Uvod u obradu prirodnog jezika

8.1. Logistička regresija

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Uvod

Cilj logističke regresije

- učenje klasifikatora koji može donijeti binarnu odluku za klasu nekog novog ulaznog promatranja

Komponente logističke regresije

- Skup za treniranje od m promatranja
  - promatranje je par ulaza i izlaza $(x^{(i)}, y^{(i)})$
- Strojno učenje za klasifikaciju ima sljedeće komponente:
  1. **Reprezentacija osobina** za ulaze
     - svaki ulaz $x^{(i)}$ je predstavljen vektorom osobina $[x_1, x_2, \ldots, x_n]$
     - osobina i za ulaz $x^{(j)}$ je $x_i^{(j)}$ (ili f_i ili $f_i(x)$)
  2. **Funkcija klasifikacije** koja računa $\hat{y}$ - procjenu klase pomoću $p(y|x)$
     - sigmoid, softmax, …
  3. **Aktivacijska funkcija** za učenje koja obično uključuje minimizaciju greške
     - Unakrsna entropija gubitka
  4. **Algoritam za optimizaciju** aktivacijske funkcije:
     - stohastičko opadanje gradijenta

Faze logističke regresije

1. Učenje (treniranje)
   - pomoću stohastičkog opadanja gradijenta i gubitka unakrsne entropije
2. Testiranje
   - Za dani testni primjer x računa se $p(y|x)$ i vraća klasa s većom vjerojatnošću $y = 1$ ili $y = 0$

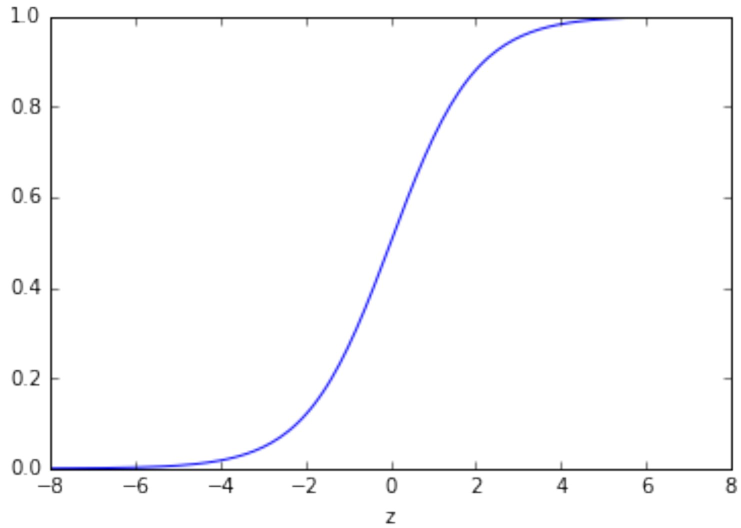
Klasifikacija

Uči se **vektor težina** $w = [w_1\ w_2 \ldots w_n]$ i **pristranost** b

- Težina w_i govori koliko je osobina x_i bitna za odluku
- $z = \sum_{i=1}^{n} w_1 x_i + b$ - težinska suma
- $z = w \cdot x + b$ gdje je $\cdot$ skalarni produkt

Težinska suma z je realni broj iz $\langle -\infty, +\infty \rangle$
kojeg treba prebaciti u vjerojatnostni prostor $[0, 1]$

- Sigmoid – logistička funkcija $y = \sigma(z) = \frac{1}{1+e^{-z}}$



Klasifikacija

Sigmoid klasifikator

- x promatranje (ulaz)
- $[x_1\ x_2\ \ldots\ x_n]$ vektor osobina za x
- $y = 1$ ili $y = 0$ klasa (izlaz)

Želimo izračunati $\mathrm{p}(y = 1|x)$

Primjer: Odluka "pozitivan sentiment" ili "negativan sentiment" za osobinu koja prebrojava riječi u dokumentu:

- $\mathrm{p}(y = 1|x)$ je vjerojatnost da je dokument "pozitivan"
- $\mathrm{p}(y = 0|x)$ je vjerojatnost da je dokument "negativan"

Klasifikacija

Izračun vjerojatnosti:

$$\mathrm{p}(y = 1|x) = \partial(w \cdot x + b) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

$$\mathrm{p}(y = 0|x) = 1 - \mathrm{p}(y = 1|x) = 1 - \frac{1}{1 + e^{-(w \cdot x + b)}} = \frac{e^{-(w \cdot x + b)}}{1 + e^{-(w \cdot x + b)}}$$

Kako odlučiti?

Neka je 0.5 **granična vrijednost**

$$\hat{y} = \begin{cases} 1 \text{ ako } p(y = 1|x) > 0.5 \\ 0 \text{ u suprotnom} \end{cases}$$

Klasifikacija

Primjer: Binarna klasifikacija sentimenta kritike filma

Je li kritika pozitivna (+) ili negativna (–)

Osobine promatranja dokumenta d

Osobina	Definicija	Vrijednost
x_1	$broj(\text{pozitivni leksikon}) \in d$	3
x_2	$broj(\text{negativni leksikon}) \in d$	2
x_3	$\begin{cases} 1, & \text{ako "}ne\text{"} \in d \\ 0, & \text{inaće} \end{cases}$	1
x_4	$broj(\text{zamjenice prvog i drugog lica} \in d)$	3
x_5	$\begin{cases} 1, & \text{ako "! "} \in d \\ 0, & \text{inaće} \end{cases}$	0
x_6	$\log(\text{broj riječi od } d)$	$\ln(64) = 4.15$

