Αναγνώριση Προτύπων

2η Σειρά Αναλυτικών Ασκήσεων

Δημήτριος Κοκκίνης
03118896
dimkok00@gmail.com

Άσκηση 2.1

1. Έχουμε το δίκτυο $f(x) = f^L(f^{L-1}(f^{L-2}(\dots f^1(x))))$ όπου $1, \dots, L-2, L-1, L$ τα επίπεδα αυτού και $f^l = \sigma(W_l x + b_l)$ η συνάρτηση που χαρακτηρίζει κάθε layer l, με παραμέτρους W_l και b_l. Θεωρώντας γραμμική συνάρτηση ενεργοποίησης $\sigma(x) = x$, τότε το δίκτυο γράφεται στη μορφή:

$$f(x) = f^L(f^{L-1}(f^{L-2}(\dots f^4(f^3(f^2(f^1(x))))))) =$$

$$= f^L(f^{L-1}(f^{L-2}(\dots f^4(f^3(f^2(\sigma(W_1 x + b_1))))))) =$$

$$= f^L(f^{L-1}(f^{L-2}(\dots f^4(f^3(\sigma(W_2(W_1 x + b_1) + b_2)))))) =$$

$$= \dots = f^L(f^{L-1}(f^{L-2}(\dots f^4(W_3 W_2 W_1 x + W_3 W_2 b_1 + W_3 b_2 + b_3))))$$

$$= \dots = W_{L-1} W_{L-2} \dots W_1 x + W_{L-1} W_{L-2} \dots W_2 b_1 x + W_{L-1} W_{L-2} \dots W_3 b_2 x + \dots + W_{L-1} b_{L-2} x + b_{L-1}$$

$$= (W_{L-1} W_{L-2} \dots W_1 + W_{L-1} W_{L-2} \dots W_2 b_1 + W_{L-1} W_{L-2} \dots W_3 b_2 + \dots + W_{L-1} b_{L-2}) x + b_{L-1}$$

Οπότε τώρα, η συνάρτηση $f(x)$ είναι στη μορφή $Wx + b$ και επειδή $\sigma(x) = x$ άρα $\sigma(Wx + b) = Wx + b$. Όμως η $\sigma(Wx + b)$ εκφράζει ένα single layer network $g(x) = Wx + b$.

Είναι $\|f(x) - g(x)\| < \epsilon$, $\epsilon > 0$ άρα από θεώρημα καθολικής προσέγγισης η $f(x)$ είναι αντίστοιχη με την $g(x) = Wx + b$ δηλαδή $f(x) \approx g(x) = Wx + b$

2. Από τα δεδομένα διαπιστώνουμε ότι έχουμε 2 κόμβους, x_1 x_2, στο input layer, έχουμε ένα deep layer, a, με 4 κόμβους a_1, a_2, a_3, a_4.

Για το επίπεδο αυτό ισχύει ότι:

$$\frac{\partial a}{\partial x} = \begin{bmatrix} \frac{\partial a_1}{\partial x_1} & \frac{\partial a_1}{\partial x_2} \\ \frac{\partial a_2}{\partial x_1} & \frac{\partial a_2}{\partial x_2} \\ \frac{\partial a_3}{\partial x_1} & \frac{\partial a_3}{\partial x_2} \\ \frac{\partial a_4}{\partial x_1} & \frac{\partial a_4}{\partial x_2} \end{bmatrix} = W_1 = \begin{bmatrix} \lambda & \lambda \\ -\lambda & -\lambda \\ \lambda & -\lambda \\ -\lambda & \lambda \end{bmatrix} = J^{(1)}$$ Ιακωβιανή

