו כך שה- k הראשונים מייצאים את הקירוב הליינרי הטוב ביותר מממד k למידע, נוכל להסתכל עליהם בדרך מעניינת. מבחינה מסוימת, הם הנקודות שהכי שונות אחת מהשנייה. בעקבות כך, מעוניין לראות מה הם מייצגים בתוך נקודות מידע,iziaeha חלק של ה"אות" הם תופסים.

אם נחזור למספרית ה-MNIST, נוכל להתבונן בכל גורם ראשי בתור תמונה 28 על 28. נוכל להסתכל על כל תמונה בתור הצירוף הליינרי של תМОונות אלו, כאמור $\widehat{f}(\mathbf{x}) = \sum_{i=1}^k \alpha_i \mathbf{u}_i + \bar{\mathbf{x}}$, כאשר $\mathbf{u}_1, \dots, \mathbf{u}_k$ הם k הגורמים הראשיים המוביילים ו- $\bar{\mathbf{x}}$ הוקטור המרכז של המידע. נבחין כי כיוון שמדובר בצירוף ליניארי, אנחנו בונים את המדגם \mathbf{x} על ידי הוספה והחסרת גורמים ראשיים:



ציור זה מראה את הבנייה מחדש של תמונה מס' 4 בתווך צירוף ליניארי של הגורמים הראשיים. בכל פריטים אנחנו מקובלים את הגורמים הראשיים הקודמים ועוד α_i לכל i וגורם ראשי של האיטרציה הנוכחי α_{-i} הקוארדינטות המתאימות.

1.1.6 בחירת k

מציאותית, אף אחד לא אומר לנו איזה k לבחור. אם כן כיצד נבחר את ה- k -המתאימים? אם נבחר k גדול מדי, אנחנו בתכל"ס לא עושים כלום. אם נבחר k קטן מדי, אנחנו מפספסים מידע חשוב. נבחן כי אם המידע שלנו 'באמת' נמצא בתת מרחב k ולא d , אז יהיו לנו לבדוק k גורמים ראשיים שאין אפסים. בעקבות כך, אם מצאנו את תת המרחב הקרוב ביותר, אמם יהיו יותר $m-k$ גורמים ראשיים שאינם אפסים. לכן, הדרך הפופולרית היא ליצא את הגורמים הראשיים לפי סדר יורך, ולהערך פחות או יותר כמה גורמים ראשיים חשובים יש.

שימוש לב שלא מדובר באלגוריתם כי אין כאן נוסחה סגורה איך לעשות זאת.

38

2 איגוד (Clustering) 2

אחד הכלים החזקים בלמידה בעיות הוא איחוד או Clustering. לעיתים, בין אם כחלק מחקרית מיידיע או כמטרה ראשית, אנחנו מעוניינים בחלוקת המידע שלנו לתור מספר קבוצות **משמעותיות**. למשל, בהינתן מספר תמונות נרצה לחלק אותן למספר תמונות כמו טבע, אנשים וכו'. בצורה שונה, בהינתן אוסף של גנים, נרצה לחבר יחד גנים שמתקשרים למחלות מסוימות. בשתי הדוגמאות, אנחנו רק מקבלים את הדגימות (תמונות או גנים), אבל אין לנו את המידע האמתי (התגיות). אין לנו מידע על מה יש בתוך התמונות, או אילו מחלות מתקשירות עם הגנים. בהסתמך על הדמיון בין הדגימות, נוכל לחלק את המידע שלנו לחתמי קבוצות שונים, כאשר נאחד דגימות שיתוור דומות אחד לשנייה, ביחס לדגימות אחרות.

2.1 אמצעים (k -Means)

למרות שישנן דרכים רבות על מנת ליצור איחודים, אחת הדרכים המקובלות היא להגדיר נקודות מידע נציגות של האיחוד. לאחר מכן, לאחד את כל הנקודות ביחס לנקודות המייצגות.

הגדרה

חלוקת של קבוצות מידע $\{\mathbf{x}_i\}_{i=1}^m = \bigcup_{j=1}^k C_j, \dots, C_k$ כך ש-

בהינתן חלוקה כלשהי על המידע ונקודות נציגים, נוכל להגדיר פונקציית מחיר (cost function) עבור החלוקה. עברו מטריקה כלשהי מעל המידע, $\mathcal{X} \times \mathbb{R}_+ \rightarrow \mathcal{X} \times \mathcal{X}$, נגדיר (או צנטרואיד):

³⁸בקודזה זו עברתי לסכם רק את המשפטים, ללא ההוכחות. תחנו.

$$G_d(C_1, \dots, C_k, \mu_1, \dots, \mu_k) := \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mu_j)$$

כאשר μ_1, \dots, μ_k הם הנציגים של האינטראקטיבים C_1, \dots, C_k . בהשתמש בפונקציה זו, המטריה היא למצוא את החלוקות שמצוירות את הדבר הבא:

$$\{C_1, \dots, C_k\}^* = \operatorname{argmin}_{\{\mu_j\}_{j=1}^k} G_d(C_1, \dots, C_k, \mu_1, \mu_k) = \operatorname{argmin}_{\{\mu_j\}_{j=1}^k} \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mu_j)$$

במקרה של אלגוריתם ה- k -means, קבוצות הנציגים נבחרות להיות $\mu_j(C_j) := \operatorname{argmin}_{\mu \in \mathcal{X}} \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mu)$. על מנת למצוא את המאפיינים הביטויי לעיל, علينا לעבור על כל החלוקות האפשרות של m אובייקטים ל- k קבוצות. כיוון שישנו מספר אקספוננציאלי זהה, מעורר פונקציית G הינו בעיית NP קשה, علينا להיעזר בהיוריסטיקה מסויימת. השיטה המפורסמת עבורי מזעורה G , כאשר d היא המרחק האוקלידי, היא k -means:

אלגוריתם 11 K-Means

1. תבחר מרכזים כובד (centroids) שיוגדרו על ידי μ_1, \dots, μ_k , רנדומלית.

2. כל עוד אין התכנסה:

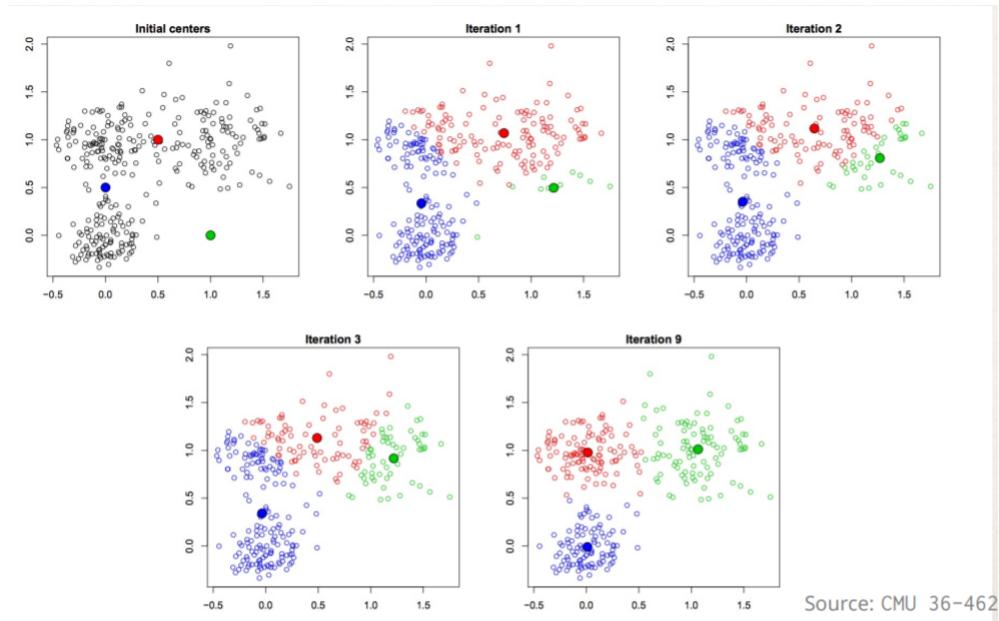
(א) התאמה: תתאים כל נקודה \mathbf{x} למרczy הקובד הקרוב ביותר אליה.

(ב) עדכון: تعدכן כל מרכז כובד. ככלומר, לכל $j \in [k]$ קיבל כי $\mathbf{x} \in C_j^{(t)}$

3. תחזיר את C_1, \dots, C_k .

שיטת אינטראקטיב זו משתמשת באלגוריתם של לירוד, למזעורה G באמצעות אסטרטגיית משתנה. קודם כל אנחנו מגדירים חלוקה של המידוע, לאחר מכן מתאיםים את הנקודות לפי מרכז הקובד הקרוב ביותר.תתי קבוצות אלו נקראים תא וורוני (Voronoi). לאחר מכן, אנחנו מחשבים מחדש את מרכז הקובד. אנחנו מקווים שהאלגוריתם יתכנס.

אם נרצה לראות דוגמה לכך:



2.1.1 התכונות למספר פתרונות ולתתי פתרונות

כיוון שאיחוד הוא בעית NP קשה, האלגוריתם של k-means משתמש בהיוריסטיקת האלגוריתם של ליניאר מינימיזציה. כיוון שמדובר אך ורק בהיוריסטיקה, האופטימליות של האלגוריתם אינה מובטחת. למעשה, כיוון שהבחירה של מרכזי הכביד הראשיים היא רנדומלית, האלגוריתם של K-Means עלול להתכנס לתה-אופטימלי, או כי יתכן וישנו יותר מפתרון אופטימלי אחד לבעה.

מה הכוונה כאשרנו אומרים תה אופטימלי? בהינתן קבוצת מידע $\{x_i\}_{i=1}^m$, אלגוריתם K-means מחייר חלוקה C_1, \dots, C_k

כך שישנה חלוקה אחרת של המידע C'_1, \dots, C'_k עם ערך קטן יותר, כלומר $G(C_1, \dots, C_k) > G(C'_1, \dots, C'_k)$.
כיוון שהחותואה של פונקציית המטרות הלו איננה קמורה, ניתן וישנה מספר נקודות מינימום מקומיות, שכל אחת מקבלת ערך שונה.

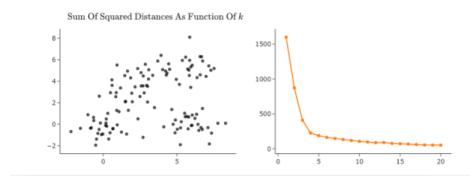
מה הכוונה כאשרנו אומרים מספר פתרונות? הכוונה היא שבהתאם קבוצת מידע $\{x_i\}_{i=1}^m$, ישנה יותר מחלוקת אפשרית של המידע כך שפונקציית המטרה תשיג אחד מינימלי. חשוב לבדוק כי בכל פעם שנדבר על איחוד, תמיד נתבונן על חלוקה לפרמוטציות עד כדי שמורות החלוקת. למשל, אם ניקח את הדגימות 1, 2, 3, 4, 5, ואת החלוקת $C_1 = \{1, 2, 3\}, C_2 = \{4, 5\}$, $C'_1 = \{1, 2, 3\}, C'_2 = \{4, 5\}$ היא זהה להחלוקת $C_1 = \{1, 2, 3\}, C_2 = \{4, 5\}$ בפונקציית המטרה. כאשרנו מדברים על מספר פתרונות אופטימלים, אנחנו מתכוונים ל:

$$\begin{aligned} &\exists \{C_1, \dots, C_k\}, \{C'_1, \dots, C'_k\} \\ &G(C_1, \dots, C_k) = G(C'_1, \dots, C'_k) \\ &\exists j \in [k] \quad \forall l \in [k] \quad C_j \neq C'_l \end{aligned}$$

2.1.2 בחרית k

כמו אלגוריתמים אחרים שראינו כבר, גם K-Means נדרש לקבל ערך k . נוכל להבחן כי מדובר גם כאן בפרמטר כוונון של השונות וההטייה - ככל ש- k גדול יותר, אז השונות של המדגם גדולה יותר. כיוון שאנו לא באמת יודעים כמה קבוצות יש לנו במידע, זו איננה משימה קללה. נבחן כי כל עוד יש לנו יותר נקודות מידע מסווגים, עבור כל פיתרון אופטימי עם k אינדיבידואלים, נוכל להשיג פיתרון עם מזער קטן יותר של פונקציית המטריה עם $1 + k$. במקרה, לrox על כל ה- k האפשריים ולבחר את הערכcis המזערים זו לא באמת משימה אפשרית.

במקרים זאת, נציג טכנית דומה לזואת שהשתמשנו ב-PCA. ניצא את הערך שהתקבל עבור הערכcis השונים של k , ובחר את הערך שהאריך השיפור הוא לא דרמטי. כמו שנראה בציור הבא, אנחנו מפעילים את האסטרטגיה על המידע. כדי שנייתן לראות, הנהו $k = 4$ שלALARICO השיפור הוא לא מאוד דרמטי. כיוון שהשיגנו את המידע מהפעלת 4 גאוסינים שונים, נראה שמצאו את הערך הנכון:



2.2 אינטראקטיבי מילוקסן (Spectral Clustering)

נרצה לעשות שני שדרוגים ביחס ל-means. ראשית, האם נוכל להתעלם ממרחוקים גדולים בעת האינטראקטיב? האם אנחנו יכולים להסתמך רק על מרחוקים בין זוגות? לפעמים המידע גדול מדי ולכן נרצה את שתי האופציות הללו. אנחנו הולכים לשלב שלושה רעיונות:

1. הולכים להתבונן במרחוקים קטנים בלבד.
2. נייצר גרף לפי מספר הדגימות, הצלעות יהיו לפי מרחוקי affinity (מרחוקים שבודקים האם דגימות מסוימות שונות מאוד).
3. נשתמש באלכסון ומיציאת k הוקטוריהם העצמיים הגדולים ביותר.

2.2.1 האלגוריתם

1. נגדיר את מטריצת הסמייכיות להיות $A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right)$ כאשר $\varepsilon > 0$ ונבחר לפי הסקללה הרצוייה.
2. נחשב את $D = D^{-1}A$ כאשר D היא אלכסונית ועל האלכסון יש את $D_{i,j} = \sum_{i=1}^m A_{i,j}$ (חלוקת למשה של כל שורה בסכום שלה).