Klasifikacija

Osobina	Definicija	Vrijednost
x_1	$broj(\text{pozitivni leksikon}) \in d$	3
x_2	$broj(\text{negativni leksikon}) \in d$	2
x_3	$\begin{cases} 1, & \text{ako "}ne\text{"} \in d \\ 0, & \text{inače} \end{cases}$	1
x_4	$broj(\text{zamjenice prvog i drugog lica} \in d)$	3
x_5	$\begin{cases} 1, & \text{ako "!"} \in d \\ 0, & \text{inače} \end{cases}$	0
x_6	$\log(\text{broj riječi od } d)$	$\ln(64) = 4.15$

x_2=2 x_3=1 x_1=3

Sve je isfolirano. Gotovo nema iznenađenja, a scenarij je drugorazredan. Pa zašto je onda bio užitak gledati?
Kao prvo, glumci su sjajni. Još jedna dobra stvar je glazba. Prevladao me nagon da se maknem s kauča i počnem plesati. Uvuklo me potpunosti, a i vas će.

x_4=3 x_5=0 x_6=4.15

Klasifikacija

Osobina	Definicija	Vrijednost
x_1	$broj$(pozitivni leksikon) $\in d$	3
x_2	$broj$(negativni leksikon) $\in d$	2
x_3	$\begin{cases} 1, & \text{ako "}ne\text{"} \in d \\ 0, & \text{inaće} \end{cases}$	1
x_4	$broj$(zamjenice prvog i drugog lica $\in d$)	3
x_5	$\begin{cases} 1, & \text{ako "! "} \in d \\ 0, & \text{inaće} \end{cases}$	0
x_6	log(broj riječi od d)	$\ln(64) = 4.15$

- Težinski vektor $w = [2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$
- Pristranost $b = 0.1$
- $\mathrm{p}(+|x) = \partial(w \cdot x + b) =$
  $$= \partial([2.5, -5.0, -1.2, 0.5, 2.0, 0.7] \cdot [3, 2, 1, 3, 0, 4.15] + 0.1)$$
  $$= \partial(1.805)$$
  $$= 0.86$$
- $\mathrm{p}(-|x) = 1 - p(+|x)$
  $$= 0.14$$

Učenje

Kako se uče parametri modela w i b?

- Želimo da $\hat{y}$ bude što bliži stvarnom y
- Odnosno da udaljenost između $\hat{y}$ i y bude što manja

Funkcija gubitka $L(\hat{y}, y)$ = koliko mnogo se $\hat{y}$ razlikuje od y

- Primjer funkcije gubitka je srednja vrijednost kvadrata (mean square error)
- $L_{\mathrm{MSE}}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$
- teško za optimizirati jer nije konveksna

Učenje

Procjena uvjetne maksimalne izglednosti:

- biramo w i b koji **maksimiziraju** $\log$ **vjerojatnost stvarnih vrijednosti** od y podataka za učenje
- dobivena funkcija gubitka je **unakrsna entropija gubitka (cross-entropy loss)**

Želimo naučiti težine koje maksimiziraju vjerojatnost točne klase za $p(y|x)$

- Imamo dvije klase (1 ili 0)
- Bernoullijeva distribucija $p(y|x) = \hat{y}^y(1-\hat{y})^{1-y}$
  - za $y = 1, p(y = 1|x) = \hat{y}$
  - za $y = 0, p(y = 0|x) = 1 - \hat{y}$

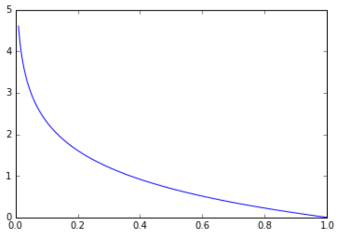
Učenje

Maksimizacija od $p(y|x)$ je isto što i maksimizacija od $\log(p(y|x))$

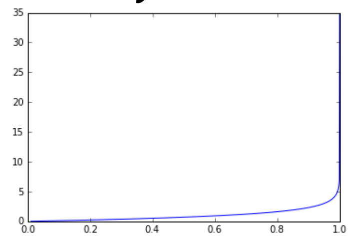
$$\log\big(p(y|x)\big) = \log(\hat{y}^{y}(1-\hat{y})^{1-y})$$
$$= y\log(\hat{y}) + (1-y)\log(1-\hat{y})$$

Funkcija gubitka se minimizira, stoga

y=1



y=0



$$L_{\mathrm{CE}}(\hat{y}, y) = -\log\big(p(y|x)\big)$$
$$= -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

Proširujemo na cijeli skup za učenje $\left\{\left(x^{(i)}, y^{(i)}\right) \mid i \in \{1, \dots, m\}\right\}$

$$\log\left(p\left(\left\{\left(x^{(i)}, y^{(i)}\right) \mid i \in \{1, \dots, m\}\right\}\right)\right) = \log\left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)})\right) =$$
$$= \sum_{i=1}^{m} \log(p(y^{(i)}|x^{(i)}))$$
$$= -\sum_{i=1}^{m} L_{CE}(\hat{y}^{(i)}, y^{(i)})$$

Učenje

Funkcija gubitka na cijelom skupu za učenje

$$cost(w,b) = \frac{1}{m}\sum_{i=1}^{m} L_{CE}\left(\hat{y}^{(i)}, y^{(i)}\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\left(\sigma\left(w \cdot x^{(i)} + b\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(w \cdot x^{(i)} + b\right)\right)$$