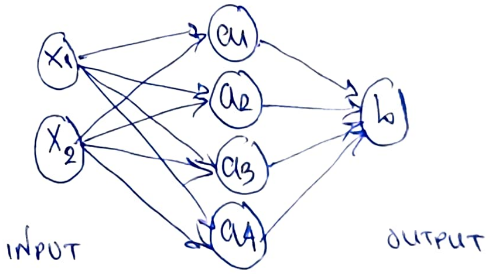
αφού είναι $a = W_1 x + b_1 = \begin{bmatrix} \lambda & \lambda \\ -\lambda & -\lambda \\ \lambda & -\lambda \\ -\lambda & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$

$$= \begin{bmatrix} \lambda(x_1 + x_2) \\ -\lambda(x_1 + x_2) \\ \lambda(x_1 - x_2) \\ -\lambda(x_1 - x_2) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

Έχουμε ένα output layer, b, με ένα κόμβο b για το οποίο ισχύει:

$$\frac{\partial b}{\partial x} = \begin{bmatrix} \frac{\partial b}{\partial a_1} & \frac{\partial b}{\partial a_2} & \frac{\partial b}{\partial a_3} & \frac{\partial b}{\partial a_4} \end{bmatrix} = W_2 = \begin{bmatrix} \mu & \mu & -\mu & -\mu \end{bmatrix} = J^{(2)}$$ ιακωβιανή

Επομένως το δίκτυο είναι το εξής:



Για το output layer ισχύει:

$$\tilde{f}(x_1, x_2) = b = W_2\, \sigma(W_1 x + b_1) = W_2\, \sigma(a) = \begin{bmatrix} \mu & \mu & -\mu & -\mu \end{bmatrix} \begin{bmatrix} \sigma(a_1) \\ \sigma(a_2) \\ \sigma(a_3) \\ \sigma(a_4) \end{bmatrix}$$

$$= \mu\sigma(a_1) + \mu\sigma(a_2) - \mu\sigma(a_3) - \mu\sigma(a_4)$$

$$= \mu\sigma(\lambda(x_1+x_2)) + \mu\sigma(-\lambda(x_1+x_2)) - \mu\sigma(\lambda(x_1+x_2)) - \mu\sigma(-\lambda(x_1-x_2))$$

3. Εκφράζουμε τη $\tilde{f}(x_1, x_2)$ ως quadratic γύρω από το $(0,0)$:

$$\tilde{f}(x_1,x_2) = \sigma(x_1,x_2) = \tilde{f}(0,0) + \tilde{f}_{x_1}(0,0)\,x_1 + \tilde{f}_{x_2}(0,0)\,x_2 + \frac{\tilde{f}_{x_1x_1}(0,0)}{2}x_1^2 + \tilde{f}_{x_1x_2}(0,0)\,x_1x_2 + \frac{\tilde{f}_{x_2x_2}}{2}x_2^2$$

- $\tilde{f}(0,0) = \mu\sigma(0) + \mu\sigma(0) - \mu\sigma(0) - \mu\sigma(0) = 0$
- $\tilde{f}_{x_1}(0,0) = \mu\lambda\dot{\sigma}(0) - \mu\lambda\dot{\sigma}(0) - \mu\lambda\dot{\sigma}(0) + \mu\lambda\dot{\sigma}(0) = 0$
- $\tilde{f}_{x_2}(0,0) = \mu\lambda\dot{\sigma}(0) - \mu\lambda\dot{\sigma}(0) + \mu\lambda\dot{\sigma}(0) - \mu\lambda\dot{\sigma}(0) = 0$
- $\tilde{f}_{x_1x_2}(0,0) = \dfrac{\partial \tilde{f}_{x_1}(0,0)}{\partial x_2} = \dots = \mu\lambda^2\ddot{\sigma}(0) + \mu\lambda^2\ddot{\sigma}(0) + \mu\lambda^2\ddot{\sigma}(0) + \mu\lambda^2\ddot{\sigma}(0)$
  $= 4\mu\lambda^2\ddot{\sigma}(0)$
- $\tilde{f}_{x_1x_1}(0,0) = \dfrac{\partial \tilde{f}_{x_1}(0,0)}{\partial x_1} = \dots = \mu\lambda^2\ddot{\sigma}(0) + \mu\lambda^2\ddot{\sigma}(0) - \mu\lambda^2\ddot{\sigma}(0) - \mu\lambda^2\ddot{\sigma}(0) = 0$
- $\tilde{f}_{x_2x_2}(0,0) = \dfrac{\partial \tilde{f}_{x_2}(0,0)}{\partial x_2} = \dots = \mu\lambda^2\ddot{\sigma}(0) + \mu\lambda^2\ddot{\sigma}(0) - \mu\lambda^2\ddot{\sigma}(0) - \mu\lambda^2\ddot{\sigma}(0) = 0$