בתוך הנוסחה הראשונה יש המונ אינפורמציה. אם הנקודות רחוקות מאוד אחת מהשנייה, אז התוצאה מאוד קטנה e^{-t} כאשר t גדול מאוד. מצד שני, אם הנקודות מאוד קרובות, אז מדובר במשה בהפרש המרחוקים הרגיל. אפשר להרים דבר זה גם הוא לא ב- \mathbb{R}^d (למשל, השוואת טקסטים). למעשה, ברגע שננדיר מרחוק, נעבד רק על פיו - זה נקרא metric learning.

המטריצה L אינה סימטרית, אבל כל התאים אי שליליים וסכום השורה הוא 1. אבל למורות זאת המטריצה L דומה למטריצה סימטרית.

המטרה שלנו היא לעשות ניתוח של רכיבי הקשרות של הגרף, כיצד נעשה זאת?

2.2.2 הקסם של הלכsoon

מכיוון שהשורות של המטריצה L בסכוםות ל-1, יש לה לפחות וקטור עצמי שמתאים לערך העצמי 1. נניח כי למטריצה יש k רכיבי קשרות - ניתן להראות כי למטריצה יש k וקטורים עצמיים שמתאימים

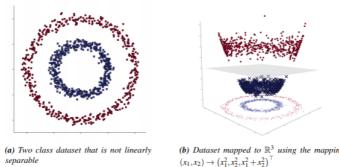
חלק VIII

דרכי קרנלייזציה (kernel methods)

נזכיר במחלקות ההיפותזה של רגרסיה ליניארית וסיווג:

$$\begin{aligned}\mathcal{H}_{reg} &:= \{\mathbf{x} \rightarrow w_0 + \mathbf{w}^\top \mathbf{x} \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d\} \\ \mathcal{H}_{reg} &:= \{\mathbf{x} \rightarrow \text{sign}(w_0 + \mathbf{w}^\top \mathbf{x}) \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d\}\end{aligned}$$

لمחלקות ההיפותזה יש כוח מוגבל. נניח במקרה בו $\mathbb{R}^2 \subset \mathcal{X}$ עם דוגמאות כפי שיש בציור הבא:



במקרה זה, המידע אינו ניתן להפרדה ליניארית, ולכן אלגוריתמים שבנויים למחלקת ההיפותזה הזאת, עלולים ליפול. נבחן כי אם במקומות החזקה המקורי של המידע, נעביר את הדוגמאות לתמונה אחרת, למשל על ידי התאמת $(x_1^2, x_2^2, x_1^2 + x_2^2) \rightarrow \mathbf{x}$, נקבל מרחב ליניארי ניתן להפרדה ב- \mathbb{R}^3 (כפי שיש בציור הימני). מעל הצגה זו נוכל להפריד זאת לשני תגיות בשימוש באלגוריתם ה-SVM (מציאת תחת מרחב, באפור).

אם כך, על מנת להעшир את יכולת הבהעה (אקספרסיונות) של מחלקת ההיפותזה, אנחנו נניח שאנו יכולים לשכן את המידע בתוך מרחב (אולי מממד גדול יותר כפי שראינו), של פיצרים, שעליו נוכל ללמידה. בדומה סכטנית:

□ בהינתן מרחב מדגם $\mathbb{R}^d \subseteq \mathcal{X}$ ואלגוריתם למידה \mathcal{A} , נבחר פונקציית שיכון $\mathcal{F} : \mathcal{X} \rightarrow \psi$ עברו מרחב פיצרים \mathcal{F} כלשהו. בהרבה מקרים נרצה כי ψ עברו $d >> k$ (ואולי אפילו ∞)

□ נאמן את המידע \mathcal{A} על קבוצת האימונו $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ בשימוש במחלקת ההיפותזה:

$$\begin{aligned}\mathcal{H}_\psi &:= \{\mathbf{x} \rightarrow w_0 + \mathbf{w}^\top \psi(\mathbf{x}) \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^k\} \\ \mathcal{H}_\psi &:= \{\mathbf{x} \rightarrow \text{sign}(w_0 + \mathbf{w}^\top \psi(\mathbf{x})) \mid w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^k\}\end{aligned}$$

תהליך זה נראה מאד אטרקטיבי, כיון שההנחה שהמידע שלנו ניתן לתיאור באמצעות פונקציה ליניארית כלשהי (עבור רגרסיה או סיווג), במרחב פיצ'רים כלשהו, נראית שלא מוגבלת יותר מדי. האמת היא שכבר התעסkeno בדוגמה של תהליך זה. במקרה של התאמה פולינומית, קיבלנו דגימה $\psi(x) \in \mathbb{R}^m$ וקיים כל הטמענו זאת בתוך מרחב פיצ'רים חדש, כאשר $\psi(x)_i = \psi(x)$, ואז למדנו זאת באמצעות מחלוקת ההיפותזות של רגרסיה ליניארית. בהמשך, נרჩיב ונכליל דבר זה להטאה של פולינומים מרובים משתנים.

באופן כללי, על מנת להשתמש באסטרטגיה זו, علينا להתמודד עם שני אתגרים מרכזיים. ראשית, אנחנו חייבים למצוא העתקה ψ שתאפשר לנו ללמידה באמצעות מבנה ליניארי. שנית, נרצה כי האלגוריתם שלנו יהיהiesel חישובית. הרבה פעמים העתקה של הדגימות ל- \mathcal{F} כוללת הגדלה של המרחב $d > k$ ולכן זה יכול להיות מאוד מרכיב חישובית: חישוב $(\psi(x))$ עלול להיות מאוד יקר. היינו רוצים למצוא דרך לבוחר $\mathcal{H} \in h_S$ בלי להעריך מפורשות את הממדים הגדולים יותר.

1 בעית למידה חלופית

נבחין כי הרבה מהבעיות שראינו בפרק הקודמים יכולים להכתב בצורה:

$$\underset{\mathbf{w} \in \mathcal{F}}{\operatorname{argmin}} f(\langle \mathbf{w}, \psi(\mathbf{x}_1) \rangle, \dots, \langle \mathbf{w}, \psi(\mathbf{x}_m) \rangle) + R(\|\mathbf{w}\|)$$

עבור פונקציית f שרירותית ו- $R : \mathbb{R}_+ \rightarrow \mathbb{R}$ פונקציה מונוטונית לא עולה. למשל, במקרה של אופטימיזציה הריבועים הפחותים, f הייתה פונקציית RSS $\sum (y_i - \hat{y}_i)^2$ והצירה לנו את \hat{y}_i . במקרה של אופטימיזציה Soft-SVM חישבנו את mean hinge loss.

טענה

יהי \mathcal{X} קבוצת דגימות לא ריקה ו- $\mathcal{F} \rightarrow \mathcal{X} : \psi$ העתקה לתוכו מרחב הילברט. תהי פונקציית אופטימיזציה מעלה $\alpha \in \mathcal{F}$ מהתוצרה שמופיעת קודם. אזי ישנו $\mathbf{w}^* \in \mathbb{R}^m$ כך ש- $\sum_{i=1}^m \alpha_i \psi(\mathbf{x}_i) = \mathbf{w}^*$ והוא מזעך של בעית האופטימיזציה.

בהתבסס על טענה זו, אנחנו יכולים לייצר נוסחה ביחס ל- α . נחליף את \mathbf{w} עם הביטוי הבא:

$$\langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle = \left\langle \sum_{j=1}^m \alpha_j \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \right\rangle = \sum_{j=1}^m \alpha_j \langle \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \rangle$$

וגם נבחין כי:

$$\|\mathbf{w}\|^2 = \left\langle \sum_{i=1}^m \alpha_i \psi(\mathbf{x}_i), \sum_{i=1}^m \alpha_i \psi(\mathbf{x}_i) \right\rangle = \sum_{i,j=1}^m \alpha_i \langle \psi(\mathbf{x}_i), \psi(\mathbf{x}_j) \rangle \alpha_j$$

אם נסמן את $G \in \mathbb{R}^{m \times m}$ מטריצת גראם, אשר עמודותיה הין $\langle \psi(\mathbf{x}_i), \psi(\mathbf{x}_j) \rangle = G_{i,j}$, נוכל לרשום את הביטוי סך הכל בטור:

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} f(G\alpha) + R(\alpha^\top G\alpha)$$

ההצגה הדואלית מאפשרת לנו לחפש רק את $\alpha \in \mathbb{R}^m$ (מספר הדגימות הנדרש), ואילו ההצגה הראשונה מחייבת אחר פתרון $F \in \mathcal{F}$ שיכל להיות גדול מאוד. בנוסף, נבהיר כי הבועיה הדואלית היא בעיה ריבועית ב- α ולכן אם G ניתן לחישוב בצורה ייעילה, כל הבעיה ניתנת לחישוב בצורה ייעילה. אם מצאנו את ה- α -המתאים לבעיה, נוכל לייצר את הפתרון האופטימלי עבור בעיית האופטימיזציה המקורית, ולהזות את התגובה עבור מוגן חדש באמצעות:

$$\hat{y}(\mathbf{x}) = \langle \mathbf{w}^*, \psi(\mathbf{x}) \rangle = \left\langle \sum_{i=1}^m \alpha_i \psi(\mathbf{x}_i), \psi(\mathbf{x}) \right\rangle = \sum \alpha_i \langle \psi(\mathbf{x}_i), \psi(\mathbf{x}) \rangle = \alpha^\top \mathbf{k}$$

כאשר לכל i מתקיים עבור \mathbf{k}_i כי הוא שווה ל- $\langle \psi(\mathbf{x}_i), \psi(\mathbf{x}) \rangle$.

2 אפיון פונקציות קרナル

בחלק הקודם הרأינו כי בהינתן העתקה $\mathcal{F} \rightarrow \mathcal{X}$: ψ כאשר \mathcal{F} הוא מרחב הילברט כלשהו, אנחנו יכולים לפתור את בעיית האופטימיזציה ביחס ל- \mathbb{R}^m ולחשתמש בכך על מנת לחצות תגובה של דגימה חדשה.

יחד עם זאת, כיוון שמרחב המוגן אליו מגיעה העתקה עלול להיות מממד גדול (או אפילו אינסופי), חישוב מטריצת גראם G או מציאת ערך החיזוי של דגימה חדשה, עלול להיות מורכב מאוד חישובי. על מנת להתמודד עם בעיה זו, נשים את הרעיון של **החלפת קרナル (kernel substitution)**. נחליט להשתמש במשפחה ספציפית של פונקציית, שייקראו פונקציות קרナル PSD, שלמרות שהוא עלולות למוגדים גבורים, הם ניתנות לחישוב בצורה ייעילה.

הגדרה

תהי $\mathbb{R} \rightarrow \mathcal{X} \times \mathcal{X} : k$ פונקציה כלשהי. k תיקרא פונקציית קרナル, אם k היא סימטרית. לעומת זאת, לכל $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ מתקיים כי $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$.

הגדרה

תהי $\mathcal{X} \times \mathcal{X} : k$ פונקציית קרナル. k תיקרא פונקציית קרナル חיובית למחצאה (Positive Semi-Definite) אם ורק אם לכל $\mathbb{N} \in m$ ולכל $\mathcal{X} \in \mathbf{x}_1, \dots, \mathbf{x}_m$ המטריצה $K \in \mathbb{R}^{m \times m}$ כאשר $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ היא מטריצה PSD.

על מנת להבחן האם פונקציות נתונות הן פונקציות קרナル, אנחנו יכולים:

1. להראות שפונקציית G הקשורה אליה היא מטריצה PSD.

2. למצוא העתקה ψ שמקיימת כי $\langle \psi(\mathbf{x}), \mathbf{x}' \rangle = k(\mathbf{x}, \mathbf{x}')$.

דוגמאות

תהי פונקציה $\mathbb{R} \rightarrow \mathcal{X} \times \mathcal{X} : k$ כאשר $\mathcal{X} = \mathbb{R}^d$, שוגדרת על ידי $.k(\mathbf{x}, \mathbf{x}') = 1$, $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. הבה ונראה שמדובר בפונקציית קרナル חיובית למחצאה, כאשר נמצא פונקציית העתקה ψ שמקיימת כי $\langle \psi(\mathbf{x}), \mathbf{x}' \rangle = \langle \psi(\mathbf{x}), \mathbf{x}' \rangle$. נתחל בהראות כי המטריצה G המותקשרת ל- k מעל $\mathbf{x}_1, \dots, \mathbf{x}_m$ היא מטריצה חיובית למחצאה. מטריצת גראם של k מעל קבוצה זו היא פשוטה $1 = G_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. נבהיר כי k היא סימטרית ולכן

גם G היא מטריצה סימטרית. בנוסף, הערכים העצמיים של G הם $0 \leq m$, ולכן G היא מטריצה חיובית למחצה ו- k היא מטריצת גרעין חיובית למחצה.

על מנת למצוא פונקציית העתקה ψ , עליה לקיים כי $\langle \psi(x), x' \rangle = 1$ לכל $x \in \mathbb{R}^d$, אז נבחר $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ שמחזירה תמיד וקטור יחידה v ואז נקבל כי $1 = \|\psi(x)\|_2^2 = v^\top v$. נבחן כי לא אמרנו מה המרחב, שיכל להיות למעשה מה שבא לנו.

דוגמה שנייה

נتابון במכפלה הפנימית הסטנדרטית ב- \mathbb{R}^d . נראה כי מדובר בפונקציית גרעין חיובית למחצה. יהיו $x_1, \dots, x_m \in \mathbb{R}^d$ מטריצת גראם המתאימה, כלומר מתקיים בה כי $\langle x_j, x_i \rangle = G_{i,j}$. נבחן כי נוכל לרשום את G גם בתור $X^\top X$ כאשר השורה ה- i של X היא הדגימה ה- i . אז G חיובית למחצה ולכן k היא פונקציית קרナル חיובית למחצה. על מנת למצוא פונקציית העתקה מתאימה ψ , נוכל לבחור את פונקציית הזוזות $\psi(x) = \psi(x)$.

באופן מעניין, אנחנו יכולים לקשר בין פונקציות העתקה ψ שדיברנו עליהם בקטע הקודם, ומטריצות קרナル חיוביות למחצה.

