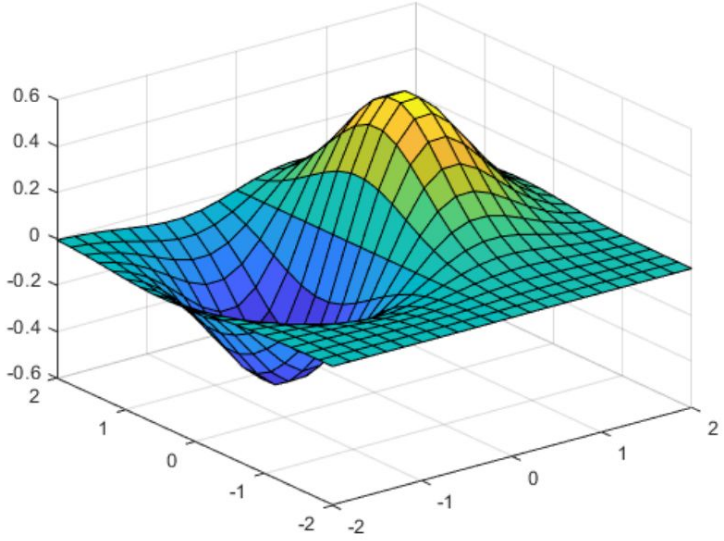
Jak zefektivnit učení NN

Aktivační funkce Softmax



Aktivační funkce Softmax

- Alternativní způsob jak zabránit zpomalení učení je změnou aktivační funkce (v poslední vrstvě)

Aktivační funkce Softmax

- Alternativní způsob jak zabránit zpomalení učení je změnou aktivační funkce (v poslední vrstvě)
- Zároveň chceme definovat nějakou aktivační funkci, která nám umožní pravděpodobnostní interpretaci (obor hodnot (0,1), součet přes všechny neurony ve vrstvy = 1)

Aktivační funkce Softmax

- Alternativní způsob jak zabránit zpomalení učení je změnou aktivační funkce (v poslední vrstvě)
- Zároveň chceme definovat nějakou aktivační funkci, která nám umožní pravděpodobnostní interpretaci (obor hodnot (0,1), součet přes všechny neurony ve vrstvy = 1)

