Intro to Machine Learning: Assignment #4

May 12, 2023

Dor Bourshan
315780122

Theory Questions

1 SVM with multiple classes.

יהיו $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ ויהיו $y_1, \ldots, y_n \in [K]$ labels מתוך קבוצה בת K labels.

נמצא מפריד $\mathbf{w}_j$ לכל אחד מהמחלקות $j \in [K]$.

נגדיר את פונקציית loss הבאה ($multiclass\ hinge - loss$) :

$$\ell\left(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i\right) = \max_{j \in [K]} \left(\langle \mathbf{w}_j, \mathbf{x}_i \rangle - \langle \mathbf{w}_{y_i}, \mathbf{x}_i \rangle + \mathbf{1}\left(j \neq y_i\right)\right)$$

נגדיא את בעיית ה multiclass SVM :

$$f\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i\right)$$

אחרי תהליך הלמידה אנחנו נסווג נקודות באופן הבא $\arg\max_{j \in [K]} \langle \mathbf{w}_j, \mathbf{x}_i \rangle$. נאמר כי אנחנו במקרה בו ה data מופרד. כלומר, קיימים $\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*$ כך ש $y_i = \arg\max_y \left\langle \mathbf{w}_y^*, \mathbf{x}_i \right\rangle$ לכל i. נראה שכל $\mathbf{w}_1, \ldots, \mathbf{w}_K$ שממנמנים את $f\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right)$ משיגים שגיאה אפס על הקלספיקציה.

אם כן, אנו נדרשים לבצע

$$\min_{\mathbf{w}_1, \ldots, \mathbf{w}_K} f\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right)$$

תחילה נפשט את הפונקציה להצגה אינפורמטיבית :

$$\begin{aligned} f\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) &= \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i\right) = \frac{1}{n} \sum_{i=1}^{n} \max_{j \in [K]} \left(\langle \mathbf{w}_j, \mathbf{x}_i \rangle - \langle \mathbf{w}_{y_i}, \mathbf{x}_i \rangle + \mathbf{1}\left(j \neq y_i\right)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \max \left\{ \max_{\substack{j \in [K] \\ j \neq y_i}} \left(\langle \mathbf{w}_j, \mathbf{x}_i \rangle - \langle \mathbf{w}_{y_i}, \mathbf{x}_i \rangle + 1\right), 0 \right\} \end{aligned}$$

כעת אנו יכולים לראות כי

$$f\left(\mathbf{w}_1, \ldots, \mathbf{w}_K\right) \geq 0 \qquad (1)$$

נראה כי

$$\ell\left(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{x}_i, y_i\right) \geq \ell_{1-0}\left(h\left(\mathbf{x}_i\right), y_i\right)$$

כלומר אנו נדרשים להראות כי :

$$\max_{j \in [K]} \left(\langle \mathbf{w}_j, \mathbf{x}_i \rangle - \langle \mathbf{w}_{y_i}, \mathbf{x}_i \rangle + \mathbf{1}\left(j \neq y_i\right)\right) \geq \mathbf{1}\left(\arg\max_{j \in [K]} \langle \mathbf{w}_j, \mathbf{x}_i \rangle \neq y_i\right)$$

נפריד למקרים :

- עבור $\arg\max_{j\in[K]}\langle \mathbf{w}_j,\mathbf{x}_i\rangle = y_i$ אנו נקבל כי

$$\max_{j\in[K]}\left(\langle \mathbf{w}_j,\mathbf{x}_i\rangle - \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle + \mathbf{1}\left(j\neq y_i\right)\right) \overset{1}{\geq} 0$$

