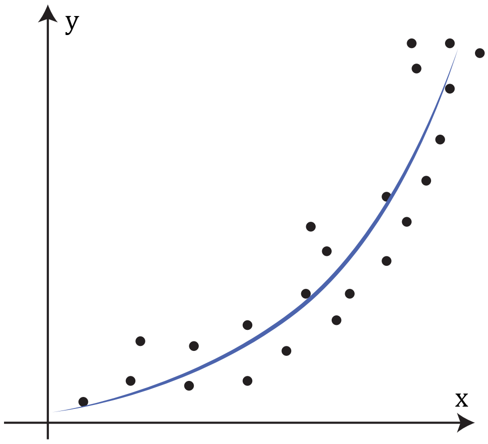
מבוא ללמידה חישובית (20942) | סיכום הרצאה 4

מנחה: שי מימון
סמסטר: 2022א'
נכתב על ידי: מתן כהן

1 רגרסיה לינארית - המשך

מה נעשה במצב שבו הנקודות לא מתיישרות סביב ישר?



נוכל לבצע טרנספורמציה מהסוג הבא:

$$x \to \{1, x, x^2, x^3, .., x^d\}$$

ובעזרתה נוכל לשבור את ההגבלה לקווים לינארים ולקבל בעצם פולינום מדרגה d:

$$\mathbf{w}^T\mathbf{x} = w_0 + w_1 x + ... + w_d x^d$$

ועדיין לא משתנה העובדה שביחס ל $\mathbf{w}$ אנחנו לינאריים (הרי הנחנו שהחזאי שלנו הוא פונקציה לינארית של **הפרמטרים** ולא של המדידות) ובעצם כפי שעשינו במודל הפרספטרון גם כאן ניתן להרחיב את הגישה לכזו שתתמודד עם מצבים בעייתיים ולא לינאריים.

במקרה הזה המטריצה X תיראה כך:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^d \end{pmatrix}$$

וגם כאן נקבל:

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

עם זאת יש להזהר מאוד שכן הדבר יכול לגרום להתאמת יתר ולפולינום שיעבור בכל הנקודות ויניב לנו $E_{in} = 0$.

2 רגרסיה לוגיסטית - Logistic Regression

רגרסיה לוגיסטית גם היא מדברת על **מאורעות בינאריים** כמו במקרה של הפרספטרון אך השוני העיקרי הוא שאלגוריתם זה רוצה להביע את החיזוי שלו בעזרת **הסתברות** - במילים אחרות החזאי שלנו יניב הסתברות בין 0 ל-1 ויאמר מה הסיכוי שהדוגמה שלנו מסווגת כ-"1".

בהקבלה לסימונים שכבר הזכרנו בקורס:

- $\hat{y} \in [0,1]$
- הסיגנל שלנו לינארי ונשאר כמו תמיד $\mathbf{w}^T\mathbf{x}$
- על הסיגנל מלבישים פונקציה לא לינארית θ
- סט ההיפותזות שלנו הוא $h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$
  - במילים אחרות זו θ אשר מורכבת על פונקציה אפינית (שוב, בגלל ה bias שיש לנו)

2.1 הפונקציה θ - sigmoid

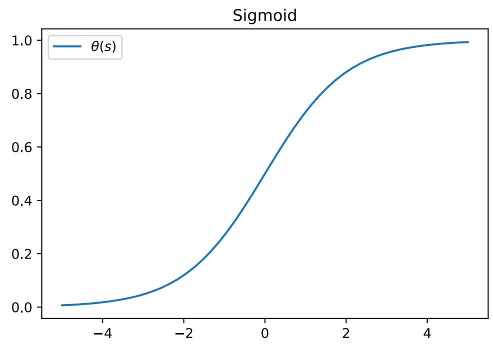
הפונקציה θ נקראת ב-2 שמות אפשריים:

(1) logistic function

(2) sigmoid

ומוגדרת:

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$$



כאשר

$$\theta(s) \xrightarrow[s\to\infty]{} 1$$

$$\theta(s) \xrightarrow[s\to-\infty]{} 0$$

ו-

$$\theta(0) = \frac{1}{2}$$

2.1.1 תכונות חשובות של הפונקציה θ

(1) $1-\theta(s)=\theta(-s)$:

$$1-\theta(s)=1-\frac{e^s}{1+e^s}=\frac{1+e^s-e^s}{1+e^s}=\frac{1}{1+e^s}=\theta(-s)$$

