Short HW5 – Logistic regression and Deep learning

.1

1.1. עבור פונקציית הסיגמואיד המוגדרת כך:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

נקבל:

$$1-\sigma(z) = 1-\frac{1}{1+e^{-z}} = \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} \cdot \frac{e^{z}}{e^{z}} = \frac{1}{e^{z}\left(1+e^{-z}\right)} = \frac{1}{1+e^{z}} = \sigma(-z)$$

1.2. נראה את הנדרש:

$$\sigma'(z) = \frac{0\cdot\left(1+e^{-z}\right)-1\cdot\left(-e^{-z}\right)}{\left(1+e^{-z}\right)^2} = \frac{e^{-z}}{\left(1+e^{-z}\right)^2} = \frac{1}{1+e^{-z}}\cdot\frac{e^{-z}}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-z}}\cdot\frac{\left(1+e^{-z}\right)-1}{1+e^{-z}} = \frac{1}{1+e^{-z}}\cdot\left(\frac{1+e^{-z}}{1+e^{-z}}-\frac{1}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}}\cdot\left(1-\frac{1}{1+e^{-z}}\right)$$

$$= \sigma(z)\cdot(1-\sigma(z))$$

1.3. נפתח את הנגזרת הנדרשת, עבור הפונקציות הבאות:

$$\hat{p} = \sigma\left(w^T x\right) = \sigma(z),\ z(w) = w^T x$$

מתקיים:

$$\nabla_w \hat{p} = \nabla_w \sigma(z(w)) = \sigma'(z(w))\cdot\nabla_w z(w)$$

ידוע כי $\nabla_w\left(w^T x + b\right) = x$ ולכן:

$$\sigma'(z(w))\cdot\nabla_w z(w) = \sigma(z)\cdot(1-\sigma(z))\cdot x$$

$$\Longrightarrow \nabla_w \hat{p} = \sigma(z)\cdot(1-\sigma(z))\cdot x$$

1.4. נוכיח כי מתקיים הנדרש:

$$\nabla_w \ell^{CE}(y,\hat{p}) \underbrace{=}_{\substack{\text{לפי הגדרת}\\ \ell^{CE}}} \nabla_w(-y\cdot ln(\hat{p})-(1-y)\cdot ln(1-\hat{p}))$$

$$= -y\cdot\nabla_w(ln(\hat{p}))-(1-y)\cdot\nabla_w(ln(1-\hat{p}))$$

$$= -y\cdot\nabla_w\cdot\frac{1}{\hat{p}}-(1-y)\cdot(-\nabla_w\hat{p})\cdot\frac{1}{1-p}$$

$$= -y\cdot\hat{p}\cdot(1-\hat{p})\cdot x\cdot\frac{1}{\hat{p}}+(1-y)\cdot(\hat{p}\cdot(1-\hat{p})\cdot x)\cdot\frac{1}{1-p}$$

$$= -y\cdot(1-\hat{p})\cdot x+(1-y)\cdot\hat{p}\cdot x = (-y+y\cdot\hat{p}+\hat{p}-y\cdot\hat{p})$$

$$= (-y+\hat{p})\cdot x$$

1.5. הפונקציה $\ell^{CE}(y,\hat{p})$ קמורה וגזירה ביחס ל-w, הבעיה (P) היא בעיית מינימיזציה של סכום סופי של פונקציות קמורות וגזירות ביחס ל-w ולכן הבעיה כולה קמורה (סכום סופי של פונקציות קמורות הוא קמור) וגזירה ביחס ל-w (סכום של גזירות היא פונקציה גזירה).
בנוסף, על מנת למצוא מינימום של פונקציה קמורה ניתן לגזור ולהשוות ל-0 כי נקבל את נקודת המינימום היחידה של הפוקנציה (קיימת נקודת מינימום יחידה עבור פונקציות קמורות) כאשר ערך הפונקציה בנקודת המינימום הזו הוא פתרון הבעיה (P).
לשם כך נגזור את הביטוי:

$$\nabla_w \sum_{i=1}^{m} \ell^{CE}(y_i, \hat{p}_i) = \sum_{i=1}^{m} \nabla_w \ell^{CE}(y_I, \hat{p}_i) = \sum_{i=1}^{m} \left(-y_i + \hat{p_i}\right) \cdot x_i$$