Za ovu funkciju je potrebno pronaći minimum

Opadanje gradijenta

Neka su θ parametri po kojima se minimizira

$\theta = (w, b)$ kod logističke regresije

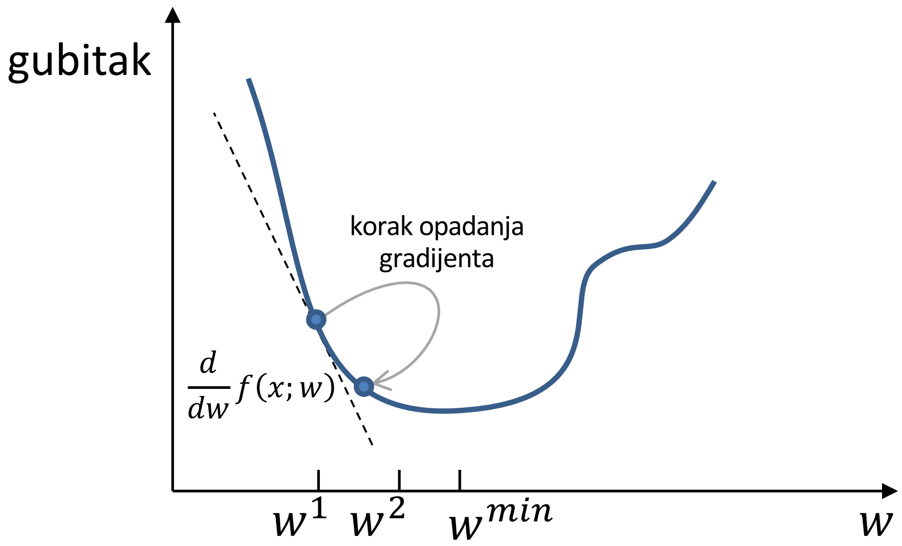
$$\hat{\theta} = \mathrm{argmin}_{\theta} \frac{1}{m} \sum_{i=1}^{m} L_{\mathrm{CE}}\left(y^{(i)}, \hat{y}^{(i)}; \theta\right)$$

Po dogovoru, funkcija gubitka je konveksna funkcija (jedan minimum)

Metoda opadanja gradijenta garantira da će se minimum pronaći

Opadanje gradijenta

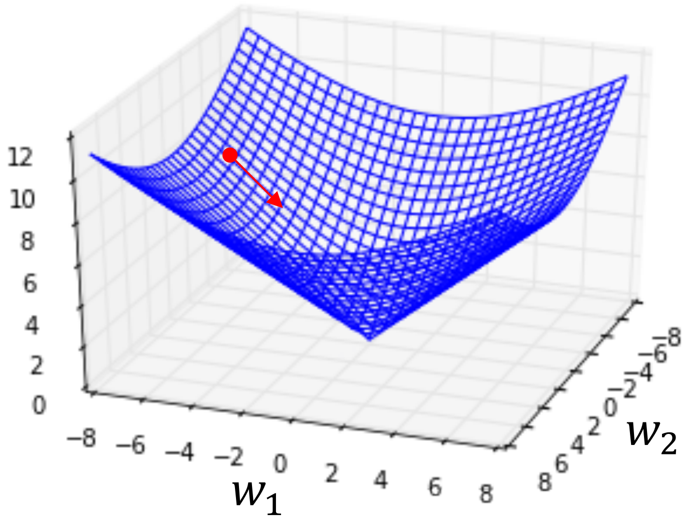
Pretpostavimo da je funkcija gubitka $f(x; w)$ od jednog parametra w



$w^{t+1} = w^t - \eta \frac{d}{dw} f(x; w)$ gdje je η **stopa učenja**

Opadanje gradijenta

Poopćimo funkciju gubitka $f(x;\theta)$ na više parametra w_i



$\theta^{t+1} = \theta^t - \eta\nabla L(f(x;\theta), y)$ gdje je

$$\nabla L(f(x;\theta), y) = \begin{bmatrix} \frac{\partial}{\partial w_1} L(f(x;\theta), y) \\ \frac{\partial}{\partial w_2} L(f(x;\theta), y) \\ \vdots \\ \frac{\partial}{\partial w_m} L(f(x;\theta), y) \end{bmatrix}$$

Opadanje gradijenta

Opadanje gradijenta kod logističke regresije

Kako izračunati $\nabla L(f(x;\theta), y)$

$$L_{CE}(w,b) = -[y\log\big(\sigma(w\cdot x+b)\big) + (1-y)\log\big(1-\sigma(w\cdot x+b)\big)]$$

$$\frac{\partial L_{CE}(w,b)}{\partial w_j} = \frac{\partial}{\partial w_j} - \big[y\log\big(\sigma(w\cdot x+b)\big) + (1-y)\log\big(1-\sigma(w\cdot x+b)\big)\big]$$
$$= -\left[\frac{\partial}{\partial w_j} y\log\big(\sigma(w\cdot x+b)\big) + \frac{\partial}{\partial w_j}(1-y)\log\big(1-\sigma(w\cdot x+b)\big)\right]$$
$$= -\frac{y}{\partial(w\cdot x+b)}\frac{\partial}{\partial w_j}\sigma(w\cdot x+b) - \frac{1-y}{1-\sigma(w\cdot x+b)}\frac{\partial}{\partial w_j}\big(1-\sigma(w\cdot x+b)\big)$$
$$= -\left[\frac{y}{\partial(w\cdot x+b)} - \frac{1-y}{1-\sigma(w\cdot x+b)}\right]\frac{\partial}{\partial w_j}\sigma(w\cdot x+b)$$
$$= -\left[\frac{y-\sigma(w\cdot x+b)}{\partial(w\cdot x+b)[1-\sigma(w\cdot x+b)]}\right]\sigma(w\cdot x+b)[1-\sigma(w\cdot x+b)]\frac{\partial\sigma(w\cdot x+b)}{\partial w_j}$$
$$= -[y-\sigma(w\cdot x+b)]x_j$$
$$= [\sigma(w\cdot x+b)-y]x_j$$

Opadanje gradijenta kod grupnog treniranja

Grupno treniranje (batch training)

– određivanje gradijenta za cijeli skup podataka

Treniranje u mini grupama (mini-batch tranining)

– određivanje gradijenta za $\boldsymbol{m}$ podatak iz skupa podataka ($m = 512, 1024, \ldots$)

$$cost(w,b) = \frac{1}{m}\sum_{i=1}^{m} L_{CE}(\hat{y}^{(i)}, y^{(i)})$$

$$cost(w,b) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\left(\sigma\left(w \cdot x^{(i)} + b\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(w \cdot x^{(i)} + b\right)\right)$$

$$\frac{\partial cost(w,b)}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m} [\sigma\left(w \cdot x^{(i)} + b\right) - y^{(i)}]{x_j}^{(i)}$$