οπότε $\tilde{f}(x_1,x_2) = 4\mu\lambda^2\ddot{\sigma}(0)x_1x_2$ (1) και $\lim\limits_{\lambda\to 0}\tilde{f}(x) = \lim\limits_{\lambda\to 0}\tilde{f}(x_1,x_2)$

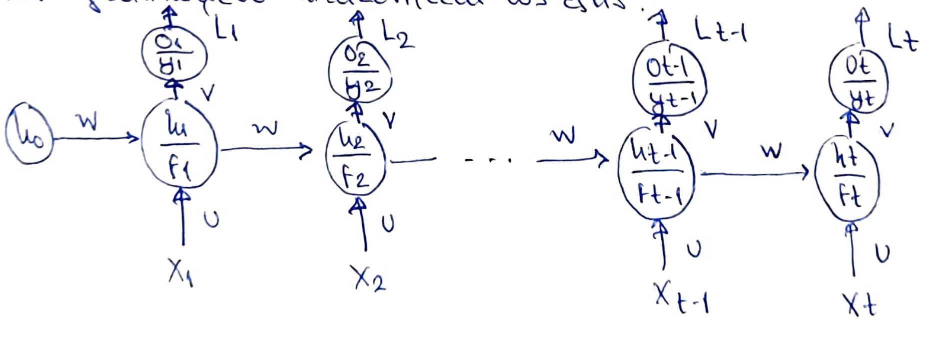
$\overset{(1)}{=} \lim\limits_{\lambda\to 0} 4\mu\lambda^2\ddot{\sigma}(0)x_1x_2$

$\underline{\ddot{\sigma}(0) = 1/4\lambda^2\mu}$ $\lim\limits_{\lambda\to 0} x_1x_2 = f(x)$

Επομένως για $\lambda\to 0$ ισχύει ότι $\tilde{f}(x) \to f(x)$

ΑΣΚΗΣΗ 2.2

Το RNN ξεδιπλωμένο απεικονίζεται ως εξής:



1. Σύμφωνα με το κανόνα της αλυσίδας η παράγωγος $\frac{\partial L_t}{\partial v}$ μπορεί να γραφεί ως εξής: $\dfrac{\partial L_t}{\partial v} = \dfrac{\partial L_t}{\partial o_t}\cdot\dfrac{\partial o_t}{\partial y_t}\cdot\dfrac{\partial y_t}{\partial v}$ (1)

Είναι $\frac{\partial y_t}{\partial v} = h_t$

Επομένως η (1) γράφεται: $\boxed{\frac{\partial L_t}{\partial v} = \frac{\partial L_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial y_t}\cdot h_t}$

2. Η $\frac{\partial L_t}{\partial w}$ γράφεται: $\frac{\partial L_t}{\partial w} = \frac{\partial L_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial y_t}\cdot\frac{\partial y_t}{\partial h_t}\cdot\frac{\partial h_t}{\partial h_\ell}$, $\frac{\partial h_\ell}{\partial w} = \frac{\partial L_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial h_t}\cdot\frac{\partial h_t}{\partial h_\ell}\cdot\frac{\partial h_\ell}{\partial w}$ (2)

ο όρος $\frac{\partial h_t}{\partial h_\ell}$ είναι η παράγωγος του hidden state τη χρονική στιγμή t αντίστοιχα με το hidden state τη χρονική στιγμή ℓ.