נשווה את הגרדיאנט ל-0 ונקבל:

$$\sum_{i=1}^{m} \left(-y_i + \hat{p_i}\right) \cdot x_i = 0$$

נתון כי $\{x_i\}_{i=1}^{m}$ הם בלתי תלויים לינארית ולכן לא קיים צ"ל שבו לא כל המקדמים הם 0 של סט ערכי ה-x אשר שווה ל-0, כלומר ערך הסכום של השוות הגרדיאנט יכול להיות 0 אם ורק אם כל המקדמים הם 0.
ידוע כי $\forall 1 \leqslant i \leqslant m : y_i \in \{0,1\}$, ידוע בנוסף כי $\hat{p_i} = \sigma\left(w^T x_i\right)$ ולכן נקבל כי:

$$\begin{aligned} \forall 1 \leqslant i \leqslant m:\ -y_i + \hat{p_i} = 0 &\Longrightarrow y_i = \hat{p_i} \\ &\Longrightarrow \hat{p_i} = 0 \vee \hat{p_i} = 1 \\ &\Longrightarrow \sigma\left(w^T x_i\right) = 0 \vee \sigma\left(w^T x_i\right) = 1 \end{aligned}$$

אך לפי הגדרת הפונקציה $\sigma(z)$ לכל $z \in \mathbb{R}$ מתקיים כי $\sigma(z) \in (0,1)$ ולכן שתי האפשרויות שקיבלנו לא אפשרויות, כלומר לבעיה (P) אין פתרון עבור סט של x_i בלתי תלויים לינארית.

2.

2.1 נראה כי מתקיים $F_{\alpha\cdot\Theta}(x) = c\cdot F_\Theta(x)$ עבור סקלר $c\in\mathbb{R}$.

נסתכל על רשת נוירונים בעלת עומק L בעלת שכבות ליניאריות בעלת המשקולות הבאות:

$$\Theta = \left(W^{(1)}, W^{(2)},\ldots, W^{(L-1)}, W^{(L)}\right)$$

עבור פונקציית אקטיבציה $\sigma = ReLU$ הנתונה בשאלה.

נתון כי:

$$h^{(1)}(x) = \sigma\left({W^{(1)}}^T x\right)$$
$$h^{(\ell)}(x) = \sigma\left({W^{(l-1)}}^T\cdot h^{(l-1)}(x)\right)$$
$$F_\Theta(x) = {w^{(L)}}^T\cdot h^{(L-1)}(x)$$

עבור $\alpha\in\mathbb{R}_{>0}$ מתקיים:

$$F_{\alpha\cdot\Theta}(x) = \alpha\cdot {w^{(L)}}^T\cdot h^{(L-1)}(x) = \alpha\cdot {w^{(L)}}^T\cdot\sigma\left(\alpha\cdot {W^{(L-1)}}^T\cdot h^{(L-2)}(x)\right)$$
$$= \alpha^2\cdot {w^{(L)}}^T\cdot\sigma\left({W^{(L-1)}}^T\cdot h^{(L-2)}(x)\right)$$
$$= \alpha^2\cdot {w^{(L)}}^T\cdot\sigma\left({W^{(L-1)}}^T\cdot\sigma\left(\alpha\cdot {W^{(L-2)}}^T\cdot h^{(L-3)}(x)\right)\right)$$
$$= \alpha^3\cdot {w^{(L)}}^T\cdot\sigma\left({W^{(L-1)}}^T\cdot\sigma\left({W^{(L-2)}}^T\cdot h^{(L-3)}(x)\right)\right)$$
$$= \ldots\ = \alpha^k\cdot {w^{(L)}}^T\cdot\sigma\left({W^{(L-1)}}^T\cdot\sigma\left(\sigma\left(\ldots\left({W^{(L-k+1)}}^T\cdot h^{(L-k)}(x)\right)\right)\right)\right)$$
$$= \alpha^k\cdot F_\Theta(x)$$

כלומר התנאי הנדרש מתקיים עבור $c = \alpha^k\in\mathbb{R}$.

2.2 נחשב את ההתסברות להתכנסות עבור $\alpha\to\infty$:

$$\lim_{\alpha\to\infty}\frac{1}{1+e^{-F_{\alpha\cdot\Theta}(x)}} = \lim_{\alpha\to\infty}\frac{1}{1+e^{-\alpha^L F_\Theta(x)}} = \lim_{\alpha\to\infty}\frac{1}{1+\frac{1}{e^{\alpha^L F_\Theta(x)}}}$$

נחלק למקרים לפי הערך של $F_\Theta(x)$:

$$\lim_{\alpha\to\infty}\frac{1}{1+\frac{1}{e^{\alpha^L F_\Theta(x)}}} = \begin{cases}\frac{1}{1+\text{"}\infty\text{"}} = 0 & F_\Theta(x) < 0\\ \frac{1}{1+1} = \frac{1}{2} & F_\Theta(x) = 0\\ \frac{1}{1+\frac{1}{\text{"}\infty\text{"}}} = \frac{1}{1} = 1 & F_\Theta(x) > 0\end{cases}$$

כלומר נקבל כי ההסתברות להתכנסות הפלט תהיה:

$$\lim_{\alpha\to\infty}\frac{1}{1+e^{-F_{\alpha\cdot\Theta}(x)}} = \begin{cases}0 & F_\Theta(x) < 0\\ \frac{1}{2} & F_\Theta(x) = 0\\ 1 & F_\Theta(x) > 0\end{cases}$$