Stohastičko opadanje gradijenta

Algoritam stohastičkog opadanja gradijenta

$\theta \leftarrow 0$

ponovi T puta

  za svaki $(x^{(i)}, y^{(i)})$ po slučajnom redoslijedu

  izračunaj $\hat{y}^{(i)} = f(x^{(i)}; \theta)$

  izračunaj gubitak $L(\hat{y}^{(i)}, y^{(i)})$

  $\theta \leftarrow \theta - \eta \nabla L(f\left(x^{(i)}; \theta\right), y^{(i)})$

vrati θ

Stohastičko opadanje gradijenta

Primjer: neka je

- $x = [x_1, x_2] = [3, 2]$
- za θ^0 imamo $w = [w_1, w_2] = [0,0], b = 0$
- $\eta = 0.1$

Znamo $\theta^{t+1} = \theta^t - \eta \nabla_\theta L(f\left(x^{(i)}; \theta\right), y^{(i)})$ stoga

$$\nabla_{w,b} = \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\mathrm{CE}}(w, b) \\ \frac{\partial}{\partial w_2} L_{\mathrm{CE}}(w, b) \\ \frac{\partial}{\partial b} L_{\mathrm{CE}}(w, b) \end{bmatrix} = \begin{bmatrix} (\sigma(w \cdot x + b) - y)x_1 \\ (\sigma(w \cdot x + b) - y)x_2 \\ \sigma(w \cdot x + b) - y \end{bmatrix}$$

$$= \begin{bmatrix} (\sigma(0) - 1)x_1 \\ (\sigma(0) - 1)x_2 \\ \sigma(0) - 1 \end{bmatrix} = \begin{bmatrix} -0.5x_1 \\ -0.5x_2 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.5 \cdot 3 \\ -0.5 \cdot 2 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -1.0 \\ -0.5 \end{bmatrix}$$

Stohastičko opadanje gradijenta

Primjer:

- $x = [x_1, x_2] = [3, 2]$
- $w = [w_1, w_2] = [0, 0],\ b = 0$
- $\eta = 0.1$

$$\theta^1 = \theta^0 - \eta \nabla_{w,b}$$

$$\theta^1 = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} - \eta \nabla_{w,b} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -1.5 \\ -1.0 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.15 \\ -0.10 \\ -0.05 \end{bmatrix}$$

Regularizacija

- Ako je osobina savršeno prediktivna (pojavljuje se samo u jednoj klasi), dobit će veliku težinu
- Težine osobina će nastojati savršeno odgovarati detaljima u podacima za učanje (**prenaučenost modela- overfitting**)
- Dobro naučen model mora moći generalizirati na dosad neviđenim podacima za testiranje.
- Regularizacija $R(w)$ se dodaje funkciji gubitka

$$\hat{w} = \operatorname{argmax}_w \sum_{i=1}^{m} \log P\left(y^{(i)} \middle| x^{i}\right) - \alpha R(w)$$

- $R(w)$ služi za "kažnjavanje" velikih težina

Regularizacija

- Dvije često korištene regularizacije
  - L2 regularizacija $R(W) = \|W\|_2^2 = \sum_{j=1}^{n} w_j^2$ - Euklidska udaljenost
  - L1 regularizacija $R(W) = \|W\|_1 = \sum_{j=1}^{n} |w_j|$ - Manhattan udaljenost
- L2 regularizacija se lakše optimizira (jednostavnija derivacija)

Primjer: Podaci

$$Train = \left\{ \left(\begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix}, y^{(1)} \right), \dots, \left(\begin{bmatrix} x_1^{(m)} \\ \vdots \\ x_n^{(m)} \end{bmatrix}, y^{(m)} \right) \right\} = \left\{ \left(\begin{bmatrix} 23 \\ 9 \\ 1 \\ 7 \\ 1 \\ 5 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 12 \\ 5 \\ 1 \\ 10 \\ 1 \\ 5 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 14 \\ 12 \\ 1 \\ 1 \\ 0 \\ 5 \end{bmatrix}, 1 \right) \right\}$$

$$X = \begin{bmatrix} x^{(1)} & \dots & x^{(m)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix} = \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \end{bmatrix}$$

$$Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Primjer: Inicijalizacija

$$\mathrm{W} = \begin{bmatrix} w_1 & \dots & w_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$b = 0$$

$$\theta = \begin{bmatrix} W & b \end{bmatrix} = \begin{bmatrix} w_1 & \dots & w_n & b \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\eta = 0.01$$

Primjer: Učenje - predikcija

$$\hat{y}^{(i)} = \sigma(W \cdot x^{(i)} + b) = \sigma\left(\begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} + b\right) = \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(i)} + b\right)$$

$$\hat{Y} = \sigma(W \cdot X + b)$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \sigma\left(\begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}\right) = \begin{bmatrix} \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(1)} + b\right) \\ \vdots \\ \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(m)} + b\right) \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) =$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \begin{bmatrix} \sigma(0 \cdot 23 + 0 \cdot 9 + 0 \cdot 1 + 0 \cdot 7 + 0 \cdot 1 + 0 \cdot 5 + 0) \\ \sigma(0 \cdot 12 + 0 \cdot 5 + 0 \cdot 1 + 0 \cdot 10 + 0 \cdot 1 + 0 \cdot 5 + 0) \\ \sigma(0 \cdot 14 + 0 \cdot 12 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 5 + 0) \end{bmatrix} = \begin{bmatrix} \sigma(0) \\ \sigma(0) \\ \sigma(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Primjer: Učenje - predikcija

$$\hat{y}^{(i)} = \sigma\left(\theta \cdot \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left([W \quad b] \cdot \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left([w_1 \quad \dots \quad w_n \quad b] \cdot \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \\ 1 \end{bmatrix}\right) = \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(i)} + b \cdot 1\right)$$

$$\hat{Y} = \sigma\left(\theta \cdot \begin{bmatrix} X \\ 1 \end{bmatrix}\right)$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \sigma\left([w_1 \quad \dots \quad w_n \quad b] \cdot \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \\ 1 & \dots & 1 \end{bmatrix}\right) = \begin{bmatrix} \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(1)} + b \cdot 1\right) \\ \vdots \\ \sigma\left(\sum_{k=1}^{n} w_k\, x_k^{(m)} + b \cdot 1\right) \end{bmatrix}$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left([0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \cdot \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \\ 1 & 1 & 1 \end{bmatrix}\right) =$$