- אם $\arg\max_{j\in[K]}\langle \mathbf{w}_j,\mathbf{x}_i\rangle \neq y_i$ מתקיים כי קיים $j\neq y_i$ כך ש $j = \arg\max_{j\in[K]}\langle \mathbf{w}_j,\mathbf{x}_i\rangle$ כלומר

$$\begin{aligned}\langle \mathbf{w}_j,\mathbf{x}_i\rangle \geq \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle &\implies \langle \mathbf{w}_j,\mathbf{x}_i\rangle - \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle \geq 0 \implies \langle \mathbf{w}_j,\mathbf{x}_i\rangle - \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle + 1 \geq 1\\ &\implies \max_{j\in[K]}\left(\langle \mathbf{w}_j,\mathbf{x}_i\rangle - \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle + \mathbf{1}\left(j\neq y_i\right)\right) \geq 1\\ &\implies \max_{j\in[K]}\left(\langle \mathbf{w}_j,\mathbf{x}_i\rangle - \langle \mathbf{w}_{y_i},\mathbf{x}_i\rangle + \mathbf{1}\left(j\neq y_i\right)\right) \geq \mathbf{1}\left(\arg\max_{j\in[K]}\langle \mathbf{w}_j,\mathbf{x}_i\rangle \neq y_i\right)\end{aligned}$$

כלומר הראנו כי

$$\ell\left(\mathbf{w}_1,\ldots,\mathbf{w}_K,\mathbf{x}_i,y_i\right) \geq \ell_{1-0}\left(h\left(\mathbf{x}_i\right),y_i\right) \tag{2}$$

נראה כי קיימים $\mathbf{w}_1^{**},\ldots,\mathbf{w}_K^{**}$ כך ש $f\left(\mathbf{w}_1^{**},\ldots,\mathbf{w}_K^{**}\right)=0$ ונקבל מ 1 כי הנ"ל מנמנמים את f.

נזכור כי ה data מופרד. כלומר, קיימים $\mathbf{w}_1^{*},\ldots,\mathbf{w}_K^{*}$ כך ש $y_i = \arg\max_y\left\langle \mathbf{w}_y^*,\mathbf{x}_i\right\rangle$ לכל i.

נסמן

$$\max_{1\leq i\leq n}\max_{\substack{j\in[K]\\ j\neq y_i}}\left(\left\langle \mathbf{w}_j^*,\mathbf{x}_i\right\rangle - \left\langle \mathbf{w}_{y_i}^*,\mathbf{x}_i\right\rangle\right) = M$$

ונשים לב כי M מספר שלילי שכן ה data מופרד וכן $y_i = \arg\max_y\left\langle \mathbf{w}_y^*,\mathbf{x}_i\right\rangle$. נסמן $\mathbf{w}_i^{**} = -\frac{1}{M}\mathbf{w}_i^*$. מתקיים כי:

$$\max_{\substack{j\in[K]\\ j\neq y_i}}\left(\left\langle \mathbf{w}_j^{**},\mathbf{x}_i\right\rangle - \left\langle \mathbf{w}_{y_i}^{**},\mathbf{x}_i\right\rangle\right) \leq -\frac{1}{M}M = -1$$

לפיכך

$$\max_{\substack{j\in[K]\\ j\neq y_i}}\left(\left\langle \mathbf{w}_j^{**},\mathbf{x}_i\right\rangle - \left\langle \mathbf{w}_{y_i}^{**},\mathbf{x}_i\right\rangle\right) + 1 \leq 0$$

ולכן

$$\max\left\{\max_{\substack{j\in[K]\\ j\neq y_i}}\left(\left\langle \mathbf{w}_j^{**},\mathbf{x}_i\right\rangle - \left\langle \mathbf{w}_{y_i}^{**},\mathbf{x}_i\right\rangle + 1\right), 0\right\} = 0$$

כלומר

$$f\left(\mathbf{w}_1^{**},\ldots,\mathbf{w}_K^{**}\right)=\frac{1}{n}\sum_{i=1}^{n}\max\left\{\max_{\substack{j\in[K]\\ j\neq y_i}}\left(\left\langle\mathbf{w}_j^{**},\mathbf{x}_i\right\rangle-\left\langle\mathbf{w}_{y_i}^{**},\mathbf{x}_i\right\rangle+1\right),0\right\}=\frac{1}{n}\sum_{i=1}^{n}0=0$$

כלומר הראנו כי מנמום הפונקציה, משמעו כי קיימים $\mathbf{w}_1,\ldots,\mathbf{w}_K$ כך ש $f\left(\mathbf{w}_1,\ldots,\mathbf{w}_K\right)=0$.

