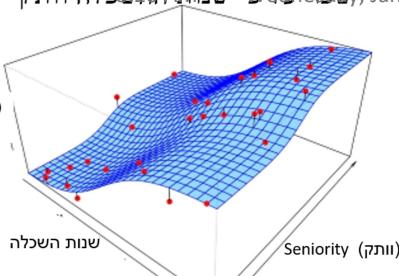


Bias vs. Variance Tradeoff

השגיאה שיש למודל על נתונים המבחן מורכבת מ 3 סוגים של שגיאה:

- על ה $\text{err}_{\text{irreducible}}$ אין שליטה, שאר השגיאות תלויות באlgorigitms הלמידה אך יש קשר ביניהם:
 - ככל שהמודל קשייך יותר (איננו גמיש) הוא עלול לא לקרב באופן הדוק את f : **Underfitting**
 - יציר שגיאת Bias
 - ככל שהמודל גמיש יותר, הוא עלול להתאים את עצמו לתבניות לא משמעותיות בנתוני האימון: **Overfitting**
 - יציר שגיאת Variance

Bias vs. Variance + MSE
שבר על ידי שנור אוניברסיטאות
Wednesday, January 20, 2021



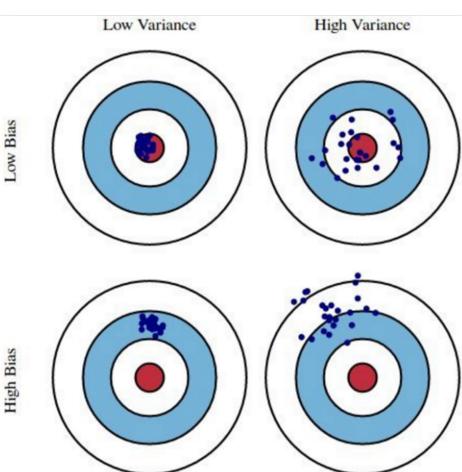
גרסתה ב 2 מימדים באמצעות **Thin Plate Splines** (ש ביכולתו להתחים את עצמו לדוגמאות שבאימון).
אך גם פה המודל שעשו שגיאות.
ישנים שגיאות שמקורם במודל לא מספיק גמיש או במדגם אימון קטן מדי.
יש גם שגיאות אינהרנטיות
מהו?

הקליטים הנתונים אינם כל הסיבות לשכר
(ישו ווש, טוויות מדדי, ישם סיבות אחרות שלא מדוין או אי אפשר למדוד אותן)

Over Fitting [High Variance] vs Under fitting [High Bias]

הגדלת הביאז

Bias up

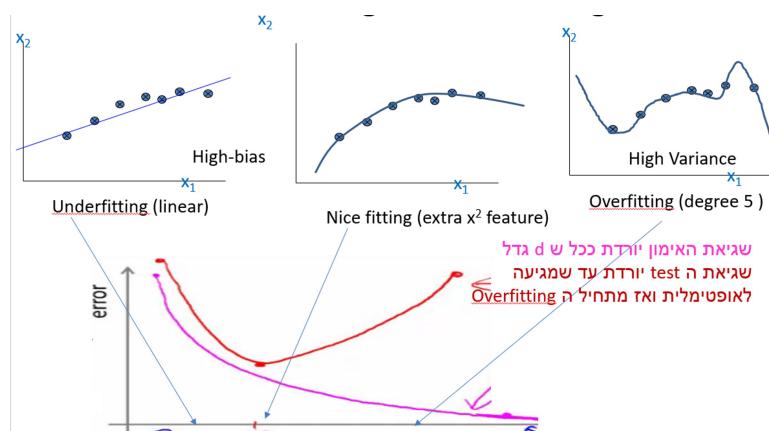


Flexibility

big

הגדלת השונות

Variance up



High Variance מרחיב ההיפותזה
גדול מידי ואין לנו מספיק data
שתספק מספיק אילוצים כדי לבחון
את ההיפותזה הטובה ביותר.
מקבלים שגיאה שנובעת מהתאמת
יתר לקבוצת אימון מסוימת

High-Bias שגיאה הנובעת
מההיפותזה פשוטה מדי
(למשל: לינארית). חוסר
גמישות.