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \end{bmatrix} = \sigma\left(\begin{bmatrix} 0 \cdot 23 + 0 \cdot 9 + 0 \cdot 1 + 0 \cdot 7 + 0 \cdot 1 + 0 \cdot 5 + 0 \cdot 1 \\ 0 \cdot 12 + 0 \cdot 5 + 0 \cdot 1 + 0 \cdot 10 + 0 \cdot 1 + 0 \cdot 5 + 0 \cdot 1 \\ 0 \cdot 14 + 0 \cdot 12 + 0 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 0 \cdot 5 + 0 \cdot 1 \end{bmatrix}\right) = \begin{bmatrix} \sigma(0) \\ \sigma(0) \\ \sigma(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \\ \frac{1}{1+e^{-0}} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Primjer: Učenje - trošak

$$L_{\mathrm{CE}}\left(\hat{y}^{(i)}, y^{(i)}\right) = -\left(y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)\right)$$

$$L_{\mathrm{CE}}\left(\hat{Y}, Y\right) = -\frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)\right)$$

$$L_{\mathrm{CE}}\left(\hat{Y}, Y\right) = L_{\mathrm{CE}}\left(\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) =$$

$$-\frac{1}{3}\begin{bmatrix} (0 \log(0.5) + (1 - 0) \log(1 - 0.5)) + \\ (0 \log(0.5) + (1 - 0) \log(1 - 0.5)) + \\ (1 \log(0.5) + (1 - 1) \log(1 - 0.5)) \end{bmatrix}$$

$$= -\frac{1}{3}[-0.69 - 0.69 - 0.69] = 0.69$$

Primjer: Učenje – opadanje gradijenta

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta}$$

$$\nabla_{\theta} = \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\text{CE}}(W,b) \\ \vdots \\ \frac{\partial}{\partial w_n} L_{\text{CE}}(W,b) \\ \frac{\partial}{\partial b} L_{\text{CE}}(W,b) \end{bmatrix} = \frac{1}{\text{m}} \begin{bmatrix} X \\ 1 \end{bmatrix} (\hat{Y} - Y) = \frac{1}{\text{m}} \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \\ 1 & \dots & 1 \end{bmatrix} \left(\begin{bmatrix} \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \right) = \frac{1}{m} \begin{bmatrix} \sum_{j=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_1^{(i)} \\ \vdots \\ \sum_{j=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_n^{(i)} \\ \sum_{j=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \end{bmatrix}$$

$$\nabla_{\theta} = \frac{1}{3} \begin{bmatrix} 23 & 12 & 14 \\ 9 & 5 & 12 \\ 1 & 1 & 1 \\ 7 & 10 & 1 \\ 1 & 1 & 0 \\ 5 & 5 & 5 \\ 1 & 1 & 1 \end{bmatrix} \left(\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = \frac{1}{3} \begin{bmatrix} (0.5-0)\cdot 23 + (0.5-0)\cdot 12 + (0.5-1)\cdot 14 \\ (0.5-0)\cdot 9 + (0.5-0)\cdot 5 + (0.5-1)\cdot 12 \\ (0.5-0)\cdot 1 + (0.5-0)\cdot 1 + (0.5-1)\cdot 1 \\ (0.5-0)\cdot 7 + (0.5-0)\cdot 10 + (0.5-1)\cdot 1 \\ (0.5-0)\cdot 1 + (0.5-0)\cdot 1 + (0.5-1)\cdot 0 \\ (0.5-0)\cdot 5 + (0.5-0)\cdot 5 + (0.5-1)\cdot 5 \\ (0.5-0)\cdot 1 + (0.5-0)\cdot 1 + (0.5-1)\cdot 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 10.5 \\ 1 \\ 0.5 \\ 8 \\ 1 \\ 2.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 0.33 \\ 0.17 \\ 2.67 \\ 0.33 \\ 0.83 \\ 0.17 \end{bmatrix}$$

Primjer: Učenje – opadanje gradijenta

$$\theta^{t+1} = \theta^{t} - \eta\nabla_{\theta^t}$$

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \\ b \end{bmatrix}^{t+1} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ b \end{bmatrix}^{t} - \eta[]\nabla_{\theta^t}$$

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ b \end{bmatrix}^{2} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ b \end{bmatrix}^{1} - \eta \begin{bmatrix} \frac{\partial}{\partial w_1} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial w_2} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial w_3} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial w_4} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial w_5} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial w_6} L_{\mathrm{CE}}(W,b) \\ \frac{\partial}{\partial b} L_{\mathrm{CE}}(W,b) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.01 \begin{bmatrix} 3.5 \\ 0.33 \\ 0.17 \\ 2.67 \\ 0.33 \\ 0.83 \\ 0.17 \end{bmatrix} = \begin{bmatrix} -0.035 \\ -0.0033 \\ -0.0017 \\ -0.0267 \\ -0.0033 \\ -0.0083 \\ -0.0017 \end{bmatrix}$$