טענה - תנאי מرسل

תהי $\mathcal{F} \rightarrow \mathcal{X} : \psi$ כאשר \mathcal{F} הוא מרחב הילברט כלשהו. אז קיימת פונקציה סימטרית $\mathbb{R} \rightarrow \mathcal{X} \times \mathcal{X} : k$ שסමמתש מכפלה פנימית ב- \mathcal{F} , אם ורק אם k היא פונקציית גרעין חיובית למחצה. כלומר לכל x_1, \dots, x_m מתקיים כי $\langle \psi(x_i), \psi(x_j) \rangle = k(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$, המטריצה G היא חיובית למחצה.

התנאי הזה למעשה אומר לנו כיצד עליינו לבחור את ψ . ראיינו כי בעיית האופטימיזציה הדואלית והחיזויים המכפלות פנימיות של ההעתקות של הדגימות $\langle x', \psi(x) \rangle$. לכן, אם נחבר ψ כך שמטריצת גראם $\langle \psi(x_i), \psi(x_j) \rangle$ היא חיובית למחצה, ישנה פונקציה גרעין חיובית למחצה $\mathbb{R} \rightarrow \mathcal{X} \times \mathcal{X} : k$ שמייצאת את אותו הפלט אבל היא אגנוסטיבית³⁹ למחרחב הפיצרים \mathcal{F} , ולכן ניתנת לחישוב בצורה ישרה.

2.1 פונקציונליות ה الكرナル ה פולינומיאליות והגאוסניות

אחד מפונקציות ה الكرナル הנפוצות היא פונקציית ה الكرナル ה פולינומיאלית. היא מיפה דוגמה נתונה למרחב הפיצרים שהקוואדריניות שלו היא כל המונומים מדרגה לכל היותר k . פונקציה זו מרחיבה את ההעתקה שראינו בהתאם פולינומית.

דוגמה

לפני שנתייחס ל מקרה הכללי, נtabון במקרה של ה الكرナル ה פולינומי ב- \mathbb{R}^2 . הוא מיפה דוגמה נתונה למרחב הפיצרים $.k = 2$ -ו $x_0, x'_0 = 1$, $x, x' \in \mathbb{R}^2$. אז מתקיים כי:

$$\begin{aligned} (1 + \langle x, x' \rangle)^2 &= (1 + x_1 x'_1 + x_2 x'_2)^2 \\ &= 1 + 2x_1 x'_1 + 2x_2 x'_2 + (x_1 x'_1)^2 + (x_2 x'_2)^2 + 2x_1 x'_1 x_2 x'_2 \end{aligned}$$

אם נגדיר ψ על ידי $\psi(x) = (1, 1 \cdot x_1, x_1 \cdot 1, 1 \cdot x_2, x_2 \cdot 1, x_1^2, x_2^2, x_1 \cdot x_2, x_2 \cdot x_1)^\top$

?³⁹

טענה

יהי $\mathcal{X} = \mathbb{R}^d$ ותהי הפונקציה $k(x, x') := (1 + \langle x, x' \rangle)^k$. זו פונקציית קרnl חיובית למחצה בהתאם להעתקה:

$$x \mapsto \left(1, \dots, x_i, \dots, x_i \cdot x_j, \dots, \prod_{\substack{i \in J \\ J \subset [d], |J|=k}} x_i \right)$$

בשימוש בקרnl הפולינומי וההעתקה המתאימה ψ נוכל לחזור רגע לשאלת החישוביות של בעיית האופטימיזציה. אם היינו צריכים לחשב כל תא במטריצת גראם, דהיינו את $G_{i,j} = \langle \psi(x_i), \psi(x_j) \rangle$, זה היה ממש מסובך. אם נשתמש במתודת עסם הגרעין, נחשב בסך הכל $(1 + \langle x, x' \rangle)^k$ פעולות עבור $x, x' \in \mathbb{R}^d$, שסך הכל זה $O(d)$ פעולות.

כלי אחר וועצמתי הוא פונקציית הקרnl הגאוסינית.

טענה - קרnlים גאוסינים

או בהחלה פונקציית קרnl תקינה:

$$\in \mathbb{R}_+ \sigma^2 \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|^2\right) =: (x, x') k$$

עם פונקציית ההעתקה ψ שמוגדרת על ידי:

$$\forall n \in \mathbb{N} \quad \psi(x)_n := \frac{1}{\sqrt{n!}} \exp(-x^2/2\sigma) x^n$$

נבחן כי ψ ממחה למוחב אינסופי ולכן הציגה הראשונה לא אפשרית לחישוב. הציגה הדואלית מאפשרת לנו לחשב עבור m פרמטרים. אבל כפי שראינו מספיק לנו לחשב את פונקציית הגרעין בלבד.

2.2 תכונות סגירות (Closure) עבור פונקציות קרnl PSD

ראינו כי פונקציית גרעין היא פונקציית גרעין PSD אם ורק אם פונקציית גראם המותאמת היא PSD, ולכן ישן נוכל לייצר מספר סגירות עבור פונקציות הגרעין PSD.

טענה

תהי k פונקציית קרnl כלשהי, אז גם הבאים הם פונקציית גרעין:

$$\text{כasher } A \text{ היא PSD בעצמה.} \quad .1$$

$$\text{כasher } c > 0 \text{ כasher } c \cdot k(x, y) \quad .2$$

$$\text{exp}(k(x, y)) \quad .3$$

$$f(x) k f(y) \quad .4$$

בשימוש בתוכנות אלו, נוכל לבנות פונקציית גרעין PSD חדשות וליצור משפחה חדשה של פונקציות.
דוגמה
בשימוש בתוכנות דלעיל, אפשר להראות שהפונקציה הגאומטרית היא אכן חיובית למחצית.

2.3 ייצור קernalים מקוונים קיימים

ניתן לבנות כל מיני סוגים של קרנליים בהתבסס על קרנליים קיימים:

- לכל פונקציה f יתקיים כי $K'(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})K(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$ היא פונקציית גרעין.
 - כל צירוף ליניארי אי שלילי של קרגנלים הוא גם קרnel $.K(\mathbf{x}, \mathbf{x}') = a_1K_1(\mathbf{x}, \mathbf{x}') + a_2K_2(\mathbf{x}, \mathbf{x}')$
 - מכפלת קרגנלים היא גם קרnel $.K(\mathbf{x}, \mathbf{x}') = a_1K_1(\mathbf{x}, \mathbf{x}') + a_2K_2(\mathbf{x}, \mathbf{x}') -$

3. אלגוריתמים מוקורנליים (Kernelized Algorithm)

אם כך, אם יש לנו אלגוריתמים בקורס הנו יכולים להחליפה בקורס מקורנקלט על ידי החלפת x עם w , ו- $\psi(x)$ ו- $\psi(w)$ נונפטור את בעיית האופטימיזציה בהתקדים. כתוב נייחם זאת עבור האלגוריתמים שראינו בעבר.

3.1 ridge לרגסית Kernel

נזכיר בבעית האופטימיזציה של גורסית ridge שמשמעות ridge הוא מינימיזציה של נורמה ℓ_2 של וקטור המקבדים w :

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$$

ונוכל באמצעות החלפת α ל- $-\beta$ ובהמרה לפונקציית גראHAM, לקבל את בעיית המזעור העדכנית:

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \|y\|^2 + 2y^\top G\alpha + \alpha^\top G^2\alpha + \lambda\alpha^\top G\alpha$$

ב尤רתו חישוב הגראדיאנט, השווה לאפס ושים שמשתוריצת $G + \lambda I_m$ היא חיובית בהחלה והפיכה, נקבל כי המזער הוא $\hat{y} = (G + \lambda I_m)^{-1} y$.

3.2 קernal לרגרסיה לוגיסטיות לאחר רגוליזציה

בדומה לרגרסיבית ridge, נוכל ליציא גרסה מקורנת לרגולרייזציה רגרסיבית לוגיסטיבית. לשם הפשטות, תהי פונקציית הרגולרייזציה נורמת של ℓ_2 של w . אם כך, התוצאה של רגולרייזציה ℓ_2 לוגיסטיבית היא:

$$f(\mathbf{w}) = \sum_{i=1}^m \left[\log \left(1 + e^{\langle \mathbf{x}_i, \mathbf{w} \rangle} \right) - y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \right] + \lambda \|\mathbf{w}\|^2$$

בדומה למה שעשינו קודם, נחליף את \mathbf{x} ב- $\psi(\mathbf{x})$ ואת \mathbf{w} ב- $\alpha^\top G\alpha + \lambda\alpha^\top G\alpha$, ונקבל את הנציגות $\alpha_1, \dots, \alpha_m$ המזעירים (עם פונקציית גראHAM). נגזר את הגרדיינט, נושא ל-0 ונקבל כי המזעירים $\alpha_1, \dots, \alpha_m$ הם הפתרונות לשוויונות:

$$\frac{1}{1 + \exp(-[G\alpha]_i)} + 2\lambda\alpha_i = y_i$$

3.3 קרNEL-PCA

אפשר לעשות קרנלייזציה גם לאלגוריתם PCA. קודם כל צריך להפוך את האלגוריתם PCA למכפלה פנימית, אז לישם את הקרןלייזציה כפי שראינו קודם.

מטריצת המרכזו במרחב פיצ'רים (Centering Matrix In Feature Space) כל מה שעשינו קודם מסתמך על כל שהמטריצה שלנו \mathbf{X} היא ממורכזת. בשונה מהמקירה של PCA, בקרNEL של PCA איננו יכולים לחשב את התוחלת ולהפחית זאת, כיון שעליינו לחשב לכל i את $(\mathbf{x}_i)^\top \phi$. אם כך, علينا לחשב זאת בצורה שונה.

למה

תהי G מטריצת גראHAM של ψ מעל קבוצת המידע X . מטריצת גראHAM הממורכזת לשימוש באלגוריתם PCA המקרונלי, נתונה על ידי:

$$\tilde{G} = G - \mathbf{1}_m G - G\mathbf{1}_m + \mathbf{1}_m G \mathbf{1}_m$$

אלגוריתם 12 kernel-PCA

$$\tilde{K} = K - \mathbf{1}_m K - K\mathbf{1}_m + \mathbf{1}_m K \mathbf{1}_m, \quad K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j).$$

יהיו $\alpha^{(1)}, \dots, \alpha^{(l)}$ הווקטורים העצמיים של \tilde{K} המתאים לערכים העצמיים הגבוהים:

(א) חשב את הווקטורים העצמיים של A על ידי $\mathbf{v}^{(j)} := \sum_{i=1}^m \alpha_i^{(j)} \mathbf{x}_i$

(ב) תחזר $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(l)}$.

לסיכום, היכולת של הקרןל היא להעשייר לנו את מחלוקת ההיפזזה ולמצוא דרך אחרת להצגת המידע. ראיינו כיצד אנחנו יכולים להמיר כל מיני אלגוריתמים שראינו בקורס לנוסחה אחרת (את עם ψ). ראיינו כי אפילו לא צריך לדעת מה ψ , אלא מספיק שתהיה זו פונקציה k חיובית למחזקה (PSD). בעקבות כך, נוכל לפטור את הקרןלייזציה בצורה יعلاה, על אף שיתכן ונחנו ממירים זאת למרחב מממד גבוה.

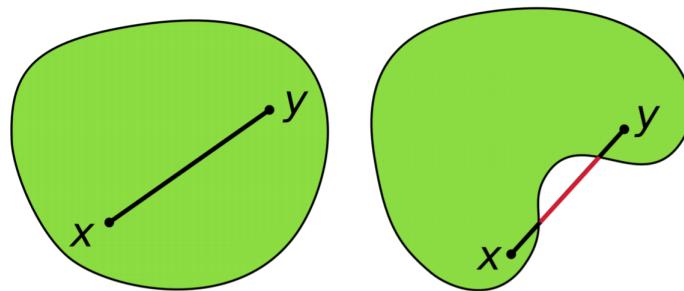
חלק IX**אופטימיזציה קמורה ולמידה عمוקה****1. קבוצות ופונקציות קמורות****1.1. הקדמה**

הרביה שימושות למידה יכולות היכتب בטור 'בעיות למידה קמורות'. בשביל בעיות אלו, נדרש את התחום של אופטימיזציה ליניארית. לשם כך נשתמש במושגים הלוקומים מאינפי, טופולוגיה מאינפי, גיאומטריה.

1.2. קבוצות קמורות**הגדרה**

יהי V מרחב וקטורי. קבוצה $C \subseteq V$ תקרא קמורה אם לכל שני וקטורים $v, u \in C$ ולכל סקלר $\alpha \in [0, 1]$ מתקיים כי $\alpha v + (1 - \alpha)u \in C$.

מבחינת גיאומטרית, קיבל כי היא קמורה אם ורק אם הקו המחבר בין שתי נקודות v, u ב- C , מוכל ב- C :



דוגמאות:

1. תת מרחב ליניארי $V \subseteq U$ הוא קמור, שהרי לכל $u, v \in U$ ו- $\alpha \in [0, 1]$ קיבל כי $\alpha u + (1 - \alpha)v$ היא קומבינציה ליניארית של וקטורים מ- U ובפרט שייכת ל- U .

2. כדור ייחידה. יהיו $B = \{v \in V \mid \|v\| = 1\}$ - איז הוא קמור, שהרי אם $u, v \in B$ ו- $\alpha \in [0, 1]$, קיבל מי שווינו המשולש:

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

3. חצאי מרחב סגורים - קבוצה מהצורה $W = \{v \mid \langle w, v \rangle \leq b\}$ כאשר w הוא וקטור שונה מ零-ב- W . או קבוצה קמורה שכן אם $u, v \in W$ ו- $b \in \mathbb{R}$, נקבל כי:

$$\begin{aligned}\langle w, \alpha u + (1 - \alpha)v \rangle &= \alpha \langle w, u \rangle + (1 - \alpha) \langle w, v \rangle \\ &\leq \alpha b + (1 - \alpha)b \\ &= b\end{aligned}$$

תרגיל

תהי V קבוצת המטריצות המשמשות הסימטריות מממד $d \times d$.

1. הראו כי V הוא מרחב וקטורי (משמעות בפעולת חיבור מטריצות).

2. הראו כי קבוצה של מטריצות שモכלות במרחב חיובית בהחלה (לכל $v^T A v \in \mathbb{R}^d$ מתקיים כי $0 > v^T A v$) היא קמורה.

הוכחה

1. פשוטה.