התוצאות יהיו דומות אחת לשניה,
אבל בMMddצען הן לא יהיו מדויקות

אלגוריתמים למציאת האיזון הטוב ביותר בין Bias ל Variance:

1. Regularization
2. Boosting
3. Bagging

Bias vs. Variance Tradeoff

השגיאה שיש למודל על נתונים המבחן מורכבת מ 3 סוגים של שגיאות:

$$\text{loss} = (\text{Bias})^2 + \text{Variance} + \text{irreducible}$$

- על ה err err אין שליטה, שאר השגיאות תלויות באלאgorיתם הלמידה אך יש קשר ביניהם:

– ככל שהמודל קשייך יותר (איננו גמיש) הוא עשוי לא לקרב באופן Underfitting

הduk את f :

יציר שגיאת Bias

– ככל שהמודל גמיש יותר, הוא עשוי להתאים את עצמו לתבניות לא שימושיות בנתוני האימון: Overfitting

ויציר שגיאת Variance

ה MSE - תוחלת ריבועי שגיאת החיזוי – ניתנת לפירוק

D קבוצת אימון: משתנה מיקרי מטריציוני - אוסף דוגמאות

x הקלט: משתנה מיקרי וקטורי שאינו תלוי ב D מתוך התפלגות לא ידועה

f פונקציה דטרמיניסטית אותה רצים לשערך בעזרת h (איננה משתנה מיקרי)

t_x = f(x) + ε ערך המטריה: משתנה מיקרי התלוי ב x וכל שגיאה בלתה נמנעת ε

ε רעש: משתנה מיקרי שאין לו שליטה בו. בעל תוחלת 0 ואי תלות בין x, ε

y = h_d(x) תוצאה החיזוי: משתנה מיקרי התלוי במדגם D ובדוגמה x

$$f(x) \approx h_d(x)$$

ה MSE התיאורתי הוא התוחלת של ריבועי השגיאה:

$$MSE = E_x, E_D [(h_D(x) - t_x)^2]$$

על כל הדגימות D ובכל הקלטים x

$$MSE = E \left[h_d(x) - f(x) \right]^2 + \text{var}(\epsilon)$$

(ε) - שגיאה בלתה ניתנת להפחיתה. השונות של הרעש

h(x) - ε שגיאה נתנת להפחיתה. תוחלת ריבועי ההפרשים בין f(x)

$$MSE = E_x E_D [(t_x - h_D(x))^2] = E[(f(x) + \epsilon - h_D(x))^2]$$

$$= E[(f(x) - h_D(x))^2 + 2(f(x) - h_D(x))\epsilon + (\epsilon - 0)^2]$$

$$= E[(f(x) - h_D(x))^2] + 2E[f(x) - h_D(x)]E[\epsilon] + E[(\epsilon - 0)^2]$$

תוחלת מכפלה = למכפלה התוחלות
אי תלות השגיאה ב x:

תוחלת ε היא 0

הגדרת [ε]

הוכחה לפירוק ה MSE

$$\text{MSE}(y, t) = E_x E_D [(h_D(x) - f(x))^2] + \text{var}(\varepsilon)$$

= ReducableError + var(ε)

ה. פונקציות דטרמיניסטיות

נגיד: $[x] \overline{h_D(x)} = E[h_D(x)]$ פונקציה תוחלת החיצויים

$$\text{ReducableError} = E[(h_D(x) - f(x))^2] = E[(h_D(x) - \overline{h_D(x)} + \overline{h_D(x)} - f(x))^2]$$

$$= E[(h_D(x) - \overline{h_D(x)})^2 + 2(h_D(x) - \overline{h_D(x)})(\overline{h_D(x)} - f(x)) + (\overline{h_D(x)} - f(x))^2]$$