Τον όρο αυτό μπορούμε να τον γράψουμε ως:

$$\frac{\partial h_i}{\partial h_{i-1}} : \quad \frac{\partial h_t}{\partial h_\ell} = \frac{\partial h_t}{\partial h_{t-1}}\cdot\frac{\partial h_{t-1}}{\partial h_{t-2}}\cdots\frac{\partial h_{\ell+1}}{\partial h_\ell} = \prod_{i=\ell+1}^{t}\frac{\partial h_i}{\partial h_{i-1}} \quad (3)$$

και λόγω της σχέσης $h_t = f(w h_{t-1} + U x_t)$ η (3) γράφεται ως εξής:

$$\frac{\partial h_t}{\partial h_\ell} = \prod_{i=\ell+1}^{t}\frac{\partial h_i}{\partial h_{i-1}} = \prod_{i=\ell+1}^{t} w^T \mathrm{diag}[f'(h_{i-1})] \quad (4)$$

Επομένως η (2) από (4),(5) γράφεται: $\boxed{\frac{\partial L_t}{\partial w} = \frac{\partial L_t}{\partial o_t}\cdot\frac{\partial o_t}{\partial h_t}\cdot\left(\prod_{i=\ell+1}^{t} W^T \mathrm{diag}[f'(h_{i-1})]\right)\cdot\frac{\partial h_\ell}{\partial w}}$

3. Θεωρούμε ότι η activation function f είναι η ταυτοτική συνάρτηση.

Άρα είναι: $\frac{\partial h_t}{\partial h_\ell} = \prod_{i=\ell+1}^{t}\frac{\partial h_i}{\partial h_{i-1}} (w^T)^{t-\ell} \overset{t-\ell=k}{=\!=} (w^T)^k$

Έστω $\lambda_1, \lambda_2, \dots, \lambda_n$ οι ιδιοτιμές του w με $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$ και τα αντίστοιχα ιδιοδιανύσματα $v_1, v_2, \dots, v_n$ που σχηματίζουν μια διανυσματική βάση.

Ο w είναι τετραγωνικός, διαγωνοποιήσιμος και μπορούμε να χρησιμοποιήσουμε τη σχέση $v_i^T (w^T)^k = \lambda_i^k v_i^T$. Μπορούμε επίσης να γράψουμε το διάνυσμα γραμμής $\frac{\partial L_t}{\partial h_t}$ χρησιμοποιώντας τη βάση $v_1, v_2, \dots, v_n$: $\frac{\partial L_t}{\partial h_t} = \sum_{i=1}^{N} c_i v_i^T$

Επιλέγουμε j τέτοιο ώστε $c_j \neq 0$ και για κάθε $j' < j$ είναι $c_{j'} = 0$.

Έχουμε: $\frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial h_\ell} = c_j \lambda_j^k v_j^T + \lambda_j^k \sum_{i=j+1}^{n} c_i \frac{\lambda_i^k}{\lambda_j^k} v_i^T$ (6)

Για $i > j$ είναι $|\lambda_i| < |\lambda_j| \Rightarrow \left|\frac{\lambda_i}{\lambda_j}\right| < 1$ επομένως $\lim_{\ell \to \infty} \left|\frac{\lambda_i}{\lambda_j}\right|^{\ell} = 0$ οπότε η (6) γράφεται ως:

$$\frac{\partial L_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_k} \approx c_j \lambda_j^{k} V_j^{T}$$

Αν $|\lambda_j| > 1$ τότε $\frac{\partial h_t}{\partial h_e} = \lambda_j^{k}$ αυξάνεται εκθετικά γρήγορα κατά τη διεύθυνση του V_j και επομένως το gradient $\frac{\partial L_t}{\partial w}$ που είναι ίσο με $\frac{\partial L_t}{\partial o_t} \cdot \frac{\partial o_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_e} \cdot \frac{\partial h_e}{\partial w}$ πηγαίνει στο άπειρο. (exploding gradient).

Αν $|\lambda_j| < 1$ τότε $\frac{\partial h_t}{\partial h_e} = \lambda_j^{k}$ μειώνεται εκθετικά γρήγορα κατά τη διεύθυνση του V_j και επομένως το gradient τείνει να εξαφανιστεί (vanishing gradient).