Primjer: Učenje

Nakon 2000 iteracija

$\theta^1 = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$

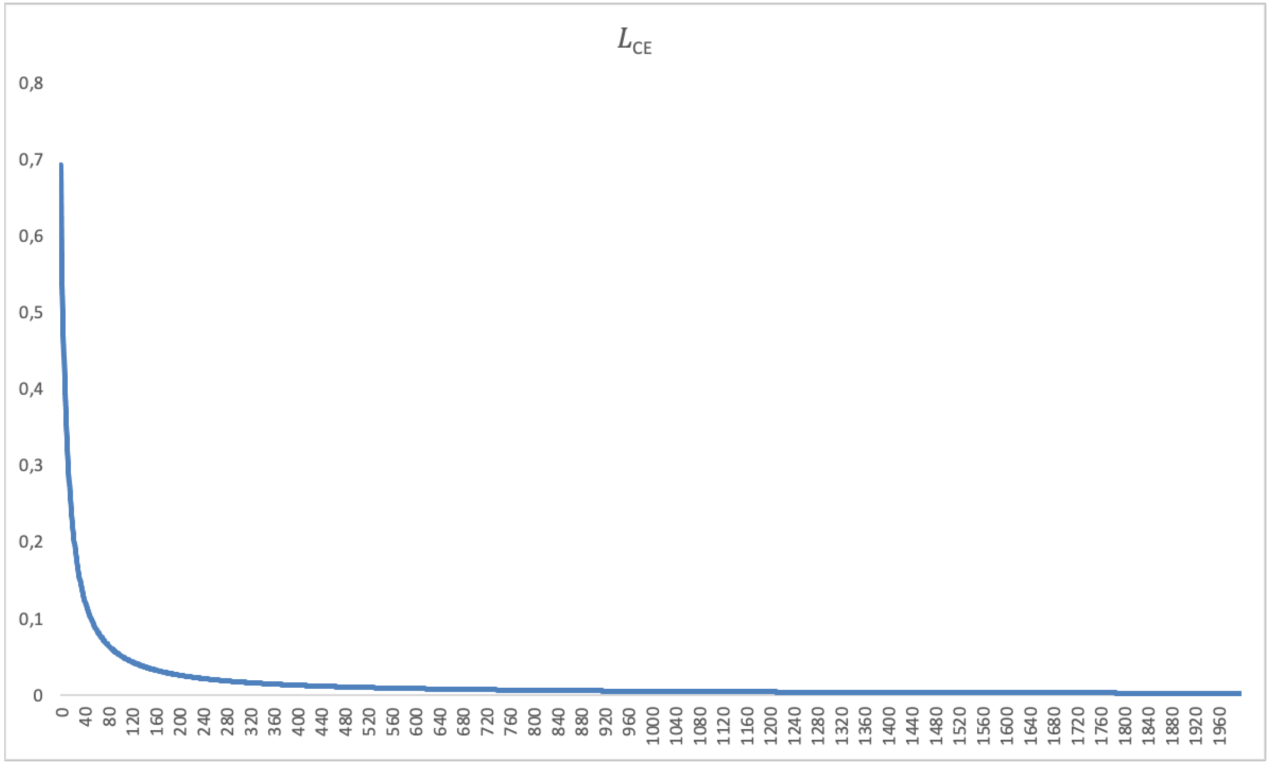
$\theta^2 = [-0.035 \quad -0.0033 \quad -0.0017 \quad -0.0267 \quad -0.0033 \quad -0.0083 \quad -0.0017]$

$\theta^3 = [-0.037 \quad -0.0095 \quad -0.0014 \quad -0.0412 \quad -0.0053 \quad -0.0072 \quad -0.0014]$

...

$\theta^{1999} = [-0.3665 \quad -0.8842 \quad -0.0293 \quad -0.8574 \quad -0.1299 \quad -0.1465 \quad -0.0293]$

$\theta^{2000} = [-0.3665 \quad -0.8843 \quad -0.0293 \quad -0.8574 \quad -0.1299 \quad -0.1465 \quad -0.0293]$



Primjer: Testiranje

$$Test = \left\{ \left(\begin{bmatrix} 18 \\ 15 \\ 1 \\ 4 \\ 1 \\ 5 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 8 \\ 4 \\ 0 \\ 0 \\ 0 \\ 5 \end{bmatrix}, 0 \right) \right\}$$

$$X = \begin{bmatrix} 18 & 8 \\ 15 & 4 \\ 1 & 0 \\ 4 & 0 \\ 1 & 0 \\ 5 & 5 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\theta = [-0.3665 \quad -0.8843 \quad -0.0293 \quad -0.8574 \quad -0.1299 \quad -0.1465 \quad -0.0293]$$

$$\hat{Y} = \sigma\left(\theta \cdot \begin{bmatrix} X \\ 1 \end{bmatrix}\right)$$

$$= \sigma\left([-0.3665 \quad -0.8843 \quad -0.0293 \quad -0.8574 \quad -0.1299 \quad -0.1465 \quad -0.0293] \begin{bmatrix} 18 & 8 \\ 15 & 4 \\ 1 & 0 \\ 4 & 0 \\ 1 & 0 \\ 5 & 5 \\ 1 & 1 \end{bmatrix} \right) = \begin{bmatrix} 0.98 \\ 0.797 \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} 0.98 \\ 0.797 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Uvod u obradu prirodnog jezika

9.2. Višeklasna logistička regresija (MaxEnt)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Višeklasna logistička regresija

- **Višeklasna logistička regresija** se još zove
  - Softmax regresija
  - Maxent klasifikator
- Klase $C = \{c_1, c_2, \dots, c_K\}$
- Funkcija klasifikacije softmax za vektor

$$z = [z_1, z_2, \dots, z_K]$$

$$\text{softmax}(z_j) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad 1 \le i \le K$$

$$\text{softmax}(z) = \left[\frac{e^{z_1}}{\sum_{j=1}^{K} e^{z_j}}, \frac{e^{z_2}}{\sum_{j=1}^{K} e^{z_j}}, \dots, \frac{e^{z_m}}{\sum_{j=1}^{K} e^{z_j}}\right]$$