(2) $\theta'(s)=\theta(s)\cdot(1-\theta(s))$:

$$\begin{aligned}\theta'(s)&=\frac{d}{ds}\left(\frac{1}{1+e^{-s}}\right)=\frac{d}{ds}(1+e^{-s})^{-1}\\&=-(1+e^{-s})^{-2}\cdot(-e^{-s})\\&=\frac{e^{-s}}{1+e^{-s}}\cdot\frac{1}{1+e^{-s}}\\&=\frac{1}{1+e^{-s}}\cdot\frac{1}{1+e^{-s}}\\&=\theta(s)\cdot(1-\theta(s))\end{aligned}$$

2.1.2 פונקציית שגיאה מתאימה לסיגמויד - cross-entropy error measure

כיוון שבידינו ההיפותזה $h(\mathbf{x})=\theta(\mathbf{w}^T\mathbf{x})$ ומכך שהדוגמאות מתויגות בצורה בינארית ואנחנו מנסים למצוא **הסתברות** $P(y=1\mid\mathbf{x})$ צריך למצוא מטריקה מתאימה לבדיקת השגיאה שלנו.

לשם כך נשתמש בפונקציית שגיאה בשם cross-entropy error measure

פונקציית השגיאה הנ"ל מוצאת מרחק בין 2 פונקציות הסתברות של משתנים בינאריים $(p,1-p),\ \ (q,1-q)$.

כעת נגדיר את פונקציית השגיאה בצורה פורמלית (I הוא אינדיקטור):

$$E_{in}(\hat{\mathbf{w}})=\frac{1}{N}\sum_{n=1}^{N}I(y_n=+1)\cdot\log\left(\frac{1}{h(\mathbf{x}_n)}\right)+I(y_n=-1)\cdot\log\left(\frac{1}{1-h(\mathbf{x}_n)}\right)$$

הסבר:

נזכור כי המטרה שלנו היא למדל את ההסתברות $P(y=1\mid\mathbf{x})$ בתור $h(\mathbf{x})$.

- במקרים בהם $y_n=+1$:
  - הביטוי: $I(y_n=+1)\cdot\log\left(\frac{1}{h(\mathbf{x}_n)}\right)$ תקף רק אם $y_n=+1$ והמחיר שנשלם הוא $\log\left(\frac{1}{h(\mathbf{x}_n)}\right)$
  - $\Leftarrow$
    - ⋆ אם $h(\mathbf{x}_n)\to 1$ הרי שנקבל שגיאה מאוד קטנה ששואפת ל-0
    - ⋆ אם $h(\mathbf{x}_n)\to\infty$ הרי שנקבל שגיאה מאוד גדולה שתשאף ל-∞
- במקרים בהם $y_n=-1$:
  - הביטוי: $I(y_n=-1)\cdot\log\left(\frac{1}{1-h(\mathbf{x}_n)}\right)$ תקף רק אם $y_n=+1$ והמחיר שנשלם הוא $\log\left(\frac{1}{1-h(\mathbf{x}_n)}\right)$
  - $\Leftarrow$
    - ⋆ אם $1-h(\mathbf{x}_n)\to 1$ הרי שנקבל שגיאה מאוד קטנה ששואפת ל-0
    - ⋆ אם $1-h(\mathbf{x}_n)\to\infty$ הרי שנקבל שגיאה מאוד גדולה שתשאף ל-∞

כעת מכך ש- $h(\mathbf{x}_n)=\theta(\hat{\mathbf{w}}^T\mathbf{x}_n)$ ומהתכונות שהוכחנו בסעיף 2.1.1 נסיק:

$$E_{in}(\hat{\mathbf{w}})=\frac{1}{N}\sum_{n=1}^{N}I(y_n=+1)\cdot\log\left(\frac{1}{\theta(\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)+I(y_n=-1)\cdot\log\left(\frac{1}{\theta(-\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)$$

