

Funkce Softmax a její využití v jazykových modelech

Materiál pro studenty učitelství matematiky

Napsalo Gemini 3.1 pro pod redakcí MŠ

Milí budoucí kolegové učitelé matematiky,

při výuce statistiky a pravděpodobnosti se často setkáte s tím, jak převádět naměřená data nebo skóre na smysluplné pravděpodobnosti. Jedním z nejdůležitějších nástrojů moderní aplikované matematiky a umělé inteligence, který přesně toto dělá, je **funkce softmax**. Pojdme si ji rozebrat krok za krokem.

1 Motivace: Výběr z N jevů

Představme si, že máme N různých jevů a u každého známe jeho četnost (nebo nějaké obecné skóre) n_k , kde $k \in \{1, \dots, N\}$. Naším cílem je z těchto jevů náhodně vybrat jeden tak, aby pravděpodobnost výběru odrážela původní četnosti – konkrétně chceme, aby jev s největší četností byl vybírán nejčastěji.

Nejjednodušším přístupem by bylo použít klasickou relativní četnost:

$$p_k = \frac{n_k}{\sum_{i=1}^N n_i}$$

To funguje skvěle, pokud jsou n_k kladná čísla. Co když ale náš model generuje skóre, která mohou být záporná, nebo chceme více zvýraznit rozdíly mezi jednotlivými jevy? Zde přichází na řadu exponenciální funkce.

Funkce **softmax** definuje pravděpodobnost k -tého jevu takto:

$$p_k = \frac{e^{n_k}}{\sum_{i=1}^N e^{n_i}}$$

Jmenovatel zde slouží jako normalizační konstanta, která zaručuje, že součet všech pravděpodobností je přesně 1.

2 Softmax ve světě LLM a generování textu

Funkce softmax je naprosto klíčová pro fungování moderních jazykových modelů. Pro pochopení kontextu si vysvětleme dva základní pojmy:

- **LLM (Large Language Model – Velký jazykový model):** Komplexní matematický model postavený na architektuře neuronových sítí. Jeho úkolem je na základě předchozího textu vypočítat, co by mělo s největší pravděpodobností následovat.
- **Token:** Jazykové modely nečtou text po písmenech, ale po tzv. tokenech (jedno písmeno, slabika, nebo i celé slovo).

Když LLM předpovídá další token, neuronová síť vygeneruje pro každý token ve slovníku hrubé skóre (tzv. logit), což odpovídá našim n_k . Aby model mohl jeden z nich smysluplně vylosovat, prožene tato hrubá skóre funkcí softmax, čímž získá pravděpodobnostní rozdělení.

3 Parametr Teplota (Temperature)

Při losování tokenů chceme mít kontrolu nad tím, jak moc má model riskovat. K tomu slouží parametr zvaný **teplota**, značený jako $T > 0$. Tento parametr vnášíme do funkce softmax tak, že jím vydělíme vstupní skóre:

$$p_k(T) = \frac{e^{\frac{n_k}{T}}}{\sum_{i=1}^N e^{\frac{n_i}{T}}}$$

3.1 Limita pro velkou teplotu ($T \rightarrow \infty$)

Zvyšujeme-li teplotu nade všechny meze, exponent se blíží nule: $\lim_{T \rightarrow \infty} \frac{n_k}{T} = 0$. Dosazením získáme:

$$\lim_{T \rightarrow \infty} p_k(T) = \frac{e^0}{\sum_{i=1}^N e^0} = \frac{1}{N}$$

Interpretace: Rozdíly ve skóre se zcela smažou. Získáme **rovnoměrné (uniformní) rozdělení** a model vybírá tokeny zcela náhodně (nesmyslný text).

3.2 Limita pro malou teplotu ($T \rightarrow 0^+$)

Předpokládejme jeden unikátní maximální prvek s hodnotou n_{max} . Vytknutím $e^{\frac{n_{max}}{T}}$ (rozšířením zlomku výrazem $e^{-\frac{n_{max}}{T}}$) získáme:

$$p_k(T) = \frac{e^{\frac{n_k - n_{max}}{T}}}{\sum_{i=1}^N e^{\frac{n_i - n_{max}}{T}}}$$

Pro $T \rightarrow 0^+$ se pro $n_i < n_{max}$ exponent blíží do $-\infty$ a $e^{-\infty} \rightarrow 0$. Pro maximální prvek je exponent 0 a $e^0 = 1$. Z toho plyne, že $p_{max} \rightarrow 1$ a ostatní $p_k \rightarrow 0$.

Interpretace: Funkce se mění na **argmax**. Rozdělení se zhroutí na nejčastější jev, model se stává naprosto deterministickým a ztrácí kreativitu.

4 Příklad: Výpočet pro skóre [5, 2, 0]

Máme $N = 3$ jevy s hrubým skóre: $n_1 = 5, n_2 = 2, n_3 = 0$.

Výpočet pro výchozí teplotu ($T = 1$)

Váhy (čitatel): $e^5 \approx 148,41$; $e^2 \approx 7,39$; $e^0 = 1$.

Normalizační součet: $\sum e^{n_i} \approx 156,80$.

Konečné pravděpodobnosti:

- $p_1 = 148,41/156,80 \approx 0,946$ (94,6 %)
- $p_2 = 7,39/156,80 \approx 0,047$ (4,7 %)
- $p_3 = 1/156,80 \approx 0,006$ (0,6 %)

Srovnání pro různé teploty

Teplota T	Jev 1 (5)	Jev 2 (2)	Jev 3 (0)	Charakteristika rozdělení
$T = 0,1$	$> 99,99 \%$	$< 0,01 \%$	$< 0,01 \%$	Tvrdé maximum: Model je deterministický, vybírá téměř výhradně favorita.
$T = 1$	$94,6 \%$	$4,7 \%$	$0,6 \%$	Standardní softmax: Favorit dominuje, ale je dán prostor i druhé variantě.
$T = 5$	$52,2 \%$	$28,6 \%$	$19,2 \%$	Zploštění: Rozdíly se smazávají, model více „riskuje“.

5 Navazující otázky pro studenty

Pro ověření porozumění odpovězte na následující otázky:

1. Proč používáme v softmaxu exponenciální funkci a ne pouze podíl původních skóre?
2. Co by se stalo s pravděpodobnostním rozdělením v našem příkladu (skóre 5, 2, 0), kdybychom ke každému skóre přičetli konstantu, například $C = 10$?
3. Představte si, že si píšete s AI asistentem a asistent začne generovat neustále se opakující fráze bez jakékoliv kreativity. Myslíte si, že jeho vnitřní parametr teploty je nastaven blízko k nule, nebo je naopak velmi vysoký?