- Nazivnik $\sum_{j=1}^{K} e^{z_j}$ služi za normalizaciju vrijednosti u vjerojatnosti

Višeklasna logistička regresija

- **Osobine**

$$f_i(x) = f_i(c, x) \text{ osobina } i \text{ za klasu } c$$

- Primjer, klasifikacija teksta u 3 klase: $\{+, -, 0\}$

osobina	definicija	W
$f_1(+, x)$	$\begin{cases} 1, & \text{ako "!"} \in d \\ 0, & \text{inaće} \end{cases}$	-4.5
$f_1(-, x)$	$\begin{cases} 1, & \text{ako "!"} \in d \\ 0, & \text{inaće} \end{cases}$	2.6
$f_1(0, x)$	$\begin{cases} 1, & \text{ako "!"} \in d \\ 0, & \text{inaće} \end{cases}$	1.3

Višeklasna logistička regresija

- **Vjerojatnost za klasu** c_i

$$\mathrm{p}(\mathrm{y} = c_i|x) = \frac{e^{w_{c_i}x+b_{c_i}}}{\sum_{j=1}^{K} e^{w_{c_j}x+b_{c_j}}}$$

- Funkcija gubitka

$$L_{CE}(\hat{y}, y) = -\sum_{k=1}^{K} 1\{y = c_k\} \log p(y = c_k|x)$$

$$= -\sum_{k=1}^{K} 1\{y = c_k\} \log \frac{e^{w_{c_k}x+b_{c_k}}}{\sum_{j=1}^{K} e^{w_{c_j}x+b_{c_j}}}$$

Višeklasna logistička regresija

- Gradijent

$$\frac{\partial L_{CE}(\hat{y}, y)}{\partial w_{c_m}} = \big(1\{y = c_k\} - p(y = c_m | x)\big) x_m$$

$$= \left(1\{y = c_k\} - \log \frac{e^{w_{c_m} x + b_{c_m}}}{\sum_{j=1}^{K} e^{w_{c_j} x + b_{c_j}}}\right) x_m$$

Primjer: Podaci

$$TrainSet = \{(d_1, c_1), (d_2, c_1), (d_3, c_2), (d_4, c_3)\}$$

$$C = \{c_1, c_2, c_3\} \qquad hot(C) = \{[1\ 0\ 0], [0\ 1\ 0], [0\ 0\ 1]\}$$

$$FeatureSet = \begin{Bmatrix} \left[f_1^{c_1}(d_1) \quad f_2^{c_1}(d_1)\right], [1\ 0\ 0] \\ \left[f_1^{c_1}(d_2) \quad f_2^{c_1}(d_2)\right], [1\ 0\ 0] \\ \left[f_1^{c_2}(d_3) \quad f_2^{c_2}(d_3)\right], [0\ 1\ 0] \\ \left[f_1^{c_3}(d_4) \quad f_2^{c_3}(d_4)\right], [0\ 0\ 1] \end{Bmatrix}$$

$$= \begin{Bmatrix} \left[x_1^{d_1} \quad x_2^{d_1}\right], [1\ 0\ 0] \\ \left[x_1^{d_2} \quad x_2^{d_2}\right], [1\ 0\ 0] \\ \left[x_1^{d_3} \quad x_2^{d_3}\right], [0\ 1\ 0] \\ \left[x_1^{d_4} \quad x_2^{d_4}\right], [0\ 0\ 1] \end{Bmatrix} = \{X, Y\}$$

Primjer: Predikcija

$\Phi = (W, B)$

$\hat{Y} = softmax(XW + B)$

$$\hat{Y} = softmax\left(\begin{bmatrix} x_1^{d_1} & x_2^{d_1} \\ x_1^{d_2} & x_2^{d_2} \\ x_1^{d_3} & x_2^{d_3} \\ x_1^{d_4} & x_2^{d_4} \end{bmatrix} \begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \end{bmatrix} + \begin{bmatrix} b^{c_1} \\ b^{c_2} \\ b^{c_3} \end{bmatrix}\right)$$

$$\hat{Y} = softmax\left(\begin{bmatrix} x_1^{d_1}w_1^{c_1} + x_2^{d_1}w_2^{c_1} + b^{c_1} & x_1^{d_1}w_1^{c_2} + x_2^{d_1}w_2^{c_2} + b^{c_2} & x_1^{d_1}w_1^{c_3} + x_2^{d_1}w_2^{c_3} + b^{c_3} \\ x_1^{d_2}w_1^{c_1} + x_2^{d_2}w_2^{c_1} + b^{c_1} & x_1^{d_2}w_1^{c_2} + x_2^{d_2}w_2^{c_2} + b^{c_2} & x_1^{d_2}w_1^{c_3} + x_2^{d_2}w_2^{c_3} + b^{c_3} \\ x_1^{d_3}w_1^{c_1} + x_2^{d_3}w_2^{c_1} + b^{c_1} & x_1^{d_3}w_1^{c_2} + x_2^{d_3}w_2^{c_2} + b^{c_2} & x_1^{d_3}w_1^{c_3} + x_2^{d_3}w_2^{c_3} + b^{c_3} \\ x_1^{d_4}w_1^{c_1} + x_2^{d_4}w_2^{c_1} + b^{c_1} & x_1^{d_4}w_1^{c_2} + x_2^{d_4}w_2^{c_2} + b^{c_2} & x_1^{d_4}w_1^{c_3} + x_2^{d_4}w_2^{c_3} + b^{c_3} \end{bmatrix}\right)$$

$$\hat{Y} = softmax\left(\begin{bmatrix} x^{d_1} \cdot w^{c_1} + b^{c_1} & x^{d_1} \cdot w^{c_2} + b^{c_2} & x^{d_1} \cdot w^{c_3} + b^{c_3} \\ x^{d_2} \cdot w^{c_1} + b^{c_1} & x^{d_2} \cdot w^{c_2} + b^{c_2} & x^{d_2} \cdot w^{c_2} + b^{c_2} \\ x^{d_3} \cdot w^{c_1} + b^{c_1} & x^{d_3} \cdot w^{c_2} + b^{c_2} & x^{d_3} \cdot w^{c_2} + b^{c_2} \\ x^{d_4} \cdot w^{c_1} + b^{c_1} & x^{d_4} \cdot w^{c_2} + b^{c_2} & x^{d_4} \cdot w^{c_2} + b^{c_2} \end{bmatrix}\right)$$