2. ניקח M, N מטריצות חיוביות בהחלה, ו- $\alpha \in [0, 1]$. מתקיים כי:

$$x^T (\alpha M + (1 - \alpha)N)x = \alpha x^T M x + (1 - \alpha) x^T N x > 0$$

כנדרש.

cutת נתעסן במספר טענות על קבוצות קמורות.

טענה

1. חיתוך $C = \bigcap_{i \in I} C_i$ של אוסף $\{C_i \mid i \in I\}$ של קבוצות קמורות, הוא קמור.

2. הזזה בוקטור והכפלת בסקלר $\{ax + b \mid a \in \mathbb{R}, b \in \mathbb{R}^d\}$ משמרת קmirות.

3. הקבוצה $\{\lambda c \mid c \in C\}$ היא קמורה, לכל קבוצה C קמורה, ולכל סקלר λ .

4. f וגם $f^{-1}(D)$ הן קמורות כאשר f אפינית.

5. הטענה הקודמת נכונה גם לפונקציות $f(x) = \frac{Ax+b}{cx+d}$,_linear fractional, כלומר מהצורה

הוכחה

1. יהיו $u, v \in C$ ותהי $\alpha \in [0, 1]$. לכל $i \in I$ מתקיים כי $u, v \in C_i$ (מהגדרת החיתוך).

מהקמירות של כל C_i , מתקיים כי $\alpha u + (1 - \alpha)v \in C_i$, נקבל גם כי $\alpha u + (1 - \alpha)v \in C$. אם כן, נקבע גם כי $\alpha u + (1 - \alpha)v \in C$.

2. יהיו $u, v \in \lambda C$ ותהי $\alpha \in [0, 1]$. אזי נקבע: $c_1, c_2 \in C$ כleshם. $\alpha c_1 + (1 - \alpha)c_2 \in \lambda C$.

$$\alpha u + (1 - \alpha)v = \lambda(\alpha c_1 + (1 - \alpha)c_2) \in \lambda C$$

תרגיל

על מישור (Hyperplane) הוא קבוצה מהצורה $W = \{v \in V \mid \langle w, v \rangle = b\}$ כאשר $w \in V$ ו- $b \in \mathbb{R}$. הראו כי כל על מישור הוא קמור.

1.3. פונקציות קמורות

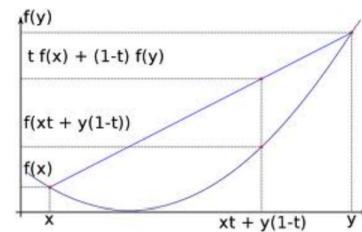
יהי V מרחב וקטורי ו- C קבוצה קמורה. פונקציה $f : C \rightarrow \mathbb{R}$ תיקרא קמורה אם לכל $u, v \in C$ ולכל $0 \leq \lambda \leq 1$ יתקיים כי:

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

פונקציה f תיקרא קמורה ממש, אם ורק אם לכל $0 < \lambda < 1$ ולכל $u \neq v \in C$ יתקיים:

$$f(\lambda u + (1 - \lambda)v) < \lambda f(u) + (1 - \lambda)f(v)$$

בחינה גיאומטרית, פונקציה מעל קבוצה קמורה היא קמורה אם הגרף של הפונקציה נמצא מתחת לכל קו המחבר בין שתי נקודות בgraf:



דוגמאות:

1. פונקציית הנורמה היא קמורה. מי שווין המשולש, קיבל לכל V ו- $\alpha \in [0, 1]$ ו- $u, v \in V$:

$$\underbrace{\triangle}_{\downarrow} \quad \| \alpha u + (1 - \alpha)v \| \leq \| \alpha u \| + \| (1 - \alpha)v \| = \alpha \| u \| + (1 - \alpha) \| v \|$$

זה נכון לכל נורמת ℓ_p . כך גם נורמת ℓ_∞ .
 2. כל פונקציה אפינית ממשית היא קמורה. יהיו $f : V \rightarrow \mathbb{R}$ על ידי $b \in \mathbb{R}$ ו- $w \in V$. נגיד $f(u) = \langle w, u \rangle + b$.
 במקרה, יהיו $\alpha \in [0, 1]$ ו- $u, v \in V$. אזי קיבל:

$$\begin{aligned}
 & \text{הגדלה} \\
 & f(\alpha u + (1 - \alpha)v) = \langle w, \alpha u + (1 - \alpha)v \rangle + b \\
 & \text{הוספה והחסרה} \\
 & = \langle w, \alpha u + (1 - \alpha)v \rangle + \alpha b + (1 - \alpha)b \\
 & \text{מכפלה פנימית} \\
 & = \alpha(\langle w, u \rangle + b) + (1 - \alpha)(\langle w, v \rangle + b) \\
 & \text{הגדלה} \\
 & = \alpha f(u) + (1 - \alpha)f(v)
 \end{aligned}$$

3. העתקת האקספוננט היא קמורה.
4. הפונקציה $x \rightarrow e^{ax}$ עבור $a \geq 1$ או $(a \leq 0)$ היא קמורה.
5. $\log(x)$ – היא פונקציה קמורה.
6. תכונות ריבועיות הן קמורות.
7. סכום הריבועים היא פונקציה קמורה.
8. פונקציית ה- \max היא גם קמורה.
9. האינדיקטור של קבוצה קמורה גם הוא קמור.
אם נתונה לנו פונקציה קמורה כלשהיא, נוכל לייצר פונקציות קמורות נוספות, בשימוש בפעולות אלגבריות.

טענה - שימוש קמורים

1. יהיו $f_i : V \rightarrow \mathbb{R}$ פונקציות נתונות ויהיו $\gamma_1, \dots, \gamma_m$ סקלרים גדולים מאפס. נניח שהפונקציה $\sum_{i=1}^m \gamma_i f_i(u)$ נתונה על ידי $f(u) = \sum_{i=1}^m \gamma_i f_i(u)$. אם f_1, \dots, f_m קמורות, אז גם f קמורה. גם המקסימום מבין הפונקציות הוא פונקציה קמורה.
2. תהינה $f : \mathbb{R}^m \rightarrow \mathbb{R}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}$ על ידי $g : \mathbb{R}^n \rightarrow \mathbb{R}$ ו- $A \in \mathbb{R}^{m \times n}$. אם f היא קמורה, אז גם g היא קמורה.
3. תהינה $f_i : V \rightarrow \mathbb{R}$ עבור $i \in I$. תה $g : V \rightarrow \mathbb{R}$ המוגדרת על ידי $g(u) = \sup_{i \in I} f_i(u)$. אם לכל $i \in I$ מתקיים כי f_i קמורה, אז g גם היא קמורה.
4. אם ישנה פונקציה שנייה לחלק לשני חלקים, למשל $\mathbb{R}^{d+k} : g$, כאשר נעשה שימוש על חלק אחד של הקוארדינטות, כלומר $g(x_1, x_2) = \min_{x_2 \in C} h(x_1)$, אז h היא גם פונקציה קמורה.
5. אם g קמורה ו- h קמורה מונוטונית לא יורדת, אז $(g \circ h)(x)$ קמורה.
6. אם h היא קמורה ומונוטונית לא יורדת בכל אחד מהארוגומנטים, ובנוסף g_i הן פונקציות קמורות, אז גם הרכבה $(g_1, \dots, g_k) \circ h$ היא קמורה.
- הוכחה**
ונוכיח רק את טענה מס' 2.
יהיו $u, v \in \mathbb{R}^d$ ו- $\alpha \in [0, 1]$. איזו מתקיים:

$$\begin{aligned}
 & \text{הגדלה} \\
 & \downarrow \\
 g(\alpha u + (1 - \alpha)v) &= f(A(\alpha u + (1 - \alpha)v) + b) \\
 &= f(\alpha(Au + b) + (1 - \alpha)(Av + b)) \\
 & \text{קמורות} \\
 & \downarrow \\
 &\leq \alpha f(Au + b) + (1 - \alpha)(f(Av + b)) \\
 & \text{הגדלה} \\
 & \downarrow \\
 &= \alpha g(u) + (1 - \alpha)g(v)
 \end{aligned}$$

דוגמה

נזכיר בפונקציית הטעות הריבועית ביחס ל- $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $(x, y) \in \mathbb{R}^d$, פונקציה f המוגדרת על ידי $f(w) = \frac{1}{2}(w - y)^2$. כיוון שהפונקציה $\langle w, x \rangle \rightarrow \langle w, x \rangle$ היא אפינית, והפונקציה הסקלרית $a \rightarrow a^2$ היא קמורה, נקבל מהטענה הקודמת כי f היא קמורה.

דוגמה נוספת (לשימוש בין תכונות שמיורות הקמורות)
הפונקציה $\log\sum_i e^{w_i^\top x + b_i}$ היא פונקציה קמורה. זהו ה- margin למשהו:

$$f(\mathbf{x}) = \log \left(\sum_{i=1}^k e^{\mathbf{w}_i^\top \mathbf{x} + b_i} \right)$$

דוגמה נוספת (hard svm)

המקסום של פונקציית הרגרסיה היא קעורה: $f(\mathbf{w}) = \min_{i \in [m]} |\langle \mathbf{w}, \mathbf{x}_i \rangle|$

דוגמה נוספת

המаксום של פונקציית הרגרסיה היא קעורה:

$$f(\mathbf{w}) = \sum_{i=1}^m \left[y_i \langle \mathbf{x}_i, \mathbf{w} \rangle - \log \left(1 + e^{\langle \mathbf{x}_i, \mathbf{w} \rangle} \right) \right]$$

1.3.1 אפייגראף (epigraph)**הגדרה**

קבוצת הנקודות הבאה תקרא האפייגראף של f :

$$\text{epigraph}(f) = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\}$$

טענה

פונקציה היא קמורה אם ורק אם האפיגרף שלה הוא קבוצה קמורה.

2 תתי גרדיאנט**2.1 תת-גרדיינט (sub - gradient)**

הרעין של תת-גרדיינטים הוא יסודי באנליה קמורה ובאופטימיזציה קמורה. זה מאפשר לנו לעבוד עם פונקציות קמורות לא גזירות.

הגדרה

תהי $f : V \rightarrow \mathbb{R}$ פונקציה. $g \in V$ יקרא **תת גראינט** של f בנקודה $u \in V$ אם לכל $v \in V$ מתקיים:

$$f(v) \geq f(u) + \langle g, v - u \rangle$$

קובוצת תת הגרדיינטים בנקודה u מוגדר על ידי $\partial f(u)$.

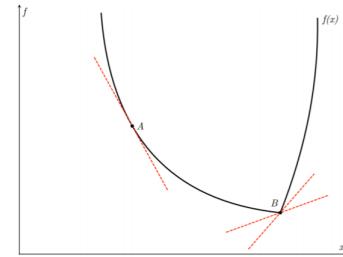


Figure 1: Subgradient

טענה

אם f קמורה וdifrenzialit בנקודה x , אז $\{\nabla f(x)\} \subseteq \partial f(x)$

באמצעות הלמה הבאה נוכל לחשב תת גראינטים.

טענה

לכל $n \geq 1$ תהי f_i פונקציה קמורה. נגדיר את $f : V \rightarrow \mathbb{R}$ על ידי $f(u) = \max_{i \in [n]} f_i(u)$. נגידר את $\partial f(u) \subseteq \partial f_i(u)$ בحينו $u \in V$, נגדיר את $j \in \operatorname{argmax}_i f_i(u)$.

הוכחה

תהי $g \in \partial f_i(u)$. לפי הבחירה של j , ההגדרה של f וההגדרה של תת גראינט, נוכל להשיג, לכל $v \in V$:

$$\begin{array}{c}
 \text{נתון} \\
 \downarrow \\
 f(v) \geq \\
 \text{הדרת תת-גרדיינט} \\
 \downarrow \\
 f_j(v) \geq \\
 \text{נתון} \\
 \downarrow \\
 f_j(u) + \langle g, v - u \rangle = f(u) + \langle g, v - u \rangle
 \end{array}$$

בעקבות כך, מתקיים כי g היא תת-גרדיינט של f בנקודה u .

דוגמה

נניח כי נוכל לכתוב את הפונקציה $f(x) = |x|$ בתור $f(x) = \max\{+x, -x\}$. מהטענה האחורונה ניתן להראות כי $\partial f(x) = \{1\}$ לכל $x > 0$ ו- $\partial f(0) = \{-1, 1\}$ ולכל $x < 0$ וגם $\partial f(x) = \{-1\}$.

ניתן להכליל זאת גם ל- $f(x) = \|x\|_1$, רק שכן קיבל כי לכל $x \in \mathbb{R}^d$ מתקיים כי $\partial f(x) = [-1, 1]^d$ ו- $\partial f(0) = \{1\}$. גם בדידות 2 נגלה כי כת הדרת הגרדיינט בכל נקודה שונה מ-0 הוא $\partial f(x) = \{x/\|x\|_2\}$, ואילו בכל נקודה שווה-

מסקנה

תהי $f : V \rightarrow \mathbb{R}$ פונקציה קמורה. נניח שעבור נקודה $\bar{w} \in V$ מתקיים כי $\partial f(\bar{w}) = \emptyset$. אז \bar{w} הוא **מיער גלובלי**.

הוכחה

לכל $w \in V$ נקבע כי:

$$\begin{array}{c}
 \text{הדרה ונתון} \\
 \downarrow \\
 f(w) \geq f(\bar{w}) + \langle 0, w - \bar{w} \rangle \\
 \text{מכפלה פנימית} \\
 \downarrow \\
 = f(\bar{w})
 \end{array}$$

2.2 תכונות של תת-גרדיינט

1. תת הגרדיינט הוא קבוצה סגורה וקמורה.

2. אם f דיפרנציאבילית ב- x אז $\{\partial f(x)\} = \{\nabla f(x)\}$ (תת הגרדיינט הוא ייחודי).

3. אם בנקודה מסוימת x תת הגרדיינט הוא ייחודי, אז f דיפרנציאבילית.