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$$

Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$

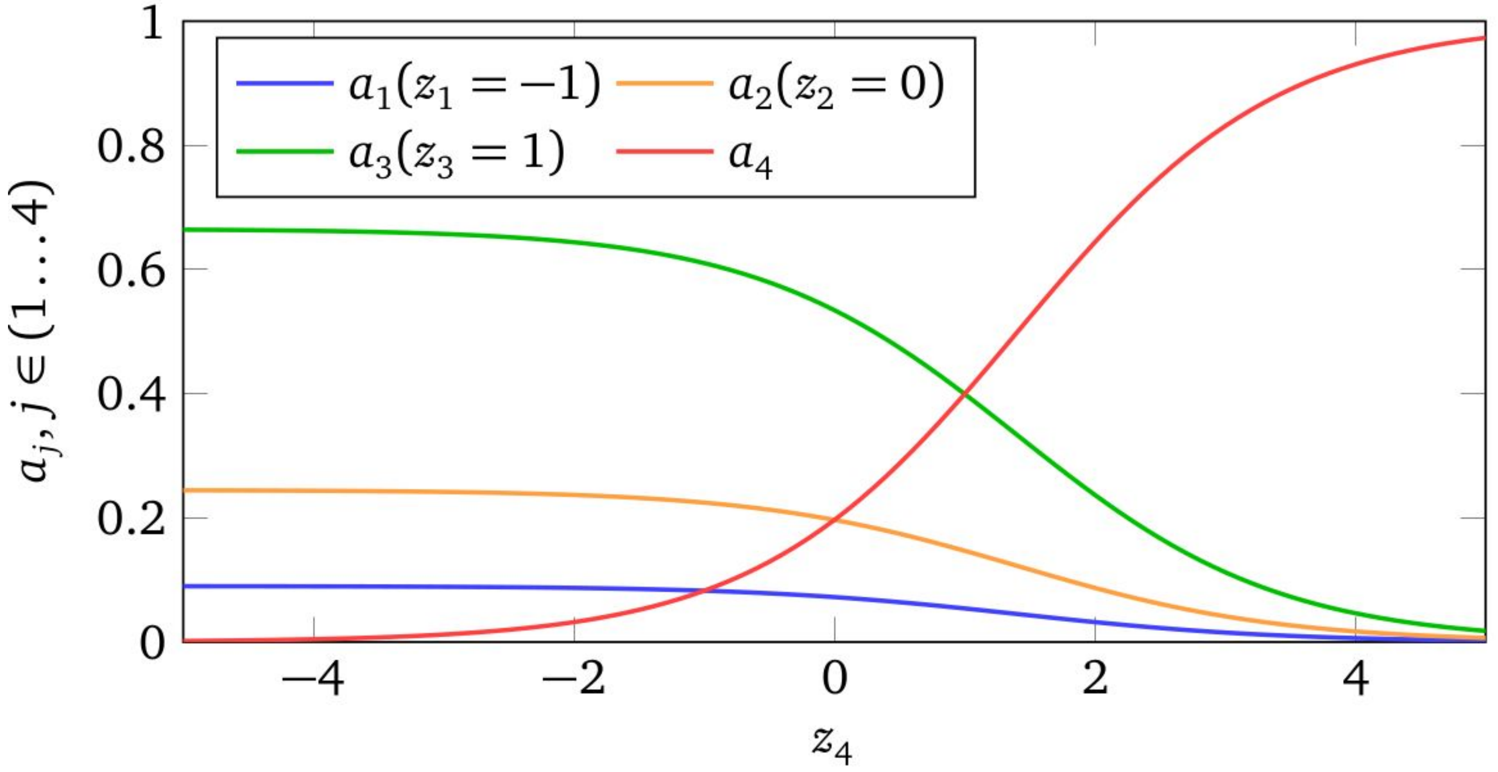
Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$
- Změna jedné aktivace vede k proporční změně ve všech ostatních aktivacích

Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$
- Změna jedné aktivace vede k proporční změně ve všech ostatních aktivacích
- např. máme aktivace 4 neuronů a1, a2, a3, a4. Když se aktivace a1 zvýší o 0.3, potom aktivace a2, a3 a a4 se sníží o 0.1

Aktivační funkce Softmax



Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$
- Změna jedné aktivace vede k proporční změně ve všech ostatních aktivacích
- např. máme aktivace 4 neuronů a1, a2, a3, a4. Když se aktivace a1 zvýší o 0.3, potom aktivace a2, a3 a a4 se sníží o 0.1

$$\sum_j a_j^L = \sum_j \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{\sum_j e^{z_j^L}}{\sum_k e^{z_k^L}} = 1$$

Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$
- Změna jedné aktivace vede k proporční změně ve všech ostatních aktivacích
- např. máme aktivace 4 neuronů a1, a2, a3, a4. Když se aktivace a1 zvýší o 0.3, potom aktivace a2, a3 a a4 se sníží o 0.1

$$\sum_j a_j^L = \sum_j \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{\sum_j e^{z_j^L}}{\sum_k e^{z_k^L}} = 1$$

- Output tedy můžeme chápat jako pravděpodobnostní rozdělení

Aktivační funkce Softmax

- Proč zrovna takový tvar? $a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$
- Změna jedné aktivace vede k proporční změně ve všech ostatních aktivacích
- např. máme aktivace 4 neuronů a1, a2, a3, a4. Když se aktivace a1 zvýší o 0.3, potom aktivace a2, a3 a a4 se sníží o 0.1

$$\sum_j a_j^L = \sum_j \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{\sum_j e^{z_j^L}}{\sum_k e^{z_k^L}} = 1$$

- Output tedy můžeme chápat jako pravděpodobnostní rozdělení - softmax vlastně přeškáluje Zj a "splácne" je, aby tvořily toto rozdělení

Aktivační funkce Softmax

- To se hodí, můžeme například sledovat, jak jistá si síť je svou predikcí (u špatně zařazených číslic často uvidíme, že síť se “rozhodovala” mezi dvěma číslicemi a zvolila tu nesprávnou)

Aktivační funkce Softmax - 1. cvičení

- Dokažte, že vrstva se sigmoidem netvoří pavděpodobnostní rozdělení (tj. že součet aktivací se nerovná 1).