אם כן יהיו $\mathbf{w}_1,\ldots,\mathbf{w}_K$ כך ש $f\left(\mathbf{w}_1,\ldots,\mathbf{w}_K\right)=0$ מתקבל כי :

$$0=f\left(\mathbf{w}_1,\ldots,\mathbf{w}_K\right)=\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathbf{w}_1,\ldots,\mathbf{w}_K,\mathbf{x}_i,y_i\right)\overset{2}{\geq}\frac{1}{n}\sum_{i=1}^{n}\ell_{1-0}\left(h\left(\mathbf{x}_i\right),y_i\right)\geq 0$$

לפיכך

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{1-0}\left(h\left(\mathbf{x}_i\right),y_i\right)=0$$

כנדרש.

□

2 Expressivity of ReLU networks.

מגדירים את ReLU פונקצית אקטיבציה, באופן הבא :

$$h(x) = x^{+} = \max\{0, x\}$$

נראה כי ניתן לממש את הפונקציה $f(x_1, x_2) = \max\{x_1, x_2\}$ באמצעות one hidden layer המורכבת מפונקציית האקטיבציה ReLU. ניתן להניח כי אין פונקציית אקטיבציה לאחר השכבה האחרונה.

אם כן

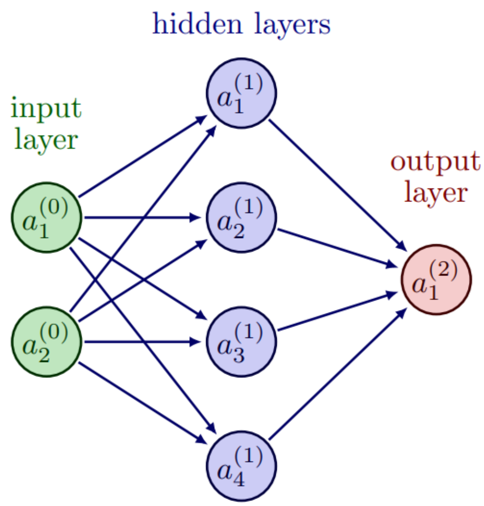

Figure 1: neural network

כאשר נגדיר :

$$a_1^{(1)} = \frac{x_1 + x_2}{2}$$
$$a_2^{(1)} = \frac{x_1 - x_2}{2}$$
$$a_3^{(1)} = \frac{-x_1 + x_2}{2}$$
$$a_4^{(1)} = \frac{-x_1 - x_2}{2}$$

וכן

$$z_1^{(1)} = \max\left\{\frac{x_1 + x_2}{2}, 0\right\}$$
$$z_2^{(1)} = \max\left\{\frac{x_1 - x_2}{2}, 0\right\}$$
$$z_3^{(1)} = \max\left\{\frac{-x_1 + x_2}{2}, 0\right\}$$
$$z_4^{(1)} = \max\left\{\frac{-x_1 - x_2}{2}, 0\right\}$$

וכן

$$a_1^{(2)} = \sum_{i=1}^{3} z_i^{(1)} - z_4^{(1)}$$

נראה עבור חלוקה למקרים:

1. מתקיים כי $x_1, x_2 \geq 0$
ולכן מתקיים כי $z_1^{(1)} = \frac{x_1+x_2}{2}$ וכן $z_4^{(1)} = 0$
(א) מתקיים כי $x_1 \geq x_2$
מתקיים כי $z_3^{(1)} = 0$ וכן $z_2^{(1)} = \frac{x_1-x_2}{2}$. כמו כן $z_2^{(1)} = \frac{x_1-x_2}{2} = \frac{|x_1-x_2|}{2}$.
(ב) מתקיים כי $x_1 < x_2$
מתקיים כי $z_2^{(1)} = 0$ וכן $z_3^{(1)} = \frac{-x_1+x_2}{2}$. כמו כן $z_3^{(1)} = \frac{-x_1+x_2}{2} = \frac{|x_1-x_2|}{2}$.
ניתן לראות אם כן כי

$$a_1^{(2)} = \sum_{i=1}^{3} z_i^{(1)} - z_4^{(1)} = \frac{x_1 + x_2}{2} + \frac{|x_1 - x_2|}{2} = \max\{x_1, x_2\} = f(x_1, x_2)$$

2. מתקיים כי $x_1, -x_2 \geq 0$
ולכן מתקיים כי $\frac{|x_1-x_2|}{2} = \frac{x_1-x_2}{2} = z_2^{(1)}$ וכן $z_3^{(1)} = 0$
(א) מתקיים כי $x_1 \geq -x_2$
מתקיים כי $z_1^{(1)} = \frac{x_1+x_2}{2}$ וכן $z_4^{(1)} = 0$. .
(ב) מתקיים כי $x_1 < -x_2$
מתקיים כי $z_1^{(1)} = 0$ וכן $z_4^{(1)} = \frac{-x_1-x_2}{2}$. ולכן $-z_4^{(1)} = \frac{x_1+x_2}{2}$
ניתן לראות אם כן כי

$$a_1^{(2)} = \sum_{i=1}^{3} z_i^{(1)} - z_4^{(1)} = \frac{x_1 + x_2}{2} + \frac{|x_1 - x_2|}{2} = \max\{x_1, x_2\} = f(x_1, x_2)$$

3. מתקיים כי $-x_1, -x_2 \geq 0$
ולכן מתקיים כי $= \frac{-x_1-x_2}{2} = z_4^{(1)}$ וכן $\frac{x_1+x_2}{2} = -z_4^{(1)}$ וכן $z_1^{(1)} = 0$
(א) מתקיים כי $-x_1 \geq -x_2$
מתקיים כי $\frac{|x_1-x_2|}{2} = \frac{-x_1+x_2}{2} = z_3^{(1)}$ וכן $z_2^{(1)} = 0$.
(ב) מתקיים כי $-x_1 < -x_2$
מתקיים כי $z_3^{(1)} = 0$ וכן $\frac{|x_1-x_2|}{2} = \frac{x_1-x_2}{2} = z_2^{(1)}$.
ניתן לראות אם כן כי

$$a_1^{(2)} = \sum_{i=1}^{3} z_i^{(1)} - z_4^{(1)} = \frac{x_1 + x_2}{2} + \frac{|x_1 - x_2|}{2} = \max\{x_1, x_2\} = f(x_1, x_2)$$

4. מתקיים כי $-x_1, x_2 \geq 0$
ולכן מתקיים כי $\frac{|x_1-x_2|}{2} = \frac{-x_1+x_2}{2} = z_3^{(1)}$ וכן $z_2^{(1)} = 0$

(א) מתקיים כי $-x_1 \geq x_2$
מתקיים כי $\frac{-x_1-x_2}{2} = z_4^{(1)}$ וכן $-z_4^{(1)} = \frac{x_1+x_2}{2}$ וכן $z_1^{(1)} = 0$.

(ב) מתקיים כי $-x_1 < x_2$
מתקיים כי $z_4^{(1)} = 0$ וכן $z_1^{(1)} = \frac{x_1+x_2}{2}$.

