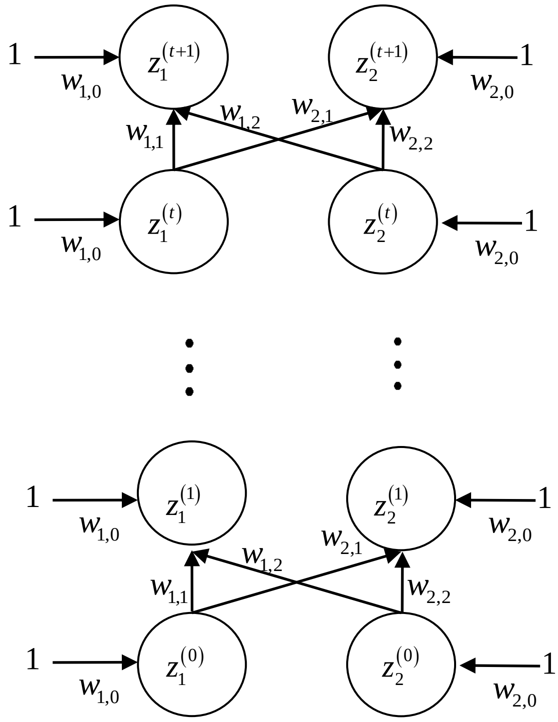
אוניברסיטת ת"א, ביה"ס להנדסת חשמל, למידת מכונה סטטיסטית

פתרון תרגיל בית 6

שאלה 1

ניתן לתאר את פעולת הרשת באופן אקויולנטי ע"י רשת feedforward בעלת שכבת כניסה (2 תאים), 4 שכבות נסתרות (2 תאים כ"א) ושכבת מוצא (2 תאים). ברשת הזאת אנחנו מאלצים שוויון בין המשקולות שמחברות בין השכבות כמתואר בציור:



נשים לב ש- $\left(z_1^{(0)}, z_2^{(0)}\right) = \left(x_1, x_2\right)$. כמו כן אצלנו $t = 4$ ו- $\left(y_1, y_2\right) \triangleq \left(z_1^{(5)}, z_2^{(5)}\right)$. לכן המרנו את הבעייה לבעית אימון רשת feedforward עם אילוצי שוויון בין פרמטרים.

אנחנו רוצים להביא למינימום את $E\left(\mathbf{w}\right) \triangleq \sum_{l=1}^{L} E_l\left(\mathbf{w}\right)$ כאשר

$E_l\left(\mathbf{w}\right) = \frac{1}{2}\left(y_1^{(l)} - t_1^{(l)}\right)^2 + \frac{1}{2}\left(y_2^{(l)} - t_2^{(l)}\right)^2$. יש להראות כיצד ניתן לחשב באופן יעיל את $\nabla E_l\left(\mathbf{w}\right)$.

לאחר מכן ניתן להשתמש בחישוב הזה באופן סטנדרטי, כפי שלמדנו בכיתה, לצורך רישום אלגוריתם גרדיינט מסוג batch או מסוג stochastic gradient descent.

קל לראות שכלל השרשרת לחישוב הגרדיינט עדיין עובד. עבור כל דגימה בבסיס הנתונים $l=1,2,\ldots,L$, ראשית נחשב את כל מוצאי התאים ע"י מעבר קדמי (forward propagation) ע"י אתחול $\left(z_1^{(0)},z_2^{(0)}\right)=\left(x_1^{(l)},x_2^{(l)}\right)$, ואח"כ שימוש בנוסחאות

$$z_1^{(t+1)}=\sigma\left(w_{1,0}+w_{1,1}z_1^{(t)}+w_{1,2}z_2^{(t)}\right)$$
$$z_2^{(t+1)}=\sigma\left(w_{2,0}+w_{2,1}z_1^{(t)}+w_{2,2}z_2^{(t)}\right)$$

עבור $t=0,1,2,3,4$. לאחר מכן נחשב את וקטורי השגיאה $\left(\delta_1^{(t)},\delta_2^{(t)}\right)$ עבור $t=5,4,3,2,1$ ע"י מעבר אחורי (backpropagation) תוך שימוש בנוסחאות:

$$\delta_1^{(5)}=z_1^{(5)}(1-z_1^{(5)})\cdot\left(z_1^{(5)}-t_1^{(l)}\right)$$
$$\delta_2^{(5)}=z_2^{(5)}(1-z_2^{(5)})\cdot\left(z_2^{(5)}-t_2^{(l)}\right)$$

ולאחר מכן

$$\delta_1^{(t)}=z_1^{(t)}(1-z_1^{(t)})\left(\delta_1^{(t+1)}w_{1,1}+\delta_2^{(t+1)}w_{2,1}\right)$$
$$\delta_2^{(t)}=z_2^{(t)}(1-z_2^{(t)})\left(\delta_1^{(t+1)}w_{1,2}+\delta_2^{(t+1)}w_{2,2}\right)$$

עבור $t=4,3,2,1$. לאחר מכן ניתן לחשב את $\frac{\partial E_l(\mathbf{w})}{\partial w_{ji}}$ ע"י שימוש בנוסחא:

$$\frac{\partial E_l}{\partial w_{ji}}=\sum_{t=1}^{5}\delta_j^{(t,l)}z_i^{(t-1,l)}$$

(כאשר $z_0^{(t-1,l)}\triangleq 1$).