Aktivační funkce Softmax - 1. cvičení

- Dokažte, že vrstva se sigmoidem netvoří pavděpodobnostní rozdělení (tj. že součet aktivací se nerovná 1).
- Stačí si představit síť s jedním vstupním a výstupním neuronem. Protože sigmoid(x)<1, je jediná aktivace výstupní vrstvy <1 a tedy i součet přes všechny (jeden) neurony výstupní vrstvy.

Aktivační funkce Softmax - 1. cvičení

- Dokažte, že vrstva se sigmoidem netvoří pavděpodobnostní rozdělení (tj. že součet aktivací se nerovná 1).
- Stačí si představit síť s jedním vstupním a výstupním neuronem. Protože sigmoid(x)<1, je jediná aktivace výstupní vrstvy <1 a tedy i součet přes všechny (jeden) neurony výstupní vrstvy.
- Podobně můžeme zkonstruovat síť, ve které bude součet outputů >1.

Aktivační funkce Softmax - 2. cvičení

- Dokažte, že $\frac{\partial a_j^L}{\partial z_k^L} > 0 \iff j = k$ a $\frac{\partial a_j^L}{\partial z_k^L} < 0 \iff j \neq k$, tj. když roste (klesá) jedna aktivace, ostatní klesají (rostou)

Aktivační funkce Softmax - 2. cvičení

- Dokažte, že $\frac{\partial a_j^L}{\partial z_k^L} > 0 \iff j = k$ a $\frac{\partial a_j^L}{\partial z_k^L} < 0 \iff j \neq k$, tj. když roste (klesá) jedna aktivace, ostatní klesají (rostou)

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{e^{z_j^L}}{e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}}$$

Aktivační funkce Softmax - 2. cvičení

- Dokažte, že $\frac{\partial a_j^L}{\partial z_k^L} > 0 \iff j = k$ a $\frac{\partial a_j^L}{\partial z_k^L} < 0 \iff j \neq k$, tj. když roste (klesá) jedna aktivace, ostatní klesají (rostou)

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{e^{z_j^L}}{e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}}$$

- $j = k:$ $$\frac{\partial a_j^L}{\partial z_j^L} = \frac{e^{z_j^L} \cdot \left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right) - e^{z_j^L} \cdot e^{z_j^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)^2} = \frac{e^{z_j^L} \cdot \sum_{k \neq j} e^{z_k^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)^2} > 0$$

Aktivační funkce Softmax - 2. cvičení

- Dokažte, že $\frac{\partial a_j^L}{\partial z_k^L} > 0 \iff j = k$ a $\frac{\partial a_j^L}{\partial z_k^L} < 0 \iff j \neq k$, tj. když roste (klesá) jedna aktivace, ostatní klesají (rostou)

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}} = \frac{e^{z_j^L}}{e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}}$$

- $j = k:$ $$\frac{\partial a_j^L}{\partial z_j^L} = \frac{e^{z_j^L} \cdot \left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right) - e^{z_j^L} \cdot e^{z_j^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)^2} = \frac{e^{z_j^L} \cdot \sum_{k \neq j} e^{z_k^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)^2} > 0$$

- $j \neq k:$ $$\frac{\partial a_j^L}{\partial z_k^L} = \frac{\partial}{\partial z_k^L} \frac{e^{z_j^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)} = -\frac{e^{z_j^L} e^{z_k^L}}{\left(e^{z_j^L} + \sum_{k \neq j} e^{z_k^L}\right)^2} < 0$$

Aktivační funkce Softmax - 3. cvičení

- V případě sigmoidu platí $a_j^L = \sigma(z_j^L)$, tj. aktivace je funkcí pouze příslušné hodnoty z. Vysvětlete, proč to v případě softmaxu neplatí.