תוחלת של מכפלת מ"מ בפונקציה דטרמיניסטיבית היא
מכפלת התוחלת בפונקציה

$$= E[(h_D(x) - \overline{h_D(x)})^2] + 2E[h_D(x) - \overline{h_D(x)}] E[\overline{h_D(x)} - f(x)] + E[(\overline{h_D(x)} - f(x))^2]$$

תוחלת הפרשיים של
מ"מ מהממוצע הוא 0

$$\begin{aligned} & E[\overline{h_D}^2] + E[f^2] - 2E[\overline{h_D}f] \\ & = (E[h_D])^2 + (E[f])^2 - 2E[h_D] E[f] \\ & = (E[h_D])^2 + (E[f])^2 - 2E[h_D] E[f] \\ & \quad (h_D \text{ is a scalar}, E[h_D] = E[h_D]) \\ & = (E[h_D] - E[f])^2 = (E[h_D - f])^2 \end{aligned}$$

$$= E_x E_D [(h_D(x) - \overline{h_D(x)})^2] + (E_x E_D [h_D(x) - f(x)])^2$$

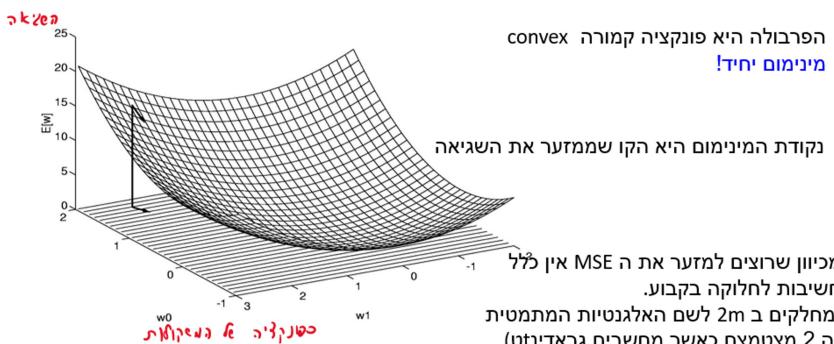
= Variance[h_D(x)] + (Bias[h,f])^2

פונקציית MSE ברגסיה לינארית של משתנה אחד היא פרבולה דו-מימדית (קערה)

D – set of given points

t, x – scalars
 w_i – variables

$$\text{MSE}_{D,h}(w_0, w_1) = \frac{1}{2m} \sum_{i \in D} (t_i - (w_1 x_i + w_0))^2$$



הפרבולה היא פונקציה קמורה convex
מינימום ייחודי!

נקודות המינימום היא הנקו שמשמער את השגיאה
מחייב שורצים למצער את MSE אין כל
חשיבות לחלקה בקבוע.
מחלקים ב 2 ומשם האלגוריות המתמטית
ה 2 מצטמצם כאשר מחשבים גראדיאנט.

דוגמה לשיטת הפיתרון הנורמלית לרוגסיה

LINARITY

Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)	
x_1	x_2	x_3	x_4	y	
2104	5	1	45	460	
1416	3	2	40	232	
1534	3	2	30	315	
852	2	1	36	178	

M=4, n=4

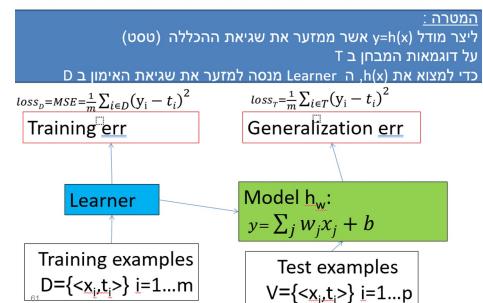
$$\left(\begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 2104 & 1416 & 1534 & 852 \\ 5 & 3 & 3 & 2 & 40 \\ 1 & 2 & 2 & 1 & 36 \\ 45 & 40 & 30 & 36 \end{bmatrix} \times \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 2104 & 1 & 1 & 1 \\ 2104 & 1416 & 1534 & 852 \\ 5 & 3 & 3 & 2 \\ 1 & 2 & 2 & 1 \\ 45 & 40 & 30 & 36 \end{bmatrix} \times \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