$\boxed{\text{Άσκηση 2.4}}$

$$\left.\begin{aligned} \mu_1 &= \vec{w}^T \vec{\mu_1}, \quad \Sigma_1 = \sum_i (\vec{x}_i - \vec{\mu_1})(\vec{x}_i - \vec{\mu_1})^T \\ \mu_2 &= \vec{w}^T \vec{\mu_2}, \quad \Sigma_2 = \sum_i (\vec{x}_i - \vec{\mu_2})(\vec{x}_i - \vec{\mu_2})^T \end{aligned}\right\} \text{(A)}$$

$$\Rightarrow \left.\begin{aligned} \sigma_1^2 &= \sum_i (y_i - \mu_1)^2 = \sum_i (\vec{w}^T \vec{x}_i - \vec{w}^T \vec{\mu_1})^2 \\ \sigma_2^2 &= \sum_i (y_i - \mu_2)^2 = \sum_i (\vec{w}^T \vec{x}_i - \vec{w}^T \vec{\mu_2})^2 \end{aligned}\right\} \Leftrightarrow \begin{aligned} \sigma_1^2 &= \sum_i \vec{w}^T (\vec{x}_i - \vec{\mu_1})(\vec{x}_i - \vec{\mu_1})^T \vec{w} \\ \sigma_2^2 &= \sum_i \vec{w}^T (\vec{x}_i - \vec{\mu_2})(\vec{x}_i - \vec{\mu_2})^T \vec{w} \end{aligned}$$

$$\Leftrightarrow \left.\begin{aligned} \sigma_1^2 &= \vec{w} \sum_i (\vec{x}_i - \vec{\mu_1})(\vec{x}_i - \vec{\mu_2})^T \vec{w} \\ \sigma_2^2 &= \vec{w} \sum_i (\vec{x}_i - \vec{\mu_1})(\vec{x}_i - \vec{\mu_2})^T \vec{w} \end{aligned}\right\} \Leftrightarrow \begin{aligned} \sigma_1^2 &= \vec{w}^T \Sigma_1 \vec{w} \\ \sigma_2^2 &= \vec{w}^T \Sigma_2 \vec{w} \end{aligned}$$

Έτσι η Fisher LDA των προβολών καθορίζεται από

$$J_Y(\vec{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{\mathbf{S_B}^y}{S_W^y} = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$$

καθώς $S_B^Y = (\mu_1 - \mu_2)^2 = (\vec{w}^T \vec{\mu_1} - \vec{w}^T \vec{\mu_2})^2 = \vec{w}^T (\vec{\mu_1} - \vec{\mu_2})(\vec{\mu_1} - \vec{\mu_2})^T \vec{w} = \vec{w}^T S_B \vec{w}$

$S_W^Y = \sigma_1^2 + \sigma_2^2 = (\vec{w}^T \Sigma_1 \vec{w}) + (\vec{w}^T \Sigma_2 \vec{w}) = \vec{w}^T (\Sigma_1 + \Sigma_2) \vec{w} = \vec{w}^T S_W \vec{w}$

Αν S_W αντιστρέψιμο:

$$\frac{dJ_1(\vec{w})}{d\vec{w}} = \frac{d}{d\vec{w}} \left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}} \right) = \frac{(\vec{w}^T S_W \vec{w})\, 2 S_B \vec{w} - (\vec{w}^T S_B \vec{w})\, 2 S_W \vec{w}}{(\vec{w}^T S_W \vec{w})^2} = 0$$

$$\Leftrightarrow 2(\vec{w}^T S_W \vec{w}) S_B \vec{w} - 2(\vec{w}^T S_B \vec{w}) S_W \vec{w} = 0 \Leftrightarrow \vec{w}^T S_W \vec{w}\, S_B \vec{w} = \vec{w}^T S_B \vec{w}\, S_W \vec{w} \Leftrightarrow$$

