Softmax: A tutorial

Michael Franke & Judith Degen

started: March 29 2022, this version compiled on: March 30, 2022

1 Motivation

- softmax is a ubiquitous helper function, frequently used as a probabilistic link function for unordered categorical data
- the softmax function occurs frequently (in the cognitive sciences), e.g.:
    - neural networks
    - multinomial regression
    - in probabilistic statistical / cognitive models for discrete choice
- the goal of this tutorial is to describe the softmax function in increasing level of mathematical and conceptual detail, so as to enable better understanding of the models in which it occurs
- different types of models require different depth of understanding of the softmax function:
    - for general-purpose models like ANNs and regression it suffices to understand what softmax does in terms of its input-output behavior, possibly with a bit of understanding of its key properties (Sections 3–4)
    - for theory-driven models like probabilistic cognitive models, whose internal parameters aspire to be meaningfully interpretable, it may be useful to additionally understand how the softmax function can be motivated conceptually and mathematically derived; we look at two such motivations and derivations here:
        - softmax as a stochastic choice function (Section 5)
        - softmax as a maximum entropy distribution (Section 6)
- a secondary goal behind increasing understanding of the softmax function is to provide better interpretability of its key parameter, the *optimality parameter*; this is very important for interpretability in at least two domains:
    - for interpreting model fits (Bayesian or point-estimates, like MLE), which yield numerical estimates for the softmax function's optimality parameter
    - for enabling the interpretation and specification of reasonable priors in Bayesian modeling

2 Softmax basics: definition, notation & terminology

- Let $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$ be a vector of n unordered *outcome categories*, and $\mathbf{s} = \langle s_1, \ldots, s_n \rangle$ be a vector of the same length giving us *scores*.
- Outcome categories could be labels (Does this picture show a dog, a cat or a goldfish? Is this text fiction, law, or science?) or actions of an agent (Will John order pizza, pasta or salad?)
- The score s_i associated with outcome x_i is a real number which we will use to define the probability of x_i.
  - The interpretability of the scores depends on the context (model) they occur in.
    - For data-driven models, like neural networks or regression models, the meaning of the scores is defined solely by their role in the softmax function itself, i.e., they are defined by which probability distributions they help create via the softmax function.
    - For theory-driven cognitive models, the scores can be intinsically meaningful, often derived from other interpretable quantities, e.g., an agent's preferences or dispositions towards different choices, or the accumulated evidence accrued for different interpretations of a given stimulus.
- The "problem" for which the softmax function provides a solution is: given the the scores $\mathbf{s}$ for categories $\mathbf{x}$ how likely is each category? I.e., the softmax function is a mapping:

  $$\text{softmax}\colon \mathbf{s} \mapsto \mathbf{p}$$

  from numerical scores to a probability vector $\mathbf{p} = \langle p_1, \ldots, p_n \rangle$, where p_i describes the probability that category x_i obtains (given the scores s_j for all j).

  - [Explain what "problem" and "solution" mean here for the two types of models we consider. Link function.]
- The definition of the softmax function is:[1]

  $$\text{SoftMax}(\mathbf{s}; \alpha) = \mathbf{p}, \quad \text{with: } p_i = \frac{\exp(\alpha\ s_i)}{\sum_j \exp(\alpha\ s_j)}$$

  Authors often use simpler notation, omitting the normalizing constant $Z = \sum_j \exp(\alpha\ s_j)$ and would just write:

  $$p_i \propto \exp(\alpha\ s_i)$$

- The softmax function has an *optimality parameter* α, which is sometimes omitted, i.e., implicitly set to 1.

3 Softmax by I/O

4 Properties of softmax

- For all i, if all s_j are finite, $p_i > 0$.

[1] We write $\exp s_i$ to mean e^{s_i}, using the same bracketing and association rules as for the inverse log operator.

  - [todo: insert proof]

- Softmax is invariant under addition: if $a \in \mathbb{R}$ is a constant, then $\text{SoftMax}(\mathbf{s}; \alpha) = \text{SoftMax}(\mathbf{s} + a; \alpha)$.

  - [todo: insert proof]

- Softmax is not invariant under multiplication: if $a \in \mathbb{R} > 0$ is a constant, then $\text{SoftMax}(\mathbf{s}; \alpha) = \text{SoftMax}(a\,\mathbf{s}; \alpha)$ only if $s_i = s_j$ for all j and j.