2.3 חשבון של תת גרדיאנטים

משמעותו של לב, מדובר בפעולות על קבוצות.

$$\partial(\alpha f) = \alpha \cdot \partial f \quad (\text{אם } \alpha > 0).$$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2.$$

$$\partial g(\mathbf{x}) = A^\top \partial f(A\mathbf{x} + \mathbf{b}) \quad \text{אזי } g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}). \quad 3.$$

דוגמה למתן גרדיאנט

אם נתבונן ב-SVM בהצגה האחורונה:

$$f(\mathbf{w}) = \lambda \|\mathbf{w}\|_2^2 + \sum_{i=1}^m \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}_i, y_i))$$

כאשר ℓ^{hinge} זה פונקציית ההפסד hinge. ניתן לרשום כי פונקציית hinge איננה דיפרנציאבילית. נוכל לחשב את תת הגרדיינט. אפשר להראות כי אם עבור w מתקיים כי $\langle w, x \rangle \geq 1$ ו- $y \in \partial f(w)$ מתקיים כי $\langle w, x \rangle < 0$ וכאשר y מתקיים כי $\langle w, x \rangle < 1$. בעקבות כך, עבור וקטור w , קיבל כי:

$$\mathbf{v} = 2\lambda \mathbf{w} - \sum_{i:y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 1} y_i \mathbf{x}_i$$

שייך לתת גרדיאנט.

3 תנאים מסדר גבוה לקמירות

טענה - תנאי מסדר ראשון

פונקציה $f : \mathbb{R}^n \rightarrow \mathbb{R}$ דיפרנציאבילית היא קמורה אם ורק אם מתקיים אי השוויון הבא לכל y, x בתחום:⁴⁰

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

דוגמה

יהי $f : \mathbb{R}^n \rightarrow \mathbb{R}$ על ידי $f(x) = \langle c, x \rangle$ ונגדיר $c \in \mathbb{R}^n$ שכל $x \in \mathbb{R}^n$ קמורה כיוון $\nabla f(x) = c$ נקבל:

$$\begin{aligned} f(y) &= c^T y = c^T x + c^T(y - x) = \\ &= f(x) + \nabla f(x)^T(y - x) \end{aligned}$$

⁴⁰ דומה למשיק ושיפוע שהוא בתיכון. צריך להחליף את $\nabla f(x)^T$ בנגזרת המוכרת לנו.

טענה - תנאי מסדר שני

פונקציה $f : \mathbb{R}^n \rightarrow \mathbb{R}$ דיפרנציאבילית פעמיים הינה קמורה אם ורק אם לכל נקודה x בתחום, פונקציית ההסיאן חיובית למחצאה. דהיינו $\nabla^2 f(x) \succeq 0$ ⁴¹

דוגמה

תהי מטריצה סימטרית ונגידר את $f(x) = x^T A x$. נחשב את הנגזרות הראשונות והשניות של f .

הוכחה⁴²

נגידר $f^\top g : \mathbb{R}^n \rightarrow \mathbb{R}$, כך שנקבל $f, g : \mathbb{R}^m \rightarrow \mathbb{R}^m$

כעת, נוכל לקבל:

$$\begin{aligned} f^\top \cdot g &= \sum_{i=1}^m f_i \cdot g_i \Rightarrow \\ \frac{\partial}{\partial x_j} f^\top g &= \sum_{i=1}^m \left(\frac{\partial}{\partial x_j} f_i \cdot g_i \right) = \sum_{i=1}^m \frac{\partial f_i}{\partial x_j} g_i + \sum_{i=1}^m \frac{\partial g_i}{\partial x_j} f_i = \\ \left[\frac{\partial f_1}{\partial x_j} \dots \frac{\partial f_m}{\partial x_j} \right] \cdot g(x) + \left[\frac{\partial g_1}{\partial x_j} \dots \frac{\partial g_m}{\partial x_j} \right] \cdot f(x) &\Rightarrow \\ = \nabla f^\top \cdot g &= \begin{bmatrix} \frac{\partial f \cdot g}{\partial x_1} \\ \vdots \\ \frac{\partial f \cdot g}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \dots \frac{\partial f_m}{\partial x_1} \\ \vdots \\ \frac{\partial f_1}{\partial x_n} \dots \frac{\partial f_m}{\partial x_n} \end{bmatrix} \cdot g(x) + \begin{bmatrix} \frac{\partial g_1}{\partial x_1} \dots \frac{\partial g_m}{\partial x_1} \\ \vdots \\ \frac{\partial g_1}{\partial x_n} \dots \frac{\partial g_m}{\partial x_n} \end{bmatrix} f(x) = \end{aligned}$$

$$(J_x(f))^\top \cdot g(x) + (J_x(g))^\top \cdot f(x) =$$

$$= \left(\frac{\partial f}{\partial x} \right)^\top g(x) + \left(\frac{\partial g}{\partial x} \right)^\top f(x) =$$

כעת, נסמן $f(x) = h^\top g$ ונקבל $h(x) = x$ -ו $g(x) = Ax$ וס"כ הכל:

$$\begin{aligned} \nabla f &= \nabla(h^\top \cdot g) = \\ \left(\frac{\partial h}{\partial x} \right)^\top g(x) + \left(\frac{\partial g}{\partial x} \right)^\top h(x) & \end{aligned}$$

נזכור כי הנגזרת של $g(x)$ הינה A (ראינו זאת) והנגזרת של $h(x)$ הינה I ולכן נקבל:

$$\begin{aligned} \left(\frac{\partial h}{\partial x} \right)^\top g(x) + \left(\frac{\partial g}{\partial x} \right)^\top h(x) &= \\ I \cdot Ax + A^\top \cdot x &= \\ (A + A^\top) x & \end{aligned}$$

⁴¹תזכורת, זו הדרך בה אנחנו מסיימים חיובית למחצאה או בהחלט במטריצות.

⁴²בhocחה שבכיתה יש טעות. זו hocחה אחרת.

ניתן להסתכל על $A + A^\top$ בטור מטריצה ולבן הנגזרת של x ($A + A^\top$), שהינה הנגזרת השנייה, כלומר ההessian של הפונקציות המקוריות, הינה $A + A^\top$ שהינה $2A$ כיוון A -סימטרית. בהשتمש בטענה לעליה, נוכל לבדוק כי f הינה קמורה אם ורק אם A חיובית למחצית, כאמור.

במקרה פשוט של $\mathbb{R} \rightarrow \mathbb{R}$: הטענה זו שוללת במקרה הפחות של $0 \geq f''$.

4 בעיות אופטימיזציה

הרבה בעיות שנתקל בהן בקורס זה הן משפחה רחבה של בעיות שנקרוות בעיות אופטימיזציה קמורה.

הגדרה

בעית אופטימיזציה מעל \mathbb{R}^d היא מהצורה הכללית - מזער (x_0) בcpf (subject) לפונקציות האילוץ $f_i \leq b_i$ עבור $1 \leq i \leq n$. x הוא משתנה האופטימיזציה, $\mathbb{R}^d \rightarrow \mathbb{R}$: $f_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ הינו פונקציות האילוץ.

מכך עולה כי בעית האופטימיזציה הינו בתחום של f_0 שמוסכל ב- \mathbb{R}^d . אם כך, בעית אופטימיזציה קמורה הינו בעית אופטימיזציה כאשר f_0, f_1, \dots, f_n , כולל פונקציות קמורות. כאשר כל הפונקציות הללו הן ליניאריות, זו נחבבת בעית אופטימיזציה ליניארית.

אם לכל נקודה $x \in D$ כאשר $x \leq g_i$ הנקודה נקראת נקודה פיזיבלית (בה האילוצים מתקיים). הערך האופטימלי (אופטימום) של f מעל כל הנקודות הפיזיבליות נקרא הערך האופטימלי והואו אנחנו רוצים לחפש. אם x היא נקודה פיזיבלית ו- $f^* = f(x)$ (הערך המינימלי שהפונקציה יכולה לקבל) אז x נקרא נקודה אופטימלית, פיתרון או מזער.

קבוצות כל הפתרונות לעית אופטימיזציה קמורה היא קבוצה קמורה. כאשר נסמן ב- C את כל הנקודות הפיזיבליות, נוכל לומר כי אנו מנסים למצער את (x) בcpf לאילוץ כי $x \in C$.

הגדרה

בעית אופטימיזציה נקראת **תכון ליניארי** אם היא יכולה להיכתב בצורה הבא: $\min c^\top x$ כך $\leq b$. כאשר $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ הם מטריצות וקטורים קבועים.

הגדרה

בעית אופטימיזציה תיקרא **תכון ריבועי** אם היא יכולה להיכתב בצורה $\min w^\top Qw + a^\top w$ כך $\leq d$. כאשר $A \in \mathbb{R}^{m \times n}$, $d \in \mathbb{R}^m$ ו- $Q \in \mathbb{R}^{n \times n}$, $c \in \mathbb{R}^n$, הם מטריצות וקטורים קבועים.

בדרכם כלל, בעית אופטימיזציה קשות לפיתרון, ולכן בעית אופטימיזציה קמורות, להן יש פיתרון יחיד. ניתן לפתור זאת באמצעות אלגוריטמי בעית אופטימיזציה, שחלקים כלליים וחילוקם פוטרים בעיות ספציפיות. האלגוריתמים האחרונים עדיפים, כיון שהם משתמשים במאפיינים של הבעיה ליעיל את האלגוריתם. בעית אופטימיזציה קמורות מתקשרות למקרה כיון שהרבה מאלגוריתם למידה יכולים להיכתב בעית אופטימיזציה, למשל מזער של כמה מסויימת (כלומר לבחור $\mathcal{H} \in h$ - מתקנת היפותזות).

לפעמים מחלוקת ההיפותזות שקופה למרחב האוקלידי ואז נוכל לפתור זאת בתור בעיית אופטימיזציה, ככלומר נבחר $\mathcal{H} \in h$ שתהיה המינימלית מעל \mathbb{R}^d או תת קבוצה של \mathbb{R}^d של פונקציית יעד כלשהיא. אם פונקציית היעד היא קמורה, נוכל להשתמש באלגוריתמי אופטימיזציה קמורה.

דוגמה

נזכיר בבעיה. נתנות לנו דוגמאות $\{y_k\}_{k=1}^n \subseteq \mathbb{R}^n$ ו- $\{x_k\}_{k=1}^n \subseteq \mathbb{R}^d$ ונניח כי $y_i = \langle w, x_i \rangle + b$ עבור $w \in \mathbb{R}^d$ ו- $b \in \mathbb{R}$. נרצה לモודר את פונקציית ה-MSE.⁴³ נניח כי $x = (3, 6, 9)$, $y = (4, 6, 8)$, נמצא את הפרמטרים של המודל באמצעות הגרדיאנט.

הוכחה

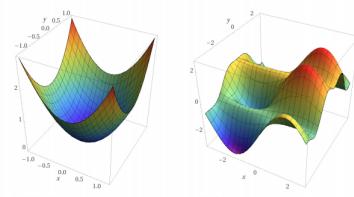
תחילה, נמצא את הגרדיאנט. נקבל:

$$\text{MSE}(w, b) = \frac{1}{N} \sum_{i=1}^N (y_i - (wx_i + b))^2$$

ולכן הנגזרת:

$$\begin{aligned} \frac{\partial \text{MSE}}{\partial w} &= \frac{2}{N} \sum_{i=1}^N -x_i (y_i - (wx_i + b)) = \\ &= \frac{2}{N} \sum_{i=1}^N -x_i y_i + wx_i^2 + x_i b = -80 + 84w + 12b \\ \frac{\partial \text{MSE}}{\partial b} &= \frac{2}{N} \sum_{i=1}^N (y_i - (wx_i + b)) = \\ &= \frac{2}{N} \sum_{i=1}^N -y_i + wx_i + b = 2b - 12 + 12w \end{aligned}$$

בכל צעדי שנעשה, נקטין את הערך של הפסד. אמן, עדין לא מובטח לנו שנגיע למינימום מקומי שאיינו מינימום גלובלי. אמן, דבר זה לא יכול לקרות כאשר f קמורה. לא יתכן שהיא מינימום מקומי שאיינו מינימום גלובלי. ניתן לראות זאת גם בציורים הבאים:



הfonקציית הימנית איינה קמורה (ולכן יש לה מינימום מקומי שאינו גלובלי), ואילו הפונקציה השמאלית הינה קמורה.

אלגוריתם לפתרון בעיות אופטימיזציה קמורה

פותר (solver) של בעית אופטימיזציה קמורה הוא אלגוריתם שמקבל כקלט בעיה קמורה ומיציא פתרון אופטימלי. מדובר בתחום עשיר מאוד. הדרך לפתרו אותו היא באמצעות סדר ראשון או שני כפי שראינו קודם.

⁴³תזכורת. טעות ריבועית ממוצעת, המוגדרת על ידי $\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$

4.1 הקשר בין למידה ובין אופטימיזציה קמורה

מה הקשר בין אופטימיזציה קמורה ובין למידת מכונה? האם בעית למידה שהיא בעית אופטימיזציה קמורה היא למידה PAC?

נubbyor לרגע להתעסק בלמידה. כדי שיכלול נוכל לדבר על אופטימיזציה קמורה בהקשר של בחירת היפותזה מתוך מחלוקת היפותזות, עלינו להזות את מחלוקת היפותזות עם \mathbb{R}^d (לדוגמא כפי שראינו את מחלוקת היפותזות של החצאי מרחבים).

נזכיר כי $\mathcal{H} \times \mathcal{Z} = Z$. נאמר כי בעית למידה (\mathcal{H}, Z, ℓ) נקרהת קמורה אם אם מחלוקת היפותזות היא **קבוצה קמורה** וגם לכל $Z \in \mathcal{Z}$ פונקציית הפסד $\ell(\cdot, \cdot)$ היא פונקציה קמורה. למשל, ראיינו שאפשר להשתמש ב-ERM, ולכן מדובר בעית למידה קמורה. מה התנאי למידה PAC? עלינו להוסף 2 תנאים:

1. \mathcal{H} היא חסומה.⁴⁴

2. פונקציית הפסד היא ליפשיצית, כלומר $|f(\mathbf{w}_1) - f(\mathbf{w}_2)| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$ עבור קבוע ρ כלשהו.

תרגיל

תהי f פונקציה קמורה. הראו כי f היא ρ -ליפשיצית אם ורק אם הנורמה של כל תת-הגרדייאנטים של f היא לכל היותר ρ .