Aktivační funkce Softmax - 3. cvičení

- V případě sigmoidu platí $a_j^L = \sigma(z_j^L)$, tj. aktivace je funkcí pouze příslušné hodnoty z. Vysvětlete, proč to v případě softmaxu neplatí.
- Evidentní, jako důkaz lze brát to, že parciální derivace podle Zk, k!=j je nenulová.

Aktivační funkce Softmax - 1. problém

- Inverzní funkce k softmaxu: ukažte, že pro známé aktivace a_j^L poslední vrstvy platí $z_j^L = \ln\left(a_j^L\right) + C$, přičemž C je konstanta nezávislá na j.

Aktivační funkce Softmax - 1. problém

- Inverzní funkce k softmaxu: ukažte, že pro známé aktivace a_j^L poslední vrstvy platí $z_j^L = \ln\left(a_j^L\right) + C$, přičemž C je konstanta nezávislá na j.
- Vyjdeme z definice:

Aktivační funkce Softmax - 1. problém

- Inverzní funkce k softmaxu: ukažte, že pro známé aktivace a_j^L poslední vrstvy platí $z_j^L = \ln\left(a_j^L\right) + C$, přičemž C je konstanta nezávislá na j.
- Vyjdeme z definice:

$$a_j^L = \frac{e^{z_j^L}}{\sum_k e^{z_k^L}}$$

$$e^{z_j^L} = a_j^L \cdot \sum_k e^{z_k^L}$$

$$z_j^L = \ln\left(a_j^L \cdot \sum_k e^{z_k^L}\right) = \ln\left(a_j^L\right) + \ln\left(\sum_k e^{z_k^L}\right), \text{ přičemž } \ln\left(\sum_k e^{z_k^L}\right) \text{ nezávisí}$$

na j (je pro každé z_j^L stejné), tedy je naší hledanou konstantou C

Aktivační funkce Softmax - jak řeší problém zpomalení učení

- definujeme účelovou funkci ***log-likelihood*** (logaritmická věrohodnostní funkce) $C_x = -\ln\left(a_y^L\right)$

Aktivační funkce Softmax - jak řeší problém zpomalení učení

- definujeme účelovou funkci ***log-likelihood*** (logaritmická věrohodnostní funkce) $C_x = -\ln\left(a_y^L\right)$
- skutečně se chová jako účelová funkce: pokud si je síť jistá predikcí a ta je správná, $a_y^L \to 1$ a tedy $C = -\ln\left(a_y^L\right) \to 0$. Naopak pokud si síť jistá nebude (predikuje např. jinou hodnotu), $a_y^L \to 0$ a tedy $C = -\ln\left(a_y^L\right) \to +\infty$

Aktivační funkce Softmax - jak řeší problém zpomalení učení

- definujeme účelovou funkci ***log-likelihood*** (logaritmická věrohodnostní funkce) $C_x = -\ln\left(a_y^L\right)$
- skutečně se chová jako účelová funkce: pokud si je síť jistá predikcí a ta je správná, $a_y^L \to 1$ a tedy $C = -\ln\left(a_y^L\right) \to 0$. Naopak pokud si síť jistá nebude (predikuje např. jinou hodnotu), $a_y^L \to 0$ a tedy $C = -\ln\left(a_y^L\right) \to +\infty$
- za pomalé učení mohly parciální derivace C podle vah a biasů

Aktivační funkce Softmax - jak řeší problém zpomalení učení

- definujeme účelovou funkci ***log-likelihood*** (logaritmická věrohodnostní funkce) $C_x = -\ln\left(a_y^L\right)$
- skutečně se chová jako účelová funkce: pokud si je síť jistá predikcí a ta je správná, $a_y^L \to 1$ a tedy $C = -\ln\left(a_y^L\right) \to 0$. Naopak pokud si síť jistá nebude (predikuje např. jinou hodnotu), $a_y^L \to 0$ a tedy $C = -\ln\left(a_y^L\right) \to +\infty$
- za pomalé učení mohly parciální derivace C podle vah a biasů
- v případě softmaxu a log-likelihoodu dopadnou stejně, jako u cross entropie a sigmoidu