  - [todo: insert proof]

- Multiplicative factors can be recovered by different optimality parameters: if $a \in \mathbb{R} > 0$ is a constant, then $\text{SoftMax}(\mathbf{s}; \alpha) = \text{SoftMax}(a\,\mathbf{s}; {}^{\alpha}/_{a})$.

  - [todo: insert proof (trivial)]

- Softmax reduces to the logistic function when $n = 2$.

  - [todo: insert proof]

- What really matters to softmax are differences between scores. In particular, odds ratios ${}^{p_i}/_{p_j}$ are just a function of score differences $s_i - s_j$, namely: ${}^{p_i}/_{p_j} = \exp\left(\alpha\ (s_i - s_j)\right)$.

$$\frac{p_i}{p_j} = \frac{\exp(\alpha\ s_i)}{\sum_k \exp(\alpha s_k)} \frac{\sum_k \exp(\alpha s_k)}{\exp(\alpha\ s_j)} = \frac{\exp(\alpha\ s_i)}{\exp(\alpha\ s_j)} = \exp(\alpha\ s_i - \alpha\ s_j) = \exp(\alpha\ (s_i -\ s_j))$$

5 Softmax as stochastic choice function

In this section, we derive the softmax function as the probability entailed by a stochastic choice mechanism, in which the choice of x_i is optimal (an x_i with the highest score is chosen), but in which the scores themselves are "wiggled" each time a decision has to be made (Luce, 1959; Train, 2009). Given specific choices about the probability of the "wiggles," we can derive the softmax function as the expected choice frequencies. [Include a visualization of this.]

Let's think of $x_1, \dots, x_n$ as an agent's available choice options (actions), each with a score $s_1, \dots, s_n$ which tells us how good each choice is. The agent in question could be a human decision maker (making perceptual or economic decisions), but it could also be an abstract system "choosing" a category based on the objective to maximize the score. An optimal (or rational) agent would always only choose x_i if $s_i = \max_j s_j$, so an optimal agent would maximize the score perfectly. But let us now assume that there is room for imperfection. Mistakes happen, but not just any mistake is equally likely. It is more likely to choose a suboptimal option whose score is almost maximal than to choose a suboptimal option which is far worse. Essentially, this leads to the desire to formulate *probabilistic choice rules* such that the probability p_i of choosing x_i is higher the higher its relative score is (Luce, 1959; Train, 2009).

The softmax choice rule can be derived as the probability p_i that an agent chooses x_i as the best choice from a noise-perturbed representation $s'_1, \dots, s'_n$ of the scores, where $s_j = s_j + epsilon_j$ and all ϵ_j are independently and identically distributed random noise perturbations of the actual scores of the available options. In other words, each time the agent chooses from $x_1, \dots, x_n$, they choose $x \in \arg\max_j s_j + \epsilon_j$ for a vector of errors $\epsilon = \left\langle \epsilon_1, \dots, \epsilon_j \right\rangle$ which are sampled anew every time a choice is made (each choice a different set of "wiggles"). The errors in ϵ can be thought of as

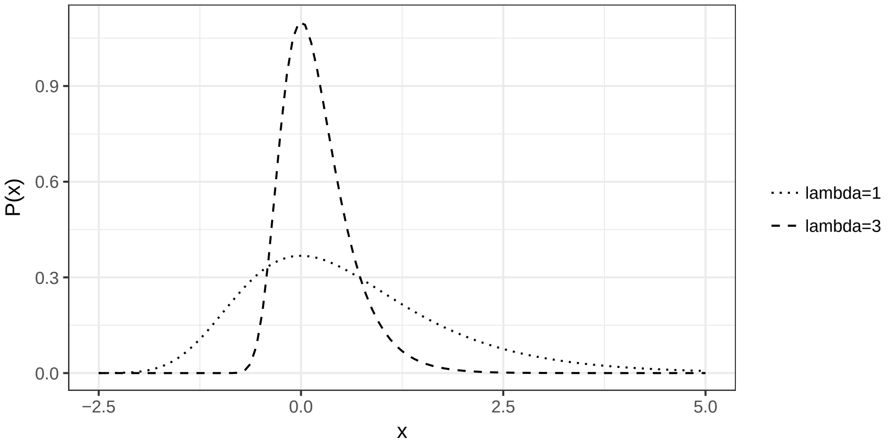

Figure 1: Examples of Gumbel probability density with location $\mu = 0$ and different values for $\beta = {}^{1}\!/_{\alpha}$: $y = f(\epsilon; \mu = 0, \beta = {}^{1}\!/_{\alpha})$ as defined in the text. [update plot, use ϵ, not x]