הגדרה

בעית למידה חסומה-קמורה-ליפשיצית היא בעית למידה (\mathcal{H}, ℓ, Z) עם הפרמטרים B, ρ אם מתקיימים התנאים הבאים:

□ \mathcal{H} היא קמורה ולכל $\mathcal{H} \in \mathcal{H}$ מתקיים כי $\|\mathbf{w}\| \leq B$.

□ לכל $Z \in \mathcal{Z}$ פונקציית הפסד $\ell(\cdot, \cdot)$ היא קמורה ו- ρ ליפשיצית.

המשפט המרכזי

בתנאי לעיל, מחלוקת היפותזות היא למידה PAC וסיבוכיות המדגם תלולה ב- $\delta, \varepsilon, \rho, B$.

5 מورد הגרדייאנט (Gradient descent)

מדוע עניין אותנו להתעסק ב-GD? יש יתרון בידעו solver בלבד. במיוחד אם יש לנו איזושהי פונקציית הפסד מוזרה. מעבר לכך, לעיתים קבות האימון שלו מאוד גדולה ובלתי אפשרי לפתור זאת באמצעות חבילות למידה, אלא יש לנו צורך לכתוב solver בעצמו. במיוחד אם משתמש במידע שורות כל הזמן, נדרש ל充满 את solver העצמנו.

חשוב לציין כי אלגוריתמי הלמידה שרצים על מידע גדול (מה שנקרא big data), משתמשים ב-GD. חלק שימושותי מהמערכות שאנו משתמשים בהם בימיום משתמשים ב-GD.

⁴⁴לא ראיינו הרבה מחלוקת היפותזות שחסומות בתחום \mathbb{R}^d .

5.1 גראדיאנט והקשר לשיפוע

נזכור כי אם f היא דיפרנציאבילית ב- x , איי (x) ∇f , הגרדיאנט מצביע על הכיוון של העלייה התולולה ביותר, ובצורה הפוכה (x) $-\nabla f$.

אנו יודעים כי f מאונכת לכל המשיקים לקווי הגובה (באנגלית - level set).

5.1.1 תנאי מסדר ראשון לאופטימליות

עבור בעיית אופטימיזציה קמורה של מזעור f , אם f היא דיפרנציאבילית, נקודה x היא אופטימלית אם ורק אם:

$$\langle \nabla f(x), y - x \rangle \geq 0 \quad \forall y \in C$$

דבר זה נקרא **תנאי ראשון לאופטימליות**.

למעשה, אנחנו משתמשים בנקודה x ובבודקים אם בכיוון כלשהו (h) הנקודה שמתקבלת היא גם פיזיבלית ($x + h$), אין רכיב שנמצא ב- $(x) \nabla f$. כלומר, אי אפשר לרדת בערך הפונקציה מ- x . המקרה המינוחד הוא בו אין אילוצים, אז הנקודה x היא אופטימלית אם ורק אם הוא מאונך לכל דבר, כלומר הוא וקטור האפס.

אם x הוא לא נקודה אופטימלית, אז יש h כך $h + x$ היא פיזיבלית אבל $\langle \nabla f(x), h \rangle < 0$. h נקראת במקרה זה "כיוון ירידה" (descent direction).

דוגמאות

אם נתבונן בתבנית הריבועית, ובבעיה האופטימיזציה ללא אילוצים, שמנסה למצוא את c , כך ש- $b + cx$ היא חיובית בהחלט, אז:

$f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ כאשר $Q \in \mathbb{R}_+^d$, $b \in \mathbb{R}^D$, $c \in \mathbb{R}$

באמצעות התנאים מסדר ראשון להראות כי אם Q היא חיובית בהחלט, אז:

■ הנקודה האופטימלית היא יחידה $x = Q^{-1}b$.

■ אם Q סינגולרית, ו- $b \notin \text{Im}(Q)$, אז אין נקודות אופטימליות.

■ אם Q היא סינגולרית, אבל $b \in \text{Im}(Q)$, אז $x = Q^\dagger b + z$ כאשר $z \in \ker(Q)$, היא אופטימלית.

מכאן ניתן להראות הרבה דברים שראינו לגבי רגסיביות ליניארית.

5.2 האלגוריתם

לאחר שראינו את התנאי מסדר ראשון לאופטימליות, נוכל להתעסק באלגוריתם עצמו של GD. אם יש אייזהו "כיוון ירידה" (x) ∇f – הוא בטוח כיון ירידה גם. השאלה היחידה היא "כמה ליכת". אם נלק יותר מדי, אנו יכולים לעبور את המינימום, ואם נלק מעט מדי, זה עלול להיות לא מועיל.

1. נתחל עם $x \in \mathbb{R}^d$.

2. בכל איטרציה t נעדכן: $x^{(t+1)} = x^{(t)} - \eta_{t+1} \nabla f(x^{(t)})$.

3. תנאי העצרה בזמן T – כאשר הגרדיאנט $\|\nabla f(x^{(t)})\|$ מספיק קטן.

4. פלט האלגוריתם: ליצא את $\mathbf{x}^{(T)}$, וקטור הממוצע $\bar{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}^{(t)}$, או הוקטור הטוב ביותר ביחס ל- t^*

$$\text{הו} \cdot \text{argmin}_{1 \leq t \leq T} f(\mathbf{x}^{(t)})$$

המספרים η נקראו גדי צעדי הגרדייאנט (gradient step sizes).

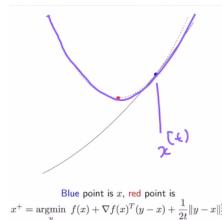
ראינו קודם לכן את תנאי מסדר שני (במקרה בו הפונקציה דיפרנציאבילית):

$$f(\mathbf{y}) \approx f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x}) (\mathbf{y} - \mathbf{x})$$

אבל במקרה לא יכולimos להניח שהפונקציה דיפרנציאבילית פעמיים או שאנו יודעים את ההסיאן ולכון נוכל להחליפה זאת עם $\frac{1}{\eta} I$:

$$f(\mathbf{y}) \approx f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle + \frac{1}{2\eta} \|\mathbf{y} - \mathbf{x}\|^2$$

מדובר בהערכת ריבועית נאיבית של f סביב א. כאשר נסתכל על קירוב, אנחנו משתמשים על $\mathbf{x}^{(t)}$ ומתחשים אחר פרבולoid משיק לנקודה:



5.3 איך נבחר את η ?

אם נבחר גודל צעד קבוע, זה עלול לא לעבוד היטב. בכל פעם אנחנו עלולים לקפוץ מעל האופטימום. אם גודל הצעד קטן מדי, אי נתקדם לאט מדי. עדיף לבחור את גודל הצעד דומה למה שעשינו בפרספורטונג, בצורה אדיטיבית.

ישנו מספר שיטות לישם זאת, אנחנו השתמש בשיטה הבאה.

5.3.1 השיטה Backtracking line search

מדובר על שיטה אדיטיבית לבחירת η ב-GD. נסמן את $(\mathbf{x} - \nabla f(\mathbf{x}))^\top \nabla f(\mathbf{x})$ ומעשה זיזים מ- \mathbf{x} ל- $\mathbf{x} - \nabla f(\mathbf{x})$. היינו רוצים לבחור את η ממשפיק גדול בשביל שנקבל ירידה בגודל הפונקציה, אבל לא גדול מדי, בשביל שלא נפספס את המינימום.

אבחנה

כיוון שהפונקציה קמורה ישנו איזושהו על מישור ממשיק שנמצא מתחת לגרף הפונקציה. בעקבות כך, אם נבחר $\alpha \leq 0$, אז בהכרח נהייה מעל גרף הפונקציה:

$$f(x) + \alpha\eta\langle \nabla f(\mathbf{x}), \Delta\mathbf{x} \rangle > f(\mathbf{x} + \eta\Delta\mathbf{x})$$

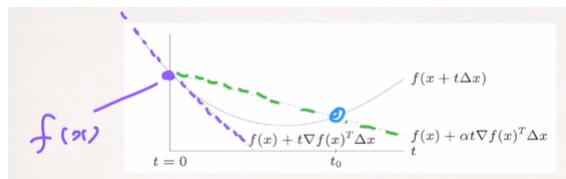
אמנם, אם נגדיל את α , נגיע למצב שנהיה מתחת לגרף הפונקציה:

$$f(x) + \alpha\eta\langle \nabla f(\mathbf{x}), \Delta\mathbf{x} \rangle < f(\mathbf{x} + \eta\Delta\mathbf{x})$$

אזי ישנו ערך η_0 שעבורו מתקיים כי:

$$f(x) + \alpha\eta_0\langle \nabla f(\mathbf{x}), \Delta\mathbf{x} \rangle = f(\mathbf{x} + \eta_0\Delta\mathbf{x})$$

בציר:



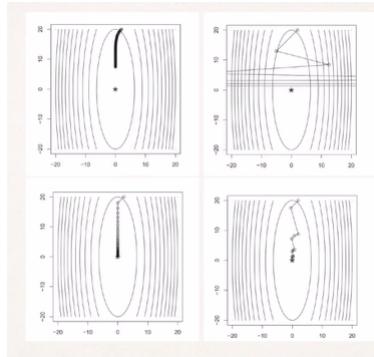
זה מה שאנחנו למבצע עושים ב-Is, מנסים למצוא קירוב לנקודה זאת. אם נעשה זאת מספר פעמים נוכל להתכנס למינימום:



כיצד נוכל לעשות זאת באופן מהיר? נבחר פרמטר $1 - \eta < \beta < 0$. נתחל מ- $t = 0$ ו בכל איטרציה נעבור מ- t ל- $t - \eta \cdot \beta$ עד שנגיע לתנאי העזירה:

$$f(x) + \alpha\eta\langle \nabla f(\mathbf{x}), \Delta\mathbf{x} \rangle < f(\mathbf{x} + \eta\Delta\mathbf{x})$$

כאן ניתן לראות את ההשוואה בין Is:



בצד שמאל למעלה הצעד קטן מדי, בצד ימין למעלה הצעד גדול מדי, הצד שמאל למטה הצעד טוב אבל קשה מדי לגילוי, ובצד ימין למטה זה לא.

5.4 התכנסות של GD

כיצד והאם ניתן להוכיח כי GD מתכנס לנקודה אופטימלית? סוג משפטים זה נקרא "משפטים של התכנסות" ולמעשה ניתן לנתח כל אלגוריתם אופטימיזציה איטרטיבי האם הוא מתכנס ותוך כמה זמן הוא מתכנס.

משפט פשוט לתהיליך התכנסות של GD
 תהי f קמורה ודיפרנציאבילית בתחום \mathbb{R}^d . נניח כי ∇f היא ליפשצית עם קבוע L , כלומר כי $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$ לכל $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
 יהיו \mathbf{x}^* הערך האופטימלי של f ותהי \mathbf{x}^* נקודה אופטימלית. אזי GD עם גודל צעד קבוע מקיים כי:

$$f(\mathbf{x}^{(t)}) - f^* \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2}{2\eta \cdot t}$$

כלומר, במקרה זה, כאשר $t \rightarrow \infty$ אכן יש התכנסות.

5.4.1 שימוש ב-GD לבעיות עם אילוצים

האם אפשר להשתמש ב-GD גם לבעיות עם אילוצים? כן.
 נתבונן בבעיות האופטימיזציה הכלילית הבאה:

$$\min_w f(\mathbf{w}) \quad \text{subject to } \mathbf{w} \in C$$

יהי P_C אופרטור החטלה על הקבוצה C , אשר מוגדר על ידי:

$$P_c(\mathbf{x}) = \operatorname{argmin}_{\mathbf{w} \in c} \|\mathbf{x} - \mathbf{w}\|_2$$

כלומר, הוא מוצא את הנקודה הקרובה ביותר בקבוצה הקמורה. דבר זה בעצם הוא בעיה קמורה. בעקבות כך, נוצר לייצר את ה-GD המוטל (projected GD).

$$\mathbf{x}^{(t+1)} = P_C \left(\mathbf{x}^{(t)} - \eta_{t+1} \nabla f \left(\mathbf{x}^{(t)} \right) \right)$$

אנו עושים את צעד הגרדיינט, ורק מקיים 'מטיליט' לתוך הקבוצה C . מדובר על אלגוריתם יעיל כיון שקל לחשב את P_C .

6 מورد התת-גרדיינט (Sub-Gradient descent)

מה נעשה כאשר הפונקציה לא דיפרנציאבילית? האם נוכל להשתמש בתת-גרדיינט?

נגיד תנאי דומה למזה שעשינו בגרדיינט.

נאמר כי לכל פונקציה (לא משנה אם קמורה או לא), מתקיים כי $\mathbf{x}^* = \min f(\mathbf{x})$ אם ורק אם $f(\mathbf{x}^*) \leq f(\mathbf{x}) \forall \mathbf{x} \in \partial f(\mathbf{x}^*)$.

זה נובע ישירות מההגדרות של תת גרדיינט שראינו, כי $\mathbf{x}^* \in \partial f(\mathbf{x}^*)$ אם ורק אם לכל \mathbf{v} מתקיים כי $\langle \mathbf{x} - \mathbf{x}^*, \mathbf{v} \rangle \geq 0$.

אם במקרה הפונקציה גם דיפרנציאבילית, יש לה גרדיינט, ונחזור לנקודה כי הגרדיינט בנקודה האופטימלית הוא.

תוגיל

תהי f דיפרנציאבילית ונניח שיש לנו בעיית אופטימיזציה של מזעור f . הוכיחו את תנאי אופטימליות מסדר ראשון ($\min f(\mathbf{x}) + 1_C(\mathbf{x})$) שראינו קודם, בהתבסס על התנאי לאופטימליות של תת גרדיינט, עבור הבעיה השוקלה $\langle \mathbf{x}, \mathbf{v} \rangle \geq 0$ האינדיקטור של C .

ניתן לראות דוגמה מדוודו התנאי של אופטימליות לתת-גרדיינט מעניין.

דוגמה

אם ניקח את בעיית הלאסו:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

אנו יכולים לראות שהפונקציה לא דיפרנציאבילית בכלל הנורמת 1.