Primjer: Predikcija

$$\hat{Y} = softmax\left(\begin{bmatrix} x^{d_1} \cdot w^{c_1} + b^{c_1} & x^{d_1} \cdot w^{c_2} + b^{c_2} & x^{d_1} \cdot w^{c_3} + b^{c_3} \\ x^{d_2} \cdot w^{c_1} + b^{c_1} & x^{d_2} \cdot w^{c_2} + b^{c_2} & x^{d_2} \cdot w^{c_2} + b^{c_2} \\ x^{d_3} \cdot w^{c_1} + b^{c_1} & x^{d_3} \cdot w^{c_2} + b^{c_2} & x^{d_3} \cdot w^{c_2} + b^{c_2} \\ x^{d_4} \cdot w^{c_1} + b^{c_1} & x^{d_4} \cdot w^{c_2} + b^{c_2} & x^{d_4} \cdot w^{c_2} + b^{c_2} \end{bmatrix}\right)$$

$$\hat{Y} = softmax\left(\begin{bmatrix} d_1c_1 & d_1c_2 & d_1c_3 \\ d_2c_1 & d_2c_2 & d_2c_3 \\ d_3c_1 & d_3c_2 & d_3c_3 \\ d_4c_1 & d_4c_2 & d_4c_3 \end{bmatrix}\right)$$

$$\hat{Y} = \begin{bmatrix} e^{d_1c_1} \big/ \sum_i e^{d_1c_i} & e^{d_1c_2} \big/ \sum_i e^{d_1c_i} & e^{d_1c_3} \big/ \sum_i e^{d_1c_i} \\ e^{d_2c_1} \big/ \sum_i e^{d_2c_i} & e^{d_2c_2} \big/ \sum_i e^{d_2c_i} & e^{d_2c_3} \big/ \sum_i e^{d_2c_i} \\ e^{d_3c_1} \big/ \sum_i e^{d_3c_i} & e^{d_3c_2} \big/ \sum_i e^{d_3c_i} & e^{d_3c_3} \big/ \sum_i e^{d_3c_i} \\ e^{d_4c_1} \big/ \sum_i e^{d_4c_i} & e^{d_4c_2} \big/ \sum_i e^{d_4c_i} & e^{d_4c_3} \big/ \sum_i e^{d_4c_i} \end{bmatrix}$$

Neka je $z^{d^i} = \sum_j e^{d_ic_j}$

$$\hat{Y} = \begin{bmatrix} \frac{e^{d_1c_1}}{z^{d^1}} & \frac{e^{d_1c_2}}{z^{d^1}} & \frac{e^{d_1c_3}}{z^{d^1}} \\ \frac{e^{d_2c_1}}{z^{d^2}} & \frac{e^{d_2c_2}}{z^{d^2}} & \frac{e^{d_2c_3}}{z^{d^2}} \\ \frac{e^{d_3c_1}}{z^{d^3}} & \frac{e^{d_3c_2}}{z^{d^3}} & \frac{e^{d_3c_3}}{z^{d^3}} \\ \frac{e^{d_4c_1}}{z^{d^4}} & \frac{e^{d_4c_2}}{z^{d^4}} & \frac{e^{d_4c_3}}{z^{d^4}} \end{bmatrix}$$

Primjer: Učenje

$$\Theta^{t+1} = \Theta^{t} - \eta \nabla_{\Theta} L$$

za d_1 imamo klasu $c_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$

$$\begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \\ b^{c_1} & b^{c_2} & b^{c_3} \end{bmatrix}^{t+1} = \begin{bmatrix} w_1^{c_1} & w_1^{c_2} & w_1^{c_3} \\ w_2^{c_1} & w_2^{c_2} & w_2^{c_3} \\ b^{c_1} & b^{c_2} & b^{c_3} \end{bmatrix}^{t} - \eta \begin{bmatrix} \frac{\partial L}{\partial w_1^{c_1}} & \frac{\partial L}{\partial w_1^{c_2}} & \frac{\partial L}{\partial w_1^{c_3}} \\ \frac{\partial L}{\partial w_2^{c_1}} & \frac{\partial L}{\partial w_2^{c_2}} & \frac{\partial L}{\partial w_2^{c_3}} \\ \frac{\partial L}{\partial b^{c_1}} & \frac{\partial L}{\partial b^{c_2}} & \frac{\partial L}{\partial b^{c_3}} \end{bmatrix}$$

$$\Theta^{t+1} = \Theta^{t} - \eta \begin{bmatrix} -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d^1}}\right)\right) x_1^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d^1}}\right)\right) x_1^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d^1}}\right)\right) x_1^{d_1} \\ -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d^1}}\right)\right) x_2^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d^1}}\right)\right) x_2^{d_1} & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d^1}}\right)\right) x_2^{d_1} \\ -\left(1 - \log\left(\frac{e^{d_1 c_1}}{z^{d^1}}\right)\right) 1 & -\left(0 - \log\left(\frac{e^{d_1 c_2}}{z^{d^1}}\right)\right) 1 & -\left(0 - \log\left(\frac{e^{d_1 c_3}}{z^{d^1}}\right)\right) 1 \end{bmatrix}$$