2.1.3 נגזרת הסיגמויד

<u>תחילה נסדר את הפונקציה בצורה נוחה:</u>

מכך שהביטוי $I(y_n=+1)\cdot\log\left(\frac{1}{\theta(\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)$ תקף רק עבור $y_n=+1$ ובאופן דומה $I(y_n=-1)\cdot\log\left(\frac{1}{\theta(-\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)$ תקף עבור $y_n=-1$

נוכל להסיק:

$$E_{in}(\hat{\mathbf{w}})=\frac{1}{N}\sum_{n=1}^{N}I(y_n=+1)\cdot\log\underbrace{\left(\frac{1}{\theta(y_n\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)}_{\#1}+I(y_n=-1)\cdot\log\underbrace{\left(\frac{1}{\theta(y_n\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)}_{\#2}$$

ונבחין כי הביטויים $\#1$ ו-$\#2$ שווים, נוכל להוציא מחוץ לסוגריים:

$$\begin{aligned}E_{in}(\hat{\mathbf{w}})&=\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{\theta(y_n\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)\cdot\underbrace{[I(y_n=+1)\cdot+I(y_n=-1)]}_{=1}\\&=\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{\theta(y_n\hat{\mathbf{w}}^T\mathbf{x}_n)}\right)\\&=\boxed{\frac{1}{N}\sum_{n=1}^{N}\log\left(1+e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\right)}\end{aligned}$$

נבחין כעת כי השגיאה שואפת ל-0 כאשר $y_n\hat{\mathbf{w}}\mathbf{x}_n$ שואף ל-∞

הערה: הביטוי וההתנהגות שלו מזכירים את מה שלמדנו בפרספטרון

<u>כעת נעבור לגזירה</u>

$$\begin{aligned}\frac{\partial}{\partial\hat{\mathbf{w}}}E_{in}(\hat{\mathbf{w}})&=\frac{\partial}{\partial\hat{\mathbf{w}}}\frac{1}{N}\sum_{n=1}^{N}\log\left(1+e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\right)\\&=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\left(1+e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\right)}\cdot e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\cdot(-y_n\mathbf{x}_n)\end{aligned}$$

ולכן:

$$\begin{aligned}\nabla_{\hat{\mathbf{w}}}E_{in}(\hat{\mathbf{w}})&=-\frac{1}{N}\sum_{n=1}^{N}\frac{e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}}{\left(1+e^{-y_n\hat{\mathbf{w}}^T\mathbf{x}_n}\right)}\cdot(y_n\mathbf{x}_n)\\&=\boxed{-\frac{1}{N}\sum_{n=1}^{N}\theta\left(-y_n\hat{\mathbf{w}}^T\mathbf{x}_n\right)\cdot(y_n\mathbf{x}_n)}\end{aligned}$$

כעת נרצה למצוא מינימום.

3 מציאת מינימום בשיטה איטרטיבית - Gradient Descent

הערה: שימוש ב-Gradient Descent נוח מאוד כאשר יש לנו פונקציות קונבקסיות להן יש מינימום גלובלי נתון ובפרט עבור פונקציות שהן קונבקסיות ממש להן יש רק מינימום גלובלי נתון יחיד.

נתבונן בפונקציית השגיאה בזמן t ובהפרש:

$$\Delta E_{in} = E_{in}(\hat{\mathbf{w}}(t+1)) - E_{in}(\hat{\mathbf{w}}(t))$$

ונגדיר:

$$\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \eta_t \cdot \hat{\mathbf{v}}_t$$

כאשר:

- η_t הוא איזשהו גודל צעד קטן שמבצעים בכל איטרציה
- $\hat{\mathbf{v}}$ הוא וקטור יחידה בכיוון מסוים ($\|\hat{\mathbf{v}}\| = 1$)