random errors in the computation of the scores. (We could say that the agent makes rational choices given a momentarily and subjectively distorted representation of actual scores.) The probability that an agent who maximizes (noise-perturbed) utility chooses action a_i is therefore:

$$p_i \quad = \quad P(\forall j \neq i: \ s_i + \epsilon_i > s_j + \epsilon_j) \quad = \quad P(\forall j \neq i: \ \epsilon_j < \epsilon_i + s_i - s_j) \,. \tag{1}$$

To derive the softmax function, we must make specific assumptions about the probability distribution from which each ϵ is sampled. Concretely, we assume that the error terms ϵ_j come from a **Gumbel distribution** with location $\mu = 0$ and arbitrary scale parameter $\beta > 0$. In general, the Gumbel distribution has a cumulative density function

$$F(\epsilon; \mu, \beta) = e^{-e^{-\frac{\epsilon - \mu}{\beta}}}$$

and a probability density function

$$f(\epsilon; \mu, \beta) = \frac{1}{\beta} e^{-\frac{\epsilon - \mu}{\beta}} e^{-e^{-\frac{\epsilon - \mu}{\beta}}} \,.$$

[Why different notation of exponential function?] The variance of this distribution is a function of the scale parameter: $\frac{\pi^2}{6}\beta^2$. This is important for interpreting the optimality parameter α of the softmax function, because we will set $\alpha = \frac{1}{\beta}$ to obtain:

$$F(\epsilon; \mu = 0, \beta = {}^{1}\!/_{\alpha}) = e^{-e^{-\alpha\epsilon}}$$

and a probability density function

$$f(\epsilon; \mu = 0, \beta = {}^{1}\!/_{\alpha}) = \alpha e^{-\alpha\epsilon} e^{-e^{-\alpha\epsilon}} \,.$$

Figure 1 gives examples of the latter probability density function for different values of α.

Given this assumption about the distribution of noise perturbations ϵ, we can spell out the probability p_i from (1) as a function $P(x_i; \alpha)$ of α. Let's first assume, unrealistically, that we would know the value of ϵ_i. From the right-hand side of (1), p_i would then be determined by how likely it is to sample a set of ϵ_j-s all of which are below a given threshold $\epsilon_i + s_i - s_j$. Since all ϵ_j are sampled independently, this is the product of the cumulative densities for all ϵ_j being smaller than the threshold $\epsilon_i + s_i - s_j$:

$$P(x_i;\alpha)^{|\epsilon_i} \;=\; \prod_{j \neq i} F(\epsilon_i + s_i - s_j; \mu = 0, \beta = {}^1\!/\!_\alpha) \;=\; \prod_{j \neq i} e^{-e^{-\alpha(\epsilon_i + s_i - s_j)}} .$$

But, of couse, we do not know the value of ϵ_i. We only know its distribution, so that:

$$\begin{aligned}
P(x_i;\alpha) &= \int f(\epsilon_i; \mu = 0, \beta = {}^1\!/\!_\alpha) \prod_{j \neq i} e^{-e^{-\alpha(\epsilon_i + s_i - s_j)}} \; d\epsilon_i \\
&= \int \alpha e^{-\alpha\epsilon_i} e^{-e^{-\alpha\epsilon_i}} \prod_{j \neq i} e^{-e^{-\alpha(\epsilon_i + s_i - s_j)}} \; d\epsilon_i \\
&= \alpha \int e^{-\alpha\epsilon_i} \prod_{j} e^{-e^{-\alpha(\epsilon_i + s_i - s_j)}} \; d\epsilon_i \\
&= \alpha \int e^{-\alpha\epsilon_i} \exp\left(- \sum_j e^{-\alpha(\epsilon_i + s_i - s_j)} \right) \; d\epsilon_i \\
&= \alpha \int e^{-\alpha\epsilon_i} \exp\left(-e^{-\alpha\epsilon_i} \sum_j e^{-\alpha(s_i - s_j)} \right) \; d\epsilon_i \\
&= \alpha \int e^{-\alpha\epsilon_i} \exp\left(-c e^{-\alpha\epsilon_i} \right) \; d\epsilon_i & \text{with } c = \sum_j e^{-\alpha(s_i - s_j)} \\
&= \frac{\exp(-c e^{-\alpha\epsilon_i})}{c} \Bigg|_{-\infty}^{\infty} \\
&= \lim_{\epsilon_i \to \infty} \frac{\exp(-c e^{-\alpha\epsilon_i})}{c} - \lim_{\epsilon_i \to -\infty} \frac{\exp(-c e^{-\alpha\epsilon_i})}{c} = \frac{1}{c} - 0 \\
&= \frac{1}{\sum_j e^{-\alpha(s_i - s_j)}} = \frac{1}{e^{-\alpha s_i} \sum_j e^{\alpha s_j}} = \frac{\exp(\alpha\; s_i)}{\sum_j \exp(\alpha\; s_j)} .
\end{aligned}$$