Aktivační funkce Softmax - 2. problém

- Spočtěte parciální derivace log-likelihood účelové funkce se softmax výstupní vrstvou (respektive ukažte, že vyjdou stejně jako v případě cross entropie a sigmoidu)

Cross Entropy: problém

- **Rovnice backpropagation (pro MSE)**

1. $\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'\left(z_j^L\right)$ zapíšeme jako $\nabla_a C \odot \sigma'\left(z^L\right)$

2. $\delta^l = \left(\left(w^{l+1}\right)^T \delta^{l+1}\right) \odot \sigma'\left(z^l\right)$, po prvcích $\delta_j^l = \sum_{k=1}^{n_{l+1}} w_{kj}^{l+1}\delta_k^{l+1}\sigma'\left(z_j^l\right)$

3. $\frac{\partial C}{\partial b_j^l} = \delta_j^l$

4. $\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$

Aktivační funkce Softmax - 2. problém

- Spočtěte parciální derivace log-likelihood účelové funkce se softmax výstupní vrstvou
- Využijeme 3. rovnici backpropagation. Co zkusit i 1.?

Backpropagation - 4 rovnice algoritmu: důkaz (cvičení)

- **Důkaz 3. :** $\dfrac{\partial C}{\partial b_j^l} = \delta_j^l$

- Vyjdeme z definice chyby neuronu a použijeme derivaci složené funkce (rozvineme závislost na biasu)

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_{k=1}^{n_{l+1}} \frac{\partial C}{\partial b_k^l} \frac{\partial b_k^l}{\partial z_j^l} = \frac{\partial C}{\partial b_j^l} \frac{\partial b_j^l}{\partial z_j^l}, \text{protože } b_k^l \text{ závisí jen na } z_k^l, \text{ neboli } \frac{\partial b_k^l}{\partial z_j^l} = 0 \text{ když j} \neq \text{k}$$

dále z definice $z_j^l = \sum_{k=1}^{n_l} w_{kj}^l a_k^{l-1} + b_j^l \iff b_j^l = z_j^l - \sum_{k=1}^{n_l} w_{kj}^l a_k^{l-1}$, zderivováním získáme $\dfrac{\partial b_j^l}{z_j^l} = 1$

což dosadíme do předchozí rovnice: $$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_{k=1}^{n_{l+1}} \frac{\partial C}{\partial b_k^l} \frac{\partial b_k^l}{\partial z_j^l} = \frac{\partial C}{\partial b_j^l} \frac{\partial b_j^l}{\partial z_j^l} = \frac{\partial C}{\partial b_j^l}$$

Backpropagation - 4 rovnice algoritmu: důkaz

- **Důkaz 1. :** $\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'\left(z_j^L\right)$
- Vyjdeme z definice a aplikujeme derivaci složené funkce:

$$\delta_j^L \equiv \frac{\partial C}{\partial z_j^L} = \sum_{k=1}^{n_L} \frac{\partial C}{\partial a_k^L}\frac{\partial a_k^L}{\partial z_j^L}$$

přitom aktivace k-tého neuronu poslední vrstvy závisí pouze na $z_k^L \implies \frac{\partial a_k^L}{\partial z_{j^L}} = 0 \text{ pokud } j \neq k$

Dohromady tedy $\delta_j^L \equiv \frac{\partial C}{\partial z_j^L} = \sum_{k=1}^{n_L} \frac{\partial C}{\partial a_k^L}\frac{\partial a_k^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\frac{\partial a_j^L}{\partial z_j^L}$

Dále víme, že $a_j^L = \sigma\left(z_j^L\right)$, což můžeme dosadit do předchozího vztahu

$$\delta_j^L \equiv \frac{\partial C}{\partial z_j^L} = \sum_{k=1}^{n_L} \frac{\partial C}{\partial a_k^L}\frac{\partial a_k^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\frac{\partial a_j^L}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L}\frac{\partial \sigma}{\partial z_j^L}\left(z_j^L\right) = \frac{\partial C}{\partial a_j^L}\sigma'\left(z_j^L\right)$$