נוכל לראות שתת הגרדיינט הינו:

$$\partial \left(\frac{1}{2} \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 \right) = -X^\top (\mathbf{y} - X\mathbf{w}) + \lambda \partial \|\mathbf{w}\|_1$$

ונשים את התת הגרדיינט אם ורק אם מתקיים כי יש $\mathbf{v} \in \partial \|\mathbf{w}\|_1$ כך ש- $\mathbf{v} - X^\top (\mathbf{y} - X\mathbf{w}) = \lambda \mathbf{v}$ אם ורק אם:

$$v_i \in \begin{cases} 1 & w_i > 0 \\ -1 & w_i < 0 \\ [-1, 1] & w_i = 0 \end{cases}$$

אם נסמן את (i) בטור העמודות ה- i של X , נוכל להוכיח כי w היא נקודת אופטימלית של בעיית הלאסו אם ורק אם $\lambda \cdot \text{sign}(w_i) = \lambda \cdot \langle \phi_i, y - Xw \rangle \leq 0$ וכן $\lambda \cdot \langle \phi_i, y - Xw \rangle = 0$.

ניתן לראות שהוקטורים שהלאסו לא בחר אותם מקיימים את התנאי הראשון, והקטורים שהלאסו בחר אותם מקיימים את התנאי השני. בפרט, ניתן להראות על ידי חישוב ישיר כי מדובר על הכללה של ההוכחה שראינו בתרגול במקרה זה.

אם ניקח $\lambda = 0$, נקבל כי כל הפיצ'רים מאונכים לפתרון (ראינו זאת בהרצאה של רגרסיה ליניארית). אם $\lambda \neq 0$, נקבל כי יש לפיצ'ר $-i$ זווית קבועה עם $Xw - y$ שקיים הזווית שלה הינה $\lambda \cdot \text{sign}(w_i)$.

כיצד נוכל להתאים כתעת האלגוריתם למקרה בו הפונקציה לא דיפרנציאבילית? אנחנו רוצים לנوع מסויל בו תתי הגרדיינט הולכים וירדים. לכן ניקח איזוחה תת-גרדיינט מתוך הקבוצה. בוגוד למקרה של גרדיינט, כאן אף אחד לא מבטיח שהפונקציה בהכרח תרד - לכן אי אפשר להחזיר את הוקטור האחרון, אך נוכל להחזיר את הוקטור המומוצע או את הוקטור הטוב ביותר. ראיינו למעשה את SGD כבר בפרסרטרון.

6.1 גודל הצעדים

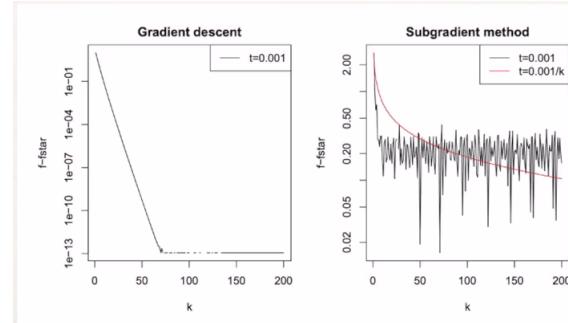
כיוון שכפי שראינו אין בהכרח כיוון ירידה, אין שיטות אדפטיביות (כמו Line Search), ולכן מראש מהו גודל הצעד. כדי לבחור צעדים שהולכים ומתקרבים ל-0, כך שכל שהאיטרציה מתקדמת, נעשו צעדים יותר ויותר קטנים. מצד שני, לא כדאי שהם יתקרבו ל-0 מדי. לכן, אפשר לבחור למשל:

$$\sum_{t=1}^{\infty} \eta_t^2 < \infty \quad \sum_{t=1}^{\infty} \eta_t = \infty$$

זה אומר שהסדרה יורדת ל-0 אבל לא יורדת מהר מדי. שאלת החתכנות כרגע היא קריטית, כיוון שלא ברור לנו האם נוכל להגיע לפונקציה. **משפט** נניח כי f היא פונקציה קמורה ונניח שהתחום שלה ב- \mathbb{R}^d . נניח גם כי f היא L -ליפשצית. יהיו x^* הערך האופטימלי של f ותהי x_{best} - הנקודה שבה פונקציה קיבלה את הערך הנמוך ביותר. מובטח בהכרח כי:

$$\lim_{t \rightarrow \infty} f(x_{\text{best}}^{(t)}) = f^*$$

שיטת זו אכן מתכנסת לאט יותר מאשר הגרדיינט. ניתן לראות כי פרופיל התכנסות של GD פחות רגולרי מאשר אצל SGD:



דבר זה מעלה שאלת על מהו תנאי עצירה נכון ל-SGD.

7 שיטת Stochastic Gradient Descent

הגדרה אבסטרקטית
יהי G משתנה מקרי מעל \mathbb{R}^d . נאמר כי האיטרציה:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \mathbf{g}^{(t)}$$

היא SGD אם לכל t קיבל כי בתוחלת המשתנה המקרי שיקף לתת הגרדיינט:

$$\mathbb{E} [G^{(t)}] \in \partial f(\mathbf{x}^{(t)})$$

כלומר, אנחנו מגרילים משהו שבממוצע הולך למזה שאנו רוצים. בסופו של דבר נעצור ונחזיר את הווקטור הממוצע (ווקטור האחרון יש הטיה גדולה מדי).

אם f דיפרנציאבילית, אז כמובן שמתקיים כי

דוגמא

נិיח למשל פונקציה $f(\mathbf{w}) = \sum_{i=1}^m f_i(\mathbf{w})$. מכיוון $\nabla f(\mathbf{w}) = \sum_{i=1}^m \nabla f_i(\mathbf{w})$ קיבל כי:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \sum_{i=1}^m \mathbf{g}_i^{(t)}$$

כאשר $\mathbf{g}_i^{(t)} \in \partial f_i(\mathbf{x}^{(t)})$ נניח כי קיימת משתנה מקרי $k(t)$ שמתפלג בצורה בלתי תלולה ושוות התפלגות מעל m , אז האיטרציה:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_{t+1} \mathbf{g}_{k(t)}^{(t)}$$

היא איטרצית SGD. כמובן, הוא מקיים את התנאי, ובתוחלת אכן נהיה בתוך התת-גרדיינט.

דוגמה נוספת

במקרה להגריל פונקציה אחת, נוכל להגריל מספר פונקציות. כמובן, באיטרציה נקבל:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\eta_{t+1}}{|K(t)|} \sum_{j \in K(t)} \mathbf{g}_j^{(t)}$$

כאשר $K(t)$ היה תת-קובוצה שמתפלגת בצורה אחידה. דבר זה נקרא Random mini-batch כיון שהוא מתפלג רנדומלית ו-mini-batchmini- מינימום וلومדים על פיו.

7.1 שימוש ב-SGD לפתרון בעיות למיניה ק她们

ראינו כי במקרה בו אנו מנסים לסייע את $L_s(w)$ (תוחלת פונקציות ההפסד) ומדובר בפונקציה קמורה, קל לנו לחסית לישם זאת. אנחנו יכולים להשתמש ב-SGD על מנת לחסוך הרצות, ונקבל, אם ניקח z_1, \dots, z_T דוגמאות שנוצרות בצורה בלתי תלולה ושוות התפלגות:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_{t+1} \mathbf{g}^{(t)}$$

ולכל t יתקיים כי $\mathbf{g}^{(t)} \in \partial \ell(\mathbf{w}^{(t)}, z_t)$. מבחינה אינטואיטיבית, אנחנו לא עושים גרדיינט על כל הנקודות אלא רק על חלק, וזה כיוון שבתוחלת זה מתקיים.⁴⁵

בכל מקרה, מה שאנו עושים הוא לבחור נקודה ובכל צעד גרדיינט אנחנו מחשבים את הגרדיינט ביחס לכל נקודה שבחרנו. כיון שאנו בוחרים נקודה יחידה, עדיף לנו לבחור מספר רב גדול יותר כדי שלא תהיה שונות, ונמצע את התוצאה:

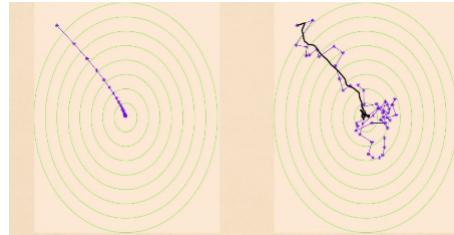
$$\frac{1}{|B|} \sum_{i \in B} \nabla \ell(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i))$$

⁴⁵עקרון זה מותיחסיפה עם הרעיון של מחשבים, כיון שאפשר למקבל זאת, ולאמון כמויות גדולות של מידע.

אפשר לפתח על דבר זה מספר וריאציות.
למשל, ישנה גרסה ציקלית בה בכל פעם זרים צד אחד במרחב המדגם, או לעשות זאת עם B נקודות בכל פעם.
כמוות האיטרציות שלוקת לנו לעבור על המידע נקראת epoch.

חשוב להזכיר כי אנחנו פעם ראשונה בקורס לא מדברים על Batch Learning. זה טוב למשל במקרה בו המידע זורם כל הזמן. כמובן, ניתן לשפר את המידע כל הזמן.

נראה כעת את ההבדל בין SGD ובין GD (בצד שמאל GD ובצד ימין SGD):



בשילוב מופיע הממוצע, וראים כיצד לבסוף מגיעים למינימום.

אופטימיזציה לשגיאת הכלכלה

בלי ששמנו לב, ניתן לראות כי אנחנו למעשה עושים אופטימיזציה לשגיאת הכלכלה.
נזכר רגע במקורו של PAC אגנוסטי. אנחנו מניחים כי $\mathcal{D} \sim \mathcal{N}(x, y) = z$ עבר פונקציית הסתברות \mathcal{D} מעל $\mathcal{Y} \times \mathcal{X}$.
המטרה שלנו היא למצוא את $\mathcal{H} \in \mathcal{W}$ שמצויר את $L_{\mathcal{D}}(\mathbf{w}) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\mathbf{w}, z)]$
מכיוון ששגיאת הכלכלה היא התוחלת של פונקציית הփס (loss)zioni הגראדיאנט של שגיאת הכלכלה, הוא הגראדיאנט של התוחלת על פונקציית הփס ולכן נקבל (מהגדרת סכום של נזורת) את התוחלת של הגראדיאנט על הփס:

$$\nabla L_{\mathcal{D}}(\mathbf{w}) = \nabla \mathbb{E}_{z \sim \mathcal{D}}[\ell(\mathbf{w}, z)] = \mathbb{E}_{z \sim \mathcal{D}}[\nabla \ell(\mathbf{w}, z)]$$

ניתן לראות כי אם ניקח נקודת מקרית ונעשה צעד בכיוון של הגראדיאנט, קיבלנו משחו שבתוחלת הוא בכיוון של שגיאת הכלכלה.

משפט

אם נרייך את SGD על בעיות קמורות-ליפשציות-חסומות עם הפרמטרים B, ρ, α לכל $0 < \epsilon$, נקבל, לאחר $T \geq \frac{B^2 \rho^2}{\epsilon^2}$ איטרציות ו- $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ כי:

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon$$

זה לא מפתיע, כי ראיינו כבר שניתנו לפטור בעיות אלו בצורה יعلاה.

אלגוריתם למימוש SGD עבור Soft-SVM

```

SGD for Solving Soft-SVM
goal: Solve Equation (15.12)
parameter: T
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 פעם אנחנו בוחרים נקודת ספציפית, ולבסוף עושים ממוצע לכל הנקודות.
ונכל גם לעשות Kernel ל-Soft-SVM