6 Softmax as maximum-entropy distribution

6.1 Informal characterization

6.2 Formal derivation

7 Interpretation of the optimality parameter

· This section seeks to answer three questions about the optimality parameter α:

1. What does a fixed value of α mean? (E.g., as returned by a point-estimate after parameter inference.) (Section 7.1)
2. How do predictions differ between two values α_1 and α_2, all else equal? Concretely, what does it mean to increase a value α_1 by a factor f, so that $\alpha_2 = f\; \alpha_1$? (Section 7.2)
3. What does all of this entail for a reasonable choice of prior probability for optimality parameters in Bayesian data analysis? (Section 7.3)

7.1 Optimality parameter as log-odds ratio for unit score difference

- We saw in Section 4 that what really matters for softmax probabilities are *differences* between scores.

$$\frac{p_i}{p_j} = \exp(\alpha\ (s_i -\ s_j))$$

- This provides an intuitive interpretation of the optimality parameter α in terms of the log-odds for a unit score difference.
    - We have a unit score difference whenever $s_i - s_j = 1$.
    - If that's the case, then:

    $$\frac{p_i}{p_j} = \exp\left(\alpha\ (s_i -\ s_j)\right) = \exp \alpha$$

    - Which means that α gives the log-odds for a unit score difference:

    $$\alpha = \log \frac{p_i}{p_j}$$

    [How best to explain what this means intuitively? Make a picture?]

- We can use this interpretation of optimality α as log-odds for unit score differences to makes sense of fitted values of α. Suppose we find that a maximum likelihood fit yields $\hat{\alpha} = 5$.

7.2 Interpreting differences between optimality parameters

- All else equal, how can we intuitively understand the difference between $\text{SoftMax}(\mathbf{s}; \alpha)$ and $\text{SoftMax}(\mathbf{s}; \alpha')$?

- We look at the factor $f = \frac{\alpha'}{\alpha}$ by which α' differs from α and note that:

$$\frac{p_i}{p_j} = \exp\left(\alpha\ (s_{i-s_j})\right)$$

$$\frac{p'_i}{p'_j} = \exp\left(\alpha'\ (s_{i-s_j})\right) = \exp\left(f\ \alpha\ (s_{i-s_j})\right) = \left[\exp\left(\alpha\ (s_{i-s_j})\right)\right]^f = \left[\frac{p_i}{p_j}\right]^f$$

- In words, if the optimality operator increases by a factor f, the odds of choosing x_i over x_j ($s_i > s_j$) increase by the power of f.

7.3 Implications for probabilistic modeling

- If we interpret α as the log odds $\log \frac{p_i}{p_j}$ for a unit difference in scores $s_i - s_j = 1$, a natural choice of family for a prior on the optimality parameter in Bayesian data analysis is a log-normal distribution (or any other log-something distribution).

- The mean of the log-normal prior distribution would correspond to the modeller's prior assumption about the expected log odds for a unit difference in scores.

- The variance of the log-normal prior distribution should be chosen based on the modeler's uncertainty about the log odds for a unit difference in scores. When choosing the variance, the result that increasing α by a factor of f increases odds by the power f can be used as guidance.

- The picture becomes more complicated, when the scores are not fixed in advance but hinge on other model parameters. In that case, priors should always be chosen "holistically" and in the light of the plausibility of the entailed prior predictive functions [insert reference to Bayesian workflow literature].

References

Luce, Duncan R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. Cambridge, MA: Cambridge University Press.