Aktivační funkce Softmax - 2. problém

- Spočtěte parciální derivace log-likelihood účelové funkce se softmax výstupní vrstvou
- Využijeme 3. rovnici backpropagation
- V důkazu 1. rovnice používáme tvar aktivační funkce, důkaz neplatí
- V důkazu 3. rovnice se nikde nevyskytuje konkrétní podoba účelové ani aktivační funkce, rovnice platí

Aktivační funkce Softmax - 2. problém

- Spočtěte parciální derivace log-likelihood účelové funkce se softmax výstupní vrstvou
- Využijeme 3. rovnici backpropagation
- V důkazu 1. rovnice používáme tvar aktivační funkce, důkaz neplatí
- V důkazu 3. rovnice se nikde nevyskytuje konkrétní podoba účelové ani aktivační funkce, rovnice platí
- Vyjdeme ze 3. rovnice a definice chyby neuronu:

Aktivační funkce Softmax - 2. problém

$$C_x = -\ln\left(a_y^L\right)$$

- Spočtěte parciální derivace log-likelihood účelové funkce se softmax výstupní vrstvou
- Využijeme 3. rovnici backpropagation
- V důkazu 1. rovnice používáme tvar aktivační funkce, důkaz neplatí
- V důkazu 3. rovnice se nikde nevyskytuje konkrétní podoba účelové ani aktivační funkce, rovnice platí
- Vyjdeme ze 3. rovnice a definice chyby neuronu:

$$\frac{\partial C}{\partial b_j^L} = \delta_j^L = \frac{\partial C}{\partial z_j^L} = \sum_k \frac{\partial C}{\partial a_k^L}\frac{\partial a_k^L}{\partial z_j^L} = \frac{\partial C}{\partial a_y^L}\frac{\partial a_y^L}{\partial z_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

Aktivační funkce Softmax - 2. problém

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

- Musíme rozlišit 2 případy:
  - $y = j$

Aktivační funkce Softmax - 2. problém

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

- Musíme rozlišit 2 případy:
  - $y = j$

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L} = -\frac{1}{a_y^L}\frac{e^{z_j^L}\cdot\sum_{k\neq j}e^{z_k^L}}{\left(e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}\right)^2} = -\frac{e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}}{e^{z_j^L}}\frac{e^{z_j^L}\cdot\sum_{k\neq j}e^{z_k^L}}{\left(e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}\right)^2} =$$

$$= -\frac{\sum_{k\neq j}e^{z_k^L}}{e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}} = -\frac{\sum_{k\neq j}e^{z_k^L}+e^{z_j^L}-e^{z_j^L}}{e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}} = -1+\frac{e^{z_j^L}}{e^{z_j^L}+\sum_{k\neq j}e^{z_k^L}} = a_j^L - 1 = a_j^L - \tilde{y}_j$$

Aktivační funkce Softmax - 2. problém

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

- Musíme rozlišit 2 případy:
  - $y \neq j$

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L} = -\frac{1}{a_y^L}\cdot\left(-\frac{e^{z_j^L}e^{z_y^L}}{\left(e^{z_y^L}+\sum_{k\neq y}e^{z_k^L}\right)^2}\right) = \frac{e^{z_y^L}+\sum_{k\neq y}e^{z_k^L}}{e^{z_y^L}}\frac{e^{z_j^L}e^{z_y^L}}{\left(e^{z_j^L}+\sum_{k\neq y}e^{z_k^L}\right)^2} =$$

$$= \frac{e^{z_j^L}}{\left(e^{z_y^L}+\sum_{k\neq y}e^{z_k^L}\right)} = \frac{e^{z_j^L}}{\left(\sum_k e^{z_k^L}\right)} = a_j^L = a_j^L - 0 = a_j^L - \tilde{y}_j$$