ניתן לראות אם כן כי

$$a_1^{(2)} = \sum_{i=1}^{3} z_i^{(1)} - z_4^{(1)} = \frac{x_1+x_2}{2} + \frac{|x_1-x_2|}{2} = \max\{x_1, x_2\} = f(x_1, x_2)$$

סה"כ עברנו על כל המקרים וקיבלנו כי

$$a_1^{(2)} = \max\{x_1, x_2\} = f(x_1, x_2)$$

$\square$

3 Soft SVM with ℓ^2 penalty.

נניח שקיימת הבעיה הבאה:

$$\min_{\mathbf{w}\in\mathbb{R}^d,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{w}\|_2^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2$$
$$\text{s.t.} y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right)\geq 1-\xi_i \;\; \forall i=1,\ldots,n$$

3.1 section (a)

נראה תחילה כי אם נוסיף את האילוץ $\xi_i \geq 0$ לא נשנה את הפיתרון המינימלי.

אם כן תהיי $I\subseteq[n]$ קבוצת אינדקסים כך שלכל $i\in I$ מתקיים

$$\xi_i<0$$

יהיי $\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n$ פתרון אופטימלי. מתקיים אם כן כי

$$y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right)\geq 1-\xi_i\geq 1+\xi_i \;\; \forall i\in I$$
$$y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right)\geq 1-\xi_i \qquad \forall i\in[n]\setminus I$$

נשים לב כי בפונקציה המנומנמת, אין חשיבות לסימן של ξ_i ולכן עבור $i\in I$ אם נחליף את ξ_i ב $-\xi_i$ נקבל שהאילוצים מתקיימים וכן ערך הפונקציה המנומנמת זהה.

כלומר נוכל לעבור לבעיה הבאה:

$$\min_{\mathbf{w}\in\mathbb{R}^d,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{w}\|_2^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2$$
$$\text{s.t.} y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right)\geq 1-\xi_i \;\; \forall i=1,\ldots,n$$
$$\xi_i\geq 0 \qquad \forall i=1,\ldots,n$$

□

3.2 section (b)

נרשום את פונקציית הלגרנג'יאן של הבעיה:

$$\min_{\mathbf{w}\in\mathbb{R}^d,b\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{w}\|_2^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2$$
$$\text{s.t.} 0\geq 1-\xi_i-y_i\left(\langle\mathbf{w},\mathbf{x}_i\rangle+b\right) \;\; \forall i=1,\ldots,n$$
$$0\geq -\xi_i \qquad \forall i=1,\ldots,n$$

אם כן

$$\mathcal{L}\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\left(\langle\mathbf{w}, \mathbf{x}_i\rangle + b\right)\right) + \sum_{i=1}^{n}\beta_i\left(-\xi_i\right)$$

כאשר אין בבעיה אילוצי שוויון.

□

3.3 section (c)

נמנמם את הפונקציה ביחס ל $\mathbf{w}, b, \boldsymbol{\xi}$. זוהי תהיה הפונקציה הדואלית.

כלומר

$$g\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \mathcal{L}\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$$

היא הפונקציה הדואלית.

ראשית נסדר את הלגרנג׳יאן.

$$\mathcal{L}\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \sum_{i=1}^{n}\alpha_i y_i\langle\mathbf{w}, \mathbf{x}_i\rangle + \frac{C}{2}\sum_{i=1}^{n}\left(\xi_i^2 - \frac{2}{C}\xi_i\left(\alpha_i + \beta_i\right)\right) + \sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i$$

נשים לב כי אם

$$\sum_{i=1}^{n}\alpha_i y_i \neq 0$$

אין לבעיה מינימום. כיוון שלכל ערך קטן כרצוננו ניתן לבחור b.

כעת ניתן לראות כי זו פונקציה של סכומים של פונקציות קמורות, ולכן קמורה (מתרגיל 3). והפונקציה דיפרנציאבלית כסכום של פונקציות דיפרנציאבליות.