וכעת נשתמש בטור טיילור לפיתוח ובכיוון מינוס הגרדיאנט על מנת לקבל ירידה תלולה ביותר:

$$\begin{aligned}
\Delta E_{in} &= E_{in}\left(\hat{\mathbf{w}}(t) + \eta_t \cdot \hat{\mathbf{v}}_t\right) - E_{in}\hat{\mathbf{w}}(t) \\
\backslash taylor &\approx E_{in}(\hat{\mathbf{w}}(t)) + \eta_t \cdot \nabla^T E_{in}(\hat{\mathbf{w}}(t))\hat{\mathbf{v}}_t - E_{in}\hat{\mathbf{w}}(t) \\
&= \eta_t \cdot \nabla^T E_{in}(\hat{\mathbf{w}}(t))\hat{\mathbf{v}}_t \\
\backslash \hat{\mathbf{v}}_t = \frac{-\nabla E_{in}(\hat{\mathbf{w}}(t))}{\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|} &\geq \eta_t \cdot \nabla^T E_{in}(\hat{\mathbf{w}}(t)) \cdot \frac{-\nabla E_{in}(\hat{\mathbf{w}}(t))}{\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|} \\
&= -\eta_t \cdot \|\nabla E_{in}(\hat{\mathbf{w}}(t))\|
\end{aligned}$$

3.1 מה גודל הצעד הטוב ביותר?

צריך להבחין בעובדות:

- צעדים קטנים מידי יגמרו להתכנסות מאוד איטית
- צעדים גדולים מידי יכולים לגרום לאי-התכנסות ואף להתרחקות מנקודת המינימום
- שילוב גדלים שונים הוא הטוב ביותר - נתחיל מצעדים גדולים ונקטין
  - היריסטיקה מתאימה לבחירת גודל הצעד היא בחירת η_t שיהיה פרופורציונלי לנורמה של הגרדיאנט
    - ⋆ ההירסטיקה עושה שכל כיוון שהנורמה של הגדריאנט כשאנו רחוקים מנקודת המינימום היא גדולה וכשמתקרבים היא הופכת קטנה

לכן נוכל לבחור:

$$\eta_t = \eta \cdot \|\nabla E_{in}(\hat{\mathbf{w}}(t))\|$$

והדבר גם יצמצם את המכנה שלנו ממקודם:

$$\begin{aligned}
\hat{\mathbf{w}}(t+1) &= \hat{\mathbf{w}}(t) + \eta \cdot \|\nabla E_{in}(\hat{\mathbf{w}}(t))\| \cdot \hat{\mathbf{v}}_t \\
\backslash \hat{\mathbf{v}}_t = \frac{-\nabla E_{in}(\hat{\mathbf{w}}(t))}{\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|} &= \hat{\mathbf{w}}(t) + \eta \cdot \|\nabla E_{in}(\hat{\mathbf{w}}(t))\| \cdot \frac{-\nabla E_{in}(\hat{\mathbf{w}}(t))}{\|\nabla E_{in}(\hat{\mathbf{w}}(t))\|} \\
&= \hat{\mathbf{w}}(t) - \eta \cdot \nabla E_{in}(\hat{\mathbf{w}}(t))
\end{aligned}$$

הערה: נהוג לבחור $\eta = 0.1$

3.2 אתחול וסיום האלגוריתם

- התחלה
  - בחלק מהמקרים ניתן להתחיל את האלגוריתם עם $\hat{\mathbf{w}}(0) = 0$
    - ⋆ בחלק מהמקרים זה יכול לגרום לתקיעה של האלגוריתם
    - ⋆ ברגרסיה לוגיסטית אין בעיה להתחיל עם הערך הנ"ל
    - ⋆ במקרים אחרים ניתן להתחים עם אתחול רנדומלי:
      - · מגרילים משתנה מקרי גאוס עם תוחלת 0 ושונות מאוד קטנה לכל ערך בוקטור
- סיום
  - דרך אחת היא לקבוע מספר איטרציות מקסימלי
    - ⋆ לפעמים על פי דרך זו אין שליטה על איכות התוצאה המתקבלת
  - דרך נוספת בנוסף לקביעת מספר האיטרציות המקסימלי היא הוספת threshold על הנורמה של הגרדיאנט
    - ⋆ זאת כיוון שאנחנו רוצים שהנורמה של הגרדיאנט תהיה קטנה מאיזשהו ערך מסוים

$$\|\nabla E_{in}(\hat{\mathbf{w}}(t))\| < \varepsilon$$

    - ⋆ לכן עם קביעת ערך מאוד מאוד קטן שחצייתו מסיימת את התהליך נוכל לתרום להתכנסות