Aktivační funkce Softmax - 2. problém

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

- Musíme rozlišit 2 případy:
  - $y \neq j$

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L} = -\frac{1}{a_y^L} \cdot \left(-\frac{e^{z_j^L} e^{z_y^L}}{\left(e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}\right)^2} \right) = \frac{e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}}{e^{z_y^L}} \frac{e^{z_j^L} e^{z_y^L}}{\left(e^{z_j^L} + \sum_{k \neq y} e^{z_k^L}\right)^2} =$$

$$= \frac{e^{z_j^L}}{\left(e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}\right)} = \frac{e^{z_j^L}}{\left(\sum_k e^{z_k^L}\right)} = a_j^L = a_j^L - 0 = a_j^L - \tilde{y}_j$$

- Dokázali jsme tedy, že platí $\frac{\partial C}{\partial b_j^L} = a_j^L - \tilde{y_j}$

Aktivační funkce Softmax - 2. problém

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L}$$

- Musíme rozlišit 2 případy:
  - $y \neq j$

$$\frac{\partial C}{\partial b_j^L} = -\frac{1}{a_y^L}\frac{\partial a_y^L}{\partial z_j^L} = -\frac{1}{a_y^L} \cdot \left(-\frac{e^{z_j^L} e^{z_y^L}}{\left(e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}\right)^2} \right) = \frac{e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}}{e^{z_y^L}} \frac{e^{z_j^L} e^{z_y^L}}{\left(e^{z_j^L} + \sum_{k \neq y} e^{z_k^L}\right)^2} =$$

$$= \frac{e^{z_j^L}}{\left(e^{z_y^L} + \sum_{k \neq y} e^{z_k^L}\right)} = \frac{e^{z_j^L}}{\left(\sum_k e^{z_k^L}\right)} = a_j^L = a_j^L - 0 = a_j^L - \tilde{y}_j$$

- Dokázali jsme tedy, že platí $\frac{\partial C}{\partial b_j^L} = a_j^L - \tilde{y_j}$
- Pro derivaci podle váhy obdobný postup, ale vyjdeme z $\frac{\partial C}{\partial w_{jk}^L} = a_k^{L-1}\delta_j^L$

Aktivační funkce Softmax - 2. problém

- Pro derivaci podle váhy obdobný postup, ale vyjdeme z $\frac{\partial C}{\partial w_{jk}^L} = a_k^{L-1}\delta_j^L$ tj. 4. rovnice backpropagation. Získáme platnost rovnosti

$$\frac{\partial C}{\partial w_{jk}^L} = a_k^{L-1}(a_j^L - y_j)$$

Aktivační funkce Softmax - 3. problém

- Proč název “softmax”: Definujme $a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}}, c > 0$, tj. pro c=1 je to náš standardní softmax.
  Prozkoumejte, co se stane s aktivací, pokud $c \to +\infty$ a dokažte, že taková funkce též tvoří pravděpodobnostní rozdělení.

Aktivační funkce Softmax - 3. problém

- Proč název "softmax": Definujme $a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}}, c > 0$, tj. pro c=1 je to náš standardní softmax.
- Prozkoumejte, co se stane s aktivací, pokud $c \to +\infty$ a dokažte, že taková funkce též tvoří pravděpodobnostní rozdělení.
  - evidentně stále platí $a_j^L = 1$
  - a podobně $\sum_j a_j^L = \sum_j \frac{e^{cz_j^L}}{\sum_j e^{cz_j^L}} = 1$
  - zároveň jsou aktivace <=1

Aktivační funkce Softmax - 3. problém

- Proč název “softmax”: Definujme $a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}}, c > 0$, tj. pro c=1 je to náš standardní softmax.
- Prozkoumejte, co se stane s aktivací, pokud $c \to +\infty$ a dokažte, že taková funkce též tvoří pravděpodobnostní rozdělení.
  - evidentně stále platí $a_j^L = 1$
  - a podobně $\sum_j a_j^L = \sum_j \frac{e^{cz_j^L}}{\sum_j e^{cz_j^L}} = 1$
  - zároveň jsou aktivace <=1
  - jedná se tedy o pravděpodobnostní rozdělení