Use: pseudo-Inverse matrix operation

pinv(X^T * X) * X^T * t

Pg.3

גרסיה לינארית רבת משתנים



כיצד מעריכים MSE(w0, w1, ..., wn)

עבור רוגסיה לינארית?

השווות הנגזרות החלקיים ל 0:

מיצרת מערכת (n+1) משוואות

לינאריות:

לכן קיימת נוסחה סגורה באלגברה
לינארית שמצוצת את נקודת
המינימום (באמצעות הפיכת מטריצה
גדולה)

אך, עבור מממד גבוה או כאשר
קובוצת האימון גדולה מאוד, לא מעשי
להשתמש בשיטה זו.

MSE Reduction

Normal Equation:

יש דוגמה גם בבחני הבית

פונקציית עלות $loss = \frac{1}{2} MSE$
היפותה לינארית: $y = wx + b$
רוצים למצוא (w) שימזע את MSE

אלגוריתם GD

התחל מוקטור משקלות אקראיים
בצע Epoch שוב ושוב (עד שמתיקמים תנאי עזרה)

- חשב הגרדיינטים לכל הדוגמאות p ב D
- לכל דוגמא ולכל משקל i , מחשבים:

$$\frac{\partial loss(w, p)}{\partial w_i} = (y_p - t_p)x_{pi}$$

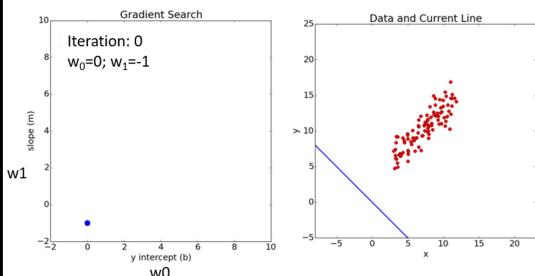
$$\Delta w_i = -\frac{\lambda}{m} \sum_{p \in D} \frac{\partial loss(w, p)}{\partial w_i} = \frac{\lambda}{m} \sum_{p \in D} (t_p - y_p)x_{pi}$$

$$w_i = w_i + \Delta w_i$$

$$w = w + \lambda \cdot mean((T - Y)X)$$

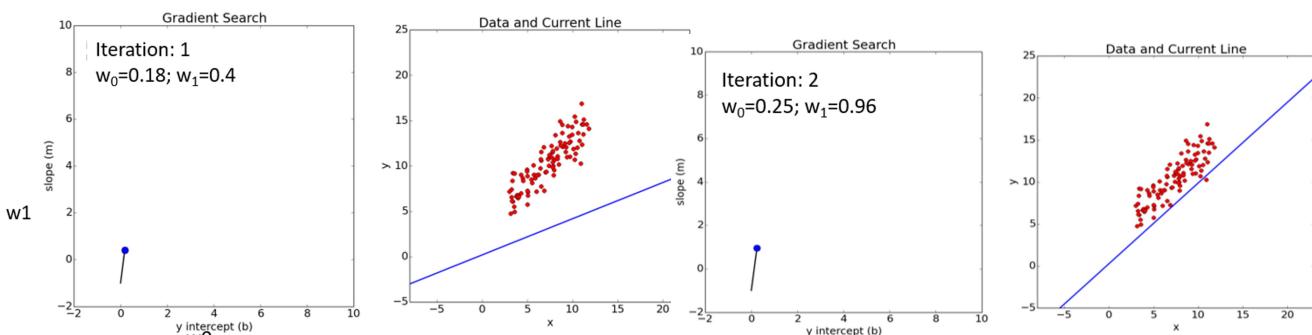
בפועלה יקטרנית אחת:

מתחילה מישר רנדומלי $w_0=0; w_1=-1$

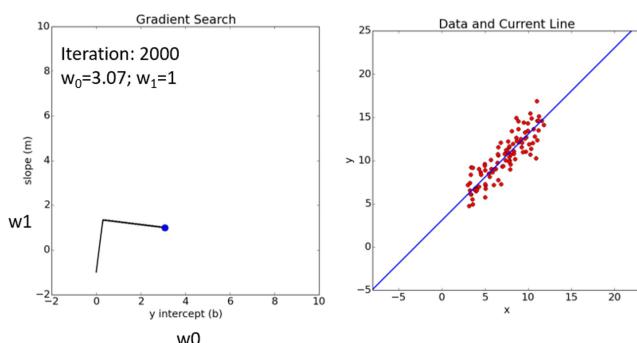


ישנן טכניקות המקטינות בהדרגה את קצב הלמידה כדי להתקרב יותר לנקודת המינימום

מחשבים צעד GD: ממוצע גראדיינט של כל הדוגמאות
ומתכנים את הקו ($\Delta w = (0.18, 1.4)$
מקבלים $w_0 = 0.18$; $w_1 = 0.4$)



באיטרציה השלישית שוב מבצעים צעד GD ומתקנים



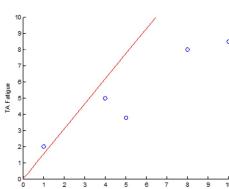
Gradient Descent- Single var with no bias

מבצע צעד GD: $\Delta w_1 = -\lambda \frac{\partial loss}{\partial w_1}$

$$\frac{\partial loss}{\partial w_1} = \frac{1}{m} \sum_{p=1}^m (y_p - t_p)x_p$$

ערך התחלתי 2

x_p	t_p	y_p	$(y_p - t_p)x_p$
10	8.5	$2 * 10 = 20$	$(20 - 8.5) * 10 = 125$
5	3.8	$2 * 5 = 10$	$(10 - 3.8) * 5 = 31$
1	2	$2 * 1 = 2$	$(2 - 2) * 1 = 0$
8	8	$2 * 8 = 16$	$(16 - 8) * 8 = 64$
4	5	$2 * 4 = 8$	$(8 - 5) * 4 = 12$

$$\frac{1}{m} \sum_{p=1}^m (y_p - t_p)x_p : 46.4$$


$$\Delta w_1 = -\lambda \frac{\partial loss}{\partial w_1}$$

מבצע צעד NOI:

$$\frac{\partial loss}{\partial w_1} = \frac{1}{m} \sum_{i=1}^m (y^i - t^i)x^i$$

עתה $w_1 = 1.54$

$$\Delta w_1 = -\lambda \frac{\partial loss}{\partial w_1} = -0.01 * 46.4 = -0.464$$

$$w_1^{new} = w_1 + \Delta w_1 = 2 - 0.464 \cong 1.54$$

x_p	t_p	y_p	$(y_p - t_p)x_p$
10	8.5	$1.54 * 10 = 15.4$	$(15.4 - 8.5) * 10 = 69$
5	3.8	$1.54 * 5 = 7.7$	$(7.7 - 3.8) * 5 = 19.5$
1	2	$1.54 * 1 = 1.54$	$(1.54 - 2) * 1 = -0.46$
8	8	$1.54 * 8 = 12.32$	$(12.32 - 8) * 8 = 34.56$
4	5	$1.54 * 4 = 6.16$	$(6.16 - 5) * 4 = 4.64$

$$\frac{1}{m} \sum_{p=1}^m (y_p - t_p)x_p = 25.45$$