$$\Leftrightarrow S_B \vec{w} = J_1(\vec{w}) S_W \vec{w} \Leftrightarrow (S_W^{-1} S_B)\vec{w} = J_1(\vec{w})\vec{w} \overset{\text{γενικευμένες}}{\underset{\text{ιδιοτιμές}}{\Leftrightarrow}} \vec{w}^* = S_W^{-1}(\vec{\mu_1} - \vec{\mu_2})$$

$$\Leftrightarrow \boxed{\vec{w}^* = (\Sigma_1 + \Sigma_2)^{-1}(\vec{\mu_1} - \vec{\mu_2})}$$ διότι αν S_W αντιστρέψιμο

$$S_W^{-1} S_B \vec{w} = \lambda \vec{w} \Leftrightarrow S_W^{-1} (\vec{\mu_1} - \vec{\mu_2}) \underbrace{(\vec{\mu_1} - \vec{\mu_2})^T \vec{w}}_{a} = \lambda \vec{w} \Leftrightarrow S_W^{-1}(\vec{\mu_1} - \vec{\mu_2}) = \frac{\lambda}{a} \vec{w} \Leftrightarrow$$

$$\Leftrightarrow \vec{w} = \frac{a}{\lambda} \cdot (\Sigma_1 + \Sigma_2)^{-1} (\vec{\mu_1} - \vec{\mu_2}) \Rightarrow \vec{w} = (\Sigma_1 + \Sigma_2)^{-1}(\vec{\mu_1} - \vec{\mu_2})$$ γιατί δεν μας ενδιαφέρει το $\|\vec{w}\|^2$

Άσκηση 2.5

1. Υπολογίζουμε $P(G=1) = P(G=1|B=1)P(B=1) + P(G=1|B=0)P(B=0)$ όπου $P(G=1|B=1)$

$= P(G=1|B=1, F=1)P(F=1) + P(G=1|B=1,F=0)P(F=0) = 0.95 \cdot 0.8 + 0.3 \cdot 0.2 = 0.82$

$P(G=1|B=0) = \ldots = 0.24$ οπότε $P(G=1) = 0.791$

- $P(D=0) = P(D=0|G=1)P(G=1) + P(D=0|G=0)P(G=0) = 0.2 \cdot 0.791 + 0.8 \cdot 0.209 = 0.3254$
- $P(G=0|F=0) = P(G=0|B=0, F=0)P(B=0|F=0) + P(G=0|B=1,F=0)P(B=1|F=0) =$

$= 0.8 \cdot 0.05 + 0.7 \cdot 0.95 = 0.705$

- $P(D=0|F=0) = 0.8 \cdot 0.705 + 0.2 \cdot 0.795 = 0.623$ άρα $P(F=0|P=0) = \ldots = 0.383$

2.

$$P(F=0|D=0, B=0) = \frac{P(F=0, D=0, B=0)}{P(D=0, B=0)} = \frac{P(D=0|F=0, B=0)P(F=0, B=0)}{P(D=0, B=0)}$$

$$= P(D=0|F=0, B=0)\frac{P(F=0|B=0)}{P(D=0|B=0)}$$

- $P(D=0|F=0, B=0) = P(D=0|G=0)P(G=0|F=0, B=0) + P(D=0|G=1)P(G=1|F=0)$

$= 0.8 \cdot 0.8 + 0.2 \cdot 0.2 = 0.68$

- $P(D=0|B=0) = P(D=0|G=0)P(G=0|B=0) + P(D=0|G=1)P(G=1|B=0)$
- $P(G=1|B=0) = P(G=1|B=0, F=0)P(F=0) + P(G=1|B=0, F=1)P(F=1)$

$= 0.2 \cdot 0.2 + 0.25 \cdot 0.8 = 0.24$

οπότε $P(D=0|B=0) = 0.8 \cdot 0.76 + 0.2 \cdot 0.24 = 0.656$

$P(F=0|D=0, B=0) = 0,207$

Δηλαδή $P(F=0|D=0, B=0) < P(F=0|D=0) = 0,383$ διότι στην $P(F=0|D=0)$ έχουμε λάβει υπόψη τη πιθανότητα $B=0$ ή $B=1$ μαζί με τη παρατήρηση του G