```

SGD for Solving Soft-SVM with Kernels
Goal: Solve Equation (16.5)
parameter: T
Initialize:  $\beta^{(1)} = \mathbf{0}$ 
for  $t = 1, \dots, T$ 
    Let  $\alpha^{(t)} = \frac{1}{\lambda t} \beta^{(t)}$ 
    Choose  $i$  uniformly at random from  $[m]$ 
    For all  $j \neq i$  set  $\beta_j^{(t+1)} = \beta_j^{(t)}$ 
    If  $(y_i \sum_{j=1}^m \alpha_j^{(t)} K(\mathbf{x}_j, \mathbf{x}_i) < 1)$ 
        Set  $\beta_i^{(t+1)} = \beta_i^{(t)} + y_i$ 
    Else
        Set  $\beta_i^{(t+1)} = \beta_i^{(t)}$ 
Output:  $\bar{\mathbf{w}} = \sum_{j=1}^m \bar{\alpha}_j \psi(\mathbf{x}_j)$  where  $\bar{\alpha} = \frac{1}{T} \sum_{t=1}^T \alpha^{(t)}$ 

```

8 למידה عمוקה (deep learning)

למידה عمוקה נמצאת בכל מקום אצלונו ביום יום. למשל, מערכות זיהוי פנים, ניצחון של אלוף עולם בשחמט על ידי מכונה, מערכות נהיגה אוטומומיות ועוד. בשנים האחרונים ישנו שינוי גדול בנושא זה.

מדובר על מודל למידה עצום, שבנוי מחלקים קטנים, כשבכל חלק נלמד על ידי רגרסיה לוגיסטיבית.

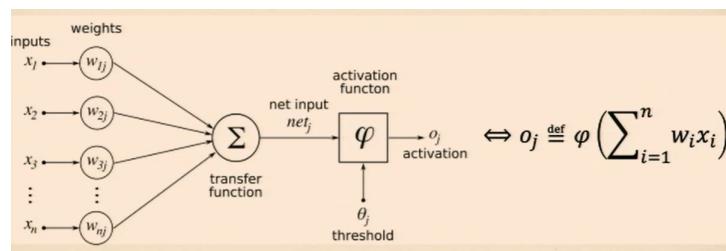
כיוון שלמידה عمוקה בניית מרשתות ניורוניים (neural networks) עצומות, נתחיל להתעסק בראשות קטנות.

8.1 רשתות ניורוניים

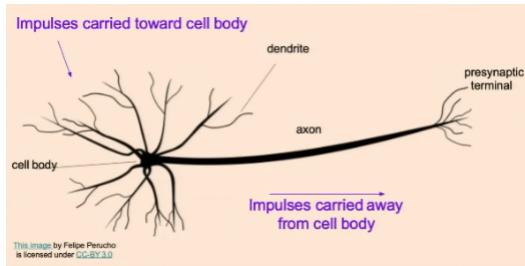
נזכיר ברגression לוגיסטיבית.
עקרונית, אנחנו מבצעים:

$$f(\mathbf{x}) = \phi(\langle \mathbf{w}, \mathbf{x} \rangle)$$

אך ניתן להסתכל על זה גם כך:



זה דומה מתמטית למה שמתתרחש בראש נוירונים בטבע:

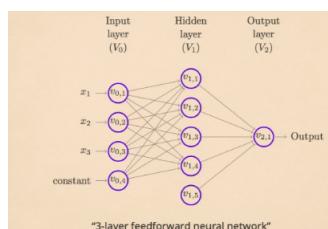


בדומה למה שקרה אצלנו בראש נוירונים אמיתיים שורדים רק הרכיבים החזקים, וכך גם אצלנו על הסכום להיות מספיק גדול.

איך זה עובד? כל נוירון מקבל קלט ופונקציות משקלות, ומתאים את המשקל לפחות מבחן חייזר:

$$\mathbf{x} \mapsto \phi \left(\langle \mathbf{x}, \mathbf{w}_t^{(1)} \rangle \right)$$

מדובר למעשה בפונקציה $\mathbb{R}^d \rightarrow \mathbb{R}^k$ - כאשר d הוא הממד של \mathbf{x} ו- k הוא מספר הנוירונים. נוכל לקח את הפלט שלנו \mathbb{R}^k וגם עליו לעשות את אותו התהליך:



על מנת להבחין בין פונקציות הביניים ובין פונקציית הסיום, לפונקציות באמצעות נקרא Activation ונסמנה עם σ .

הגדרה

רשת זרימה קדימה (Feedforward neural network) היא רשת עם שכבה חבויה אחת (שמכילה k נוירונים) ופונקציית אקטיבציה σ , היא מחלקת היפותזות מהצורה הבאה:

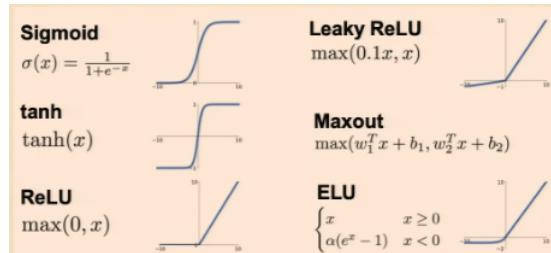
$$\mathcal{H} = \{ \phi \left(\langle \sigma(w_1^\top \mathbf{x}), \mathbf{w}_2 \rangle \right) \}$$

כאשר $\mathbb{R} \rightarrow \mathbb{R}$: σ היא פונקציית אקטיבציה, $-W_1$ היא וקטור המשקלות, ו- $\mathbf{w}_2 \in \mathbb{R}^k$ היא וקטור המשקלות. $\phi(x) = \frac{e^x}{1+e^x}$ עבור גרסיה מתקית כי $\phi(x) = x$

נשים לב כי השכבות החבויות לא מתעסקות בתוצאה הסופית של ϕ .

8.2 פונקציית אקטיבציה

ישנו אוסף של פונקציות אקטיבציה:



נדגיש כי השכבה最后一 מכילה נירון אחד. עבור בעיות רגרסיה גדולה יותר (multiple regression), עם k תוצאות, או בעיות סיווג גדולות יותר (multiclass classification) בשכבה הפלט יהיו k נירונים.

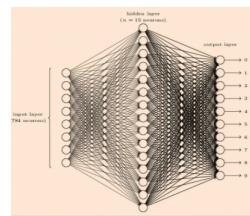
8.3 רגרסיה לוגיסטיבית ל-multiclass

כיצד נוכל לעשות רגרסיה לוגיסטיבית ל-multiclass? אם קיילנו k וקטורים, נחשב k מכפלות פנימיות, כאשר נקבל כי $\mathbf{x} \mapsto (\langle \mathbf{x}, \mathbf{w}_1 \rangle, \dots, \langle \mathbf{x}, \mathbf{w}_k \rangle)$ לבסוף, ההכללה של פונקציית logit תהיה:

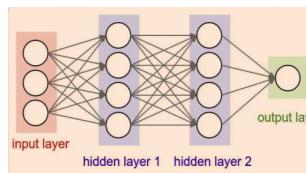
$$\frac{e^{\langle \mathbf{x}, \mathbf{w}_i \rangle}}{1 + \sum_{\alpha=1}^{k-1} e^{\langle \mathbf{x}, \mathbf{w}_{\alpha} \rangle}}$$

פונקציית הנראות המירבית עבור המחלקה ה- k יהיה $\frac{1}{1 + \sum_{\alpha=1}^{k-1} e^{\langle \mathbf{x}, \mathbf{w}_{\alpha} \rangle}}$

למעשה, נישם פונקציה זו בראשת הנוירונים ונתקבל את פונקציית ה-Soft-Max, ואומר מדויר בהכללה של logit: $-\log(\phi) : \mathbf{z} \mapsto (\log(\sum_{\alpha} e^{z_{\alpha}}) - z_1, \dots, \log(\sum_{\alpha} e^{z_{\alpha}}) - z_k)$ כך זה נראה בתמונה הגדולה:



ניתן להגדיל את מחלוקת ההיפותזה עם עוד שכבות חבויות:

**הגדרה**

רשת זרימה קדימה (Feedforward neural network) עם קלט $x \in \mathbb{R}^d$, T שכבות, פונקציה אקטיבציה σ ופלט ב- \mathbb{R}^k :

□ גראף מעגלי מכון $G = (V, E)$ כאשר כל שכבה היא באיחוד זר עם השכבות האחרות.

□ פונקציית משקל $E \rightarrow \mathbb{R}$: w מעל הצלעות.

□ פונקציה אקטיבציה $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ ופונקציית פלט $\phi : \mathbb{R}^{|V_T|} \rightarrow \mathbb{R}^{|V_T|}$ כאשר ϕ נקבע על ידי:

$$\mathbf{o}_t = \sigma(W_{t-1}^\top \mathbf{o}_{t-1}) \quad \mathbf{o}_m = \phi(\mathbf{o}_1, \dots, \mathbf{o}_T)$$

8.4 עשרות מחלוקת הhipotheses

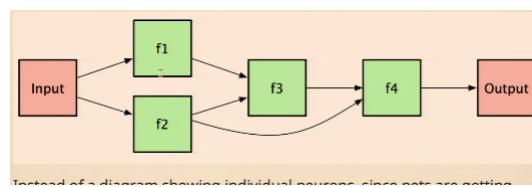
אם ניקח את $\phi(x) = \text{sign}(x)$, נרצה לבדוק את עשרות המחלוקת. במקרה בו אין אין שכבות נסתרות, מדובר פשוט במסוג חצי מרחב. אם נוסיף שכבה חבויה אחת, מדובר בחיתוך של $1 - k$ חצאי מרחבים. אם ניקח שכבה חבויה נוספת, כבר קיבל פונקציה גדולה הרבה יותר.

משמעות

תהי $\sigma(x) = \phi(x) = \text{sign}(x)$.

לכל d יש רשת נוירונים בוליאנית עם שכבה נסתרת אחת, כך שלכל פונקציה $\{ \pm 1 \}^d \rightarrow \{ \pm 1 \}^d$ קיימות השמות של משקלות w_1, w_2 שemmמשות את הפונקציה זו.

מספר הנוירונים הנדרש במצב זה הוא אקספוננציאלי במימד. למעשה, יש כאן overfit, אבל מסתבר שלא. אגב, אפשר להבחן שדרך הциור של רשת הנוירונים לא באמת מספקת מידע במקרה בו יש המון נוירונים, ולכן נוהגים להשתמש בדיאגרמת בלוקים, כאשר כל בלוק מכיל המון נוירונים:

**8.5 רשתות נוירונים عمוקות (deep neural net)**

בתחילת, הרעיון של רשתות נוירונים לא נטמע בקרב הקהילה המדעית, כיון שמדובר בעייה שאינה קמורה. אבל לאחר שהרעיון התפתח, התחילו לבנות רשתות נוירונות عمוקות, שמכילות המון שכבות חבויה.

השאלה המעניינת היא כיצד מאמנים רשתות גדולות כל כך.

8.5.1 אימון רשותות נוירוניים عمוקות

נבחון כי מחלקה הhipotезות היא כל הפונקציות מקבוצת הקשתות E , כאשר נתון לנו גוף $G = (V, E)$. דבר זה נכון, כיון שלכל קשת מאומנת יש משקל. אם כך, ניתן לראות כיצד מחלקה הhipotезות שולחה למיד אוקלידי $(\mathbb{R}^n, \|\cdot\|)$.

לכן, מדובר במודל שדומה לרגרסיה לוגיסטיות. בהינתן קבוצת אימון $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$ ופונקציית loss שמוגדרת מסתבר שאימונו על פי ERM הוא NP קשה, ולכן נאמן על פי עקרון הנראות המירבית.⁴⁶

על ידי, $\ell_{\mathbf{w}}$, נרצה לאמור את $\ell_{\mathbf{w}}(\mathbf{x}, y)$

למרות שראינו שעקרון הנראות המריבתי עבור רגסיה לוגיסטי היא קמור, עבור רשותות נוירונים עם שכבות חבויות הוא לא קמור. אולם, אחת ההפטעות הגדולות היא שלמרות זאת, אם נבצע GD עם התאמות, על פונקציות אלו, יתרחש דבר מעניין.

8.6 שימוש ב-GD בראשות נוירונים

קודם כל, חשוב להציג כי שימוש ב-GD⁴⁷ ברשותנו איננו טרויאלי - בעקבות גודל המדגם, נוצרך קבוצת אימוי ענקית והסכים על השגיאות מורכב מאוד חישובית.

מעבר לכך, חישוב הגרדינט של (S, w) L^7 הוא גזירה של כל אחד מהפרמטרים, שזו גם משימה מורכבת. לשם כך, נשתמש בשני טריקים:

נשתמש ב-SGD ולכון נctrar לחשב את הגרדיינט של ההפסד רק ביחס לדוגמה אחת.

קיטים אלגוריתם גומרי שמאפשר לנו לחשב את הגרדייאנט בנקודה, שנקרא back propagation (או prop back).

היעיון של back propagation הוא שbullet שכבת t יש $(W_{t-1} \bullet_{t-1}) = \sigma$.
 נגידיר את (u) Λ_t להיות **מינוס loglikelihood** - פונקציה שמקבלת את π ומייצאת פלט. נרצה לחשב את $(x)_0, \Lambda_0$,
 שהרי מתקיים כי $(\pi) = \Lambda_{t+1}$. נרצה למצוא את הנזירות החלקיים לפי w .
 נוכל לראות כי אנחנו רוצים לנזרות של הרכיבה, ולכן נוכל להשתמש בכלל השרשרת. בקרה זו פועל גם
 האלגוריתם:

```

SGD for Neural Networks

parameters:
number of iterations  $\tau$ 
step size sequence  $\eta_1, \eta_2, \dots, \eta_\tau$ 
regularization parameter  $\lambda > 0$ 

input:
layered graph  $(V, E)$ 
differentiable activation function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ 

initialize:
choose  $w^{(1)} \in \mathbb{R}^{|E|}$  at random
    (from a distribution s.t.  $w^{(1)}$  is close enough to 0)

for  $i = 1, 2, \dots, \tau$  do
    sample  $(x, y) \sim D$ 
    calculate gradient  $v_i = \text{backpropagation}(x, y, w, (V, E), \sigma)$ 
    update  $w^{(i+1)} = w^{(i)} - \eta_i(v_i + \lambda w^{(i)})$ 

output:
     $w$  is the best performing  $w^{(i)}$  on a validation set

```

46 הוא מכסה גם את הסיווג וגם את הרגרסיה.
47 נזכיר את הנוסחה:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_{t+1} \nabla L \left(\mathbf{w}^{(t)}; s \right),$$

הוקטור $\omega \in \mathbb{R}^{|E|}$ מייצג כאן את וקטור המשקלות.

ניתן להשתמש גם ב-mini batch - כלומר לדגום על כמה דוגמאות ולא אחת, אז לסקום.

8.6.1 טריקים בשימוש ב-SGD

כפי שאמרנו, כיוון שהבעייה לא קמורה, אז אי אפשר להשתמש בשיטות שראינו באופן ישיר ויש צורך להשתמש בכל מיני סוגים של טריקים.

שימוש במומנטום

כיוון ש-SGD עלול להיות איטי בחישוב המינימום, אנו יכולים להשתמש ב-SGD עם נסטרוב-מוונטום. למעשה, בכל פעם אנחנו לוקחים וקטור מסוים ומעדכנים אותו על פי המהירות:

$$\begin{aligned}\mathbf{v}^{(t+1)} &= \alpha \mathbf{v}^{(t)} - \eta_t \sum_{(\mathbf{x}, y) \in B_t} \nabla L(\theta^{(t)} + \alpha \mathbf{v}^{(t)}; B_t) \\ \theta^{(t+1)} &= \theta^{(t)} + \mathbf{v}^{(t)}\end{aligned}$$

המהירות מתקדמת בכל פעם, וכך גם תאוצה. כלומר, אם הلنנו לכיוון מסוים אנחנו יכול להאיץ ולהתקדם לכיוון זה.

אתחול המשקלות

עד כה לא התעסקנו בשאלת "היכןначילה" להתעסק עם המידע, כיוון שהפונקציות היו קמורות. במקרה של ממדים גבוהים זו שאלה ממשונית. ניתן את��זרה *iid*, או בכך שבכל פעם יש כמה נוירונים שונים מאפס.

קצב למידה אדפטיבי (adaptive learning rate)

כשדיברנו על GD, השתמשנו בגודל הצעד (step size), ולמעשה זהו המושג כאן.

חשוב להתחאים מקומית את גודל הצעד בו נדרש ללכט.

ישנן שלוש שיטות מרכזיות, AdaGrad, RMSProp, ADAM לחישוב גודל הצעד בצורה אדפטיבית.

רגוריזציה

במוקם למינר את פונקציית ההפסד, נוכל למינר את $\|\mathbf{w}\|_2^2$.
לכארה, זה אומר להקטין את השונות.

פרקטיקה

שם המשחק הוא למשה חישוב: ההבדל בין רשות טוביה לרשות פחות טוביה יכול להתבטא בשימוש בטריקים שונים מלפני זה.

בחירה ה-Hyperparameter

נבחן כי יש לנו מספר פרמטרים שנctrarck לבחור. מהו גודל ה-batch שנשתמש בו, מהו המומנטום σ ומהו פרמטר הרגוליזציה λ .