הסבר לנוסחא האחרונה: נוסחא זאת נובעת מכלל השרשרת לחישוב נגזרות. נניח שיש לנו פונקציה $g\left(w_1,w_2\right)$ ונניח שמתקיים $w_1=w_2=w$. לאחר שחישבנו את הנגזרות $\frac{\partial g\left(w_1,w_2\right)}{\partial w_1},\frac{\partial g\left(w_1,w_2\right)}{\partial w_2}$ נוכל לקבל על פי כלל השרשרת:

$$\frac{\partial g(w,w)}{\partial w}=\left[\frac{\partial g\left(w_1,w_2\right)}{\partial w_1}\right]_{w_1=w_2=w}\cdot\frac{\partial w_1}{\partial w}+\left[\frac{\partial g\left(w_1,w_2\right)}{\partial w_2}\right]_{w_1=w_2=w}\cdot\frac{\partial w_2}{\partial w}$$
$$=\left[\frac{\partial g\left(w_1,w_2\right)}{\partial w_1}\right]_{w_1=w_2=w}+\left[\frac{\partial g\left(w_1,w_2\right)}{\partial w_2}\right]_{w_1=w_2=w}$$

Recurrent backpropagation הוצג לראשונה ע"י Rumelhart et. al בספר Parallel Distributed Processing (פרק 8, עמודים 354-361).

שאלה 2

א. הטרנספורמציה מ- $\left(y_1, y_2, \ldots, y_{10}\right)$ ל- $\left(\frac{e^{y_1}}{\sum_{l=1}^{10} e^{y_l}}, \frac{e^{y_2}}{\sum_{l=1}^{10} e^{y_l}}, \ldots, \frac{e^{y_{10}}}{\sum_{l=1}^{10} e^{y_l}}\right)$ היא טרנספורמצית softmax. נסמן $r = \arg\max_i y_i$. אם $\forall i \neq r \quad y_r \gg y_i$ אזי הטרנספורמציה ממפה לוקטור שמקרב את וקטור היחידה ה- r-י. אפשר להתייחס אל האלמנטים של וקטור הטרנספורמציה כאל מידול ההסתברויות האפוסטריוריות של המחלקה בהינתן התמונה, ועל כן קריטריון האימון הוא של סבירות מירבית בהנחה שהדוגמאות השונות הן דגימות בת"ס של תמונות. בהתאם לכך אפשרות הגיונית לביצוע הזיהוי בהינתן $\left(y_1, y_2, \ldots, y_{10}\right)$ במוצא הרשת היא לסווג את התמונה למחלקה $r = \arg\max_i y_i$.

ב. יש להראות כיצד משתנה החישוב של $\nabla E_l\left(\mathbf{w}\right)$. נסמן את המשקולות שמחברות בין שכבת הכניסה לשכבה הנסתרת ב- $\left\{w_{ij}\right\}$ ואת המשקולות שמחברות בין השכבה הנסתרת לשכבת המוצא ב- $\left\{\bar{w}_{ij}\right\}$. בהינתן וקטור הכניסה $x_1, x_2, \ldots, x_I$, המעבר הקדמי מתבצע באופן סטנדרטי כדלקמן:

$$a_j = \sum_{i=0}^{I} w_{ji} x_i, \quad z_j = \sigma\left(a_j\right) \qquad j = 1, 2, \ldots, H$$
$$y_k = \sum_{i=0}^{H} \bar{w}_{ki} z_i \qquad k = 1, 2, \ldots, 10$$

כאשר אנחנו מגדירים $x_0 \triangleq z_0 \triangleq 1$. נסמן את אותות השגיאה $\delta_j \triangleq \frac{\partial E_l}{\partial a_j}$ ו- $\bar{\delta}_k \triangleq \frac{\partial E_l}{\partial y_k}$.

בדומה למה שלמדנו בכיתה,

$$\frac{\partial E_l}{\partial w_{ji}} = \frac{\partial E_l}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ji}} = \delta_j x_i$$
$$\frac{\partial E_l}{\partial \bar{w}_{ji}} = \frac{\partial E_l}{\partial y_j} \cdot \frac{\partial y_j}{\partial \bar{w}_{ji}} = \bar{\delta}_j z_i$$

עבור הבעיה שלנו,

$$E_l = \log \frac{\exp\left(y_{c^{(l)}}\right)}{\sum_{k=1}^{10} e^{y_k}} = y_{c^{(l)}} - \log \sum_{k=1}^{10} e^{y_k}$$

$$\bar{\delta}_j = \frac{\partial E_l}{\partial y_j} = \begin{cases} 1 - \frac{e^{y_j}}{\sum_{k=1}^{10} e^{y_k}} & \text{if } j = c^{(l)} \\ -\frac{e^{y_j}}{\sum_{k=1}^{10} e^{y_k}} & \text{if } j \neq c^{(l)} \end{cases}$$

וכמו שלמדנו בכיתה

$$\delta_j = \sigma'\left(a_j\right) \sum_k \bar{w}_{kj} \bar{\delta}_k = z_j \left(1 - z_j\right) \sum_k \bar{w}_{kj} \bar{\delta}_k$$

כך מתקבל אלגוריתם רקורסיבי יעיל לחישוב הגרדיינט. ניתן להשתמש בגרסת ה- batch או בגרסת ה- stochastic gradient descent של האלגוריתם כפי שלמדנו בכיתה.

שימו לב שהואיל ובשאלה הזאת יש לנו מקסימיזציה במקום מינימיזציה, האלגוריתם הוא למעשה gradient ascent, ז״א הולכים בכוון הגרדיאנט, ובנוסחת העדכון יש לשנות את הסימן לפני איבר הגרדיאנט ממינוס לפלוס.