Aktivační funkce Softmax - 3. problém

- Pro $c \to +\infty$:

$$a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}} = \frac{1}{e^{-cz_j^L} \sum_k e^{cz_k^L}} = \frac{1}{\sum_k e^{c\left(z_k^L - z_j^L\right)}}$$

Aktivační funkce Softmax - 3. problém

- Pro $c \to +\infty$:

$$a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}} = \frac{1}{e^{-cz_j^L} \sum_k e^{cz_k^L}} = \frac{1}{\sum_k e^{c\left(z_k^L - z_j^L\right)}}$$

- $z_j^L \neq \max\left(z_k^L\right) \implies \sum_k e^{c\left(z_k^L - z_j^L\right)} \to +\infty \implies a_j^L \to 0$

Aktivační funkce Softmax - 3. problém

- Pro $c \to +\infty$:

$$a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}} = \frac{1}{e^{-cz_j^L} \sum_k e^{cz_k^L}} = \frac{1}{\sum_k e^{c\left(z_k^L - z_j^L\right)}}$$

- $z_j^L \neq \max\left(z_k^L\right) \implies \sum_k e^{c\left(z_k^L - z_j^L\right)} \to +\infty \implies a_j^L \to 0$
- $z_j^L = \max\left(z_k^L\right)$ a n ≥ 1 z_i^L nabývá této hodnoty: $\implies \sum_k e^{c\left(z_k^L - z_j^L\right)} \to +\infty \implies a_j^L \to 0$
  - pro $k;\ z_k^L = z_j^L:\ e^{c\left(z_k^L - z_j^L\right)} = 1 \implies \lim_{c \to +\infty} e^{c\left(z_k^L - z_j^L\right)} = 1$
  - pro $k;\ z_k^L < z_j^L:\ z_k^L - z_j^L < 0 \implies \lim_{c \to +\infty} e^{c\left(z_k^L - z_j^L\right)} = 0$

Aktivační funkce Softmax - 3. problém

- Pro $c \to +\infty$:

$$a_j^L = \frac{e^{cz_j^L}}{\sum_k e^{cz_k^L}} = \frac{1}{e^{-cz_j^L} \sum_k e^{cz_k^L}} = \frac{1}{\sum_k e^{c\left(z_k^L - z_j^L\right)}}$$

- $z_j^L \neq \max\left(z_k^L\right) \implies \sum_k e^{c\left(z_k^L - z_j^L\right)} \to +\infty \implies a_j^L \to 0$
- $z_j^L = \max\left(z_k^L\right)$ a n ≥ 1 z_i^L nabývá této hodnoty:
  - pro k; $z_k^L = z_j^L$: $e^{c\left(z_k^L - z_j^L\right)} = 1 \implies \lim\limits_{c \to +\infty} e^{c\left(z_k^L - z_j^L\right)} = 1$
  - pro k; $z_k^L < z_j^L$: $z_k^L - z_j^L < 0 \implies \lim\limits_{c \to +\infty} e^{c\left(z_k^L - z_j^L\right)} = 0$

- dohromady tedy $\lim\limits_{c \to +\infty} \sum_k e^{c\left(z_k^L - z_j^L\right)} = n$ a $\lim\limits_{c \to +\infty} a_j^L = \frac{1}{n}$

Aktivační funkce Softmax - 3. problém

- Tedy čím vyšší c, tím větší váhu má největší hodnota, přestože v potaz se berou všechny

Aktivační funkce Softmax - 4. problém

- Backpropagation a log-likelihood: dokažte, že pro chybu neurony výstupní vrstvy platí

$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = a_j^L - \tilde{y_j}$$

Aktivační funkce Softmax - 4. problém

- Backpropagation a log-likelihood: dokažte, že pro chybu neurony výstupní vrstvy platí
$$\delta_j^L = \frac{\partial C}{\partial z_j^L} = a_j^L - \tilde{y_j}$$
- To jsme naštěstí dokázali už ve 2. problému :-)