אם קיים מינימום הוא נמצא בנקודה בה ההגרדיאנט מתאפס.

$$\begin{aligned}
\nabla_{\mathbf{w}}\mathcal{L} &= \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i = 0 \implies \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i \\
\frac{\partial\mathcal{L}}{\partial b} &= \sum_{i=1}^{n}\alpha_i y_i = 0 \\
\frac{\partial\mathcal{L}}{\partial\xi_i} &= C\cdot\xi_i - \left(\alpha_i + \beta_i\right) = 0 \implies \xi_i = \frac{\alpha_i + \beta_i}{C}
\end{aligned}$$

נציב ונקבל

$$\begin{aligned}
\mathcal{L}\left(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) &= \frac{1}{2}\|\mathbf{w}\|_2^2 - \mathbf{w}^\top \underbrace{\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}_{\mathbf{w}} + \sum_{i=1}^{n} \alpha_i - \frac{1}{2C}\sum_{i=1}^{n}\left(\alpha_i + \beta_i\right)^2 + b\sum_{i=1}^{n}\alpha_i y_i \\
&= -\frac{1}{2}\|\mathbf{w}\|_2^2 - \frac{1}{2C}\sum_{i=1}^{n}\left(\alpha_i + \beta_i\right)^2 + \sum_{i=1}^{n}\alpha_i \\
&= \boxed{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j - \frac{1}{2C}\sum_{i=1}^{n}\left(\alpha_i + \beta_i\right)^2 + \sum_{i=1}^{n}\alpha_i}
\end{aligned}$$

3.4 section (d)

הפונקציה הדואלית:

$$g\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j - \frac{1}{2C}\sum_{i=1}^{n}\left(\alpha_i + \beta_i\right)^2 + \sum_{i=1}^{n}\alpha_i$$

הבעיה הדואלית:

$$\begin{aligned}
&\max g\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) \\
&\sum_{i=1}^{n}\alpha_i y_i = 0 \\
&\alpha_i, \beta_i \geq 0 \quad \forall i = 1, \ldots, n
\end{aligned}$$

□

4 Gradient of cross-entropy loss over softmax.

יהיי $\mathbf{y} \in \{0,1\}^d$ וקטור בעל כניסה אחת של 1 והשאר 0. נגדיר את הפונקצית loss $\ell_{\mathbf{y}} : \mathbb{R}^d \to \mathbb{R}$ הנקראת $cross-entropy\ loss$ ע"י :

$$\ell_{\mathbf{y}}(\mathbf{w}) = -\mathbf{y}\ln(\text{softmax}(\mathbf{w}))$$

כאשר

$$\text{softmax}(\mathbf{w}) = \left(\frac{e^{w_1}}{\sum_{j=1}^{d} e^{w_j}}, \ldots, \frac{e^{w_1}}{\sum_{j=1}^{d} e^{w_j}}\right)$$

וכן הלוגריתם מודר על וקטור כאיבר איבר.

נחשב את הדרדיאנט של $\ell_{\mathbf{y}}(\mathbf{w})$:

$$\ell_{\mathbf{y}}(\mathbf{w}) = -\mathbf{y}\ln(\text{softmax}(\mathbf{w})) = -\sum_{i=1}^{n} y_i \ln\left(\frac{e^{w_i}}{\sum_{j=1}^{d} e^{w_j}}\right) = \sum_{i=1}^{n}\left(-y_i \ln(e^{w_i}) + y_i \ln\left(\sum_{j=1}^{d} e^{w_j}\right)\right) =$$
$$= \sum_{i=1}^{n} -y_i w_i + \sum_{i=1}^{n} y_i \ln\left(\sum_{j=1}^{d} e^{w_j}\right)$$

נגזור

$$\frac{\partial \ell}{\partial w_q} = -y_q + \sum_{i=1}^{n} y_i \frac{e^{w_q}}{\sum_{j=1}^{d} e^{w_j}}$$

נזכור כי יש רק כניסה אחת ל $\mathbf{y}$ ששווה ל 1 ושאר הכניסות אפסים.

$$\frac{\partial \ell}{\partial w_q} = -y_q + \frac{e^{w_q}}{\sum_{j=1}^{d} e^{w_j}}$$

ולכן

$$\nabla \ell_{\mathbf{y}}(\mathbf{w}) = \text{softmax}(\mathbf{w}) - \mathbf{y}$$

□

Programming Assignment