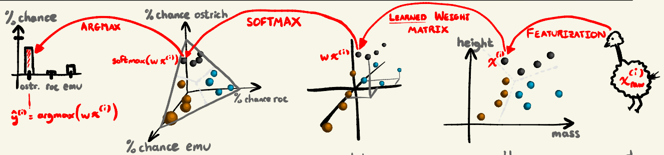
Softmax @ 5 Levels of Sophistication

Sam's Explainers
2022-06-07

A. lay person
B. MFBT developer
C. FBR developer
D. deep architect
E. researcher

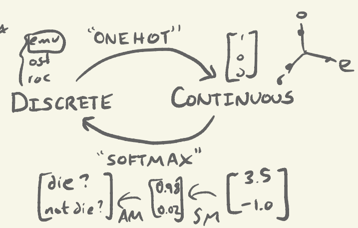


CLASSIC PARADIGM	standard (textbook-given) transform	machine-learned transform	hand designed transform
MODERN PARADIGM	standard (textbook-given) transform	machine-learned transform	hand designed transform

A. lay person

Softmax is... how we pick 'the' best answer while being fair about ties

models uncertainty: key to learning* to calibrated predictions



allows continuous methods to solve discrete tasks

B. MFBT developer

Softmax is... how we normalize real-valued scores to probabilities

$$\text{softmax}: \mathbb{R}^K \to \{(p_i : 0 \le i < K) \in \mathbb{R}^K : (\forall 0 \le i < K : 0 \le p_i) \,\&\, (\textstyle\sum p_i = 1)\} \subseteq \mathbb{R}^K$$

$$(z_i : 0 \le i < K) \longmapsto \left(\frac{\exp(z_i)}{\sum_j \exp(z_j)} : 0 \le i < K\right)$$

```
z = np.matmul(w, x[i])
e = np.exp(z)
n = np.sum(e)
p = e/n
```

one-hot

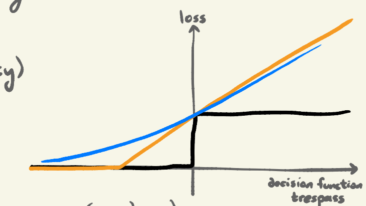
```
dl_dp = (-1./p) · y[i]
dl_dn = -np.dot(dl_dp, e)/n²
dl_de = dl_dp/n + dl_dn
dl_dz = dl_de · e
dl_dw = np.outer(dl_dz, x[i])
```

dl_dz = p - y[i]

C. FBR developer

Softmax is... a computationally convenient interface w/ likelihood loss

convex, so easy to optimize
bounded gradients (despite miracle sensitivity)
tight surrogate for 0-1 loss



$$\text{decision func trespass} = \underbrace{\left(-y^{(i)} w \cdot x^{(i)}\right)}$$

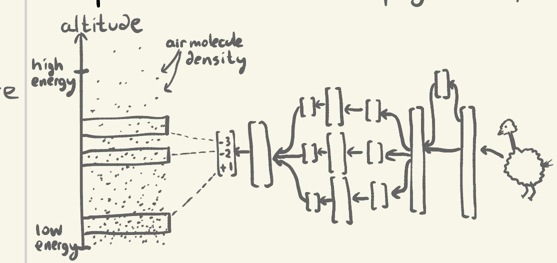
$$\text{hinge loss} = \max(0, 1 + \cdots)$$

$$\text{softplus loss} = \log(1 + \exp(\cdots)) = \log\left(\frac{\exp(-\cdots)}{1 + \exp(-\cdots)}\right) = \left(\log \text{softmax}\begin{bmatrix}\cdots \\ 0\end{bmatrix}\right)$$

D. deep architect

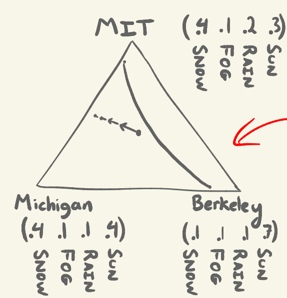
Softmax is... a physics-inspired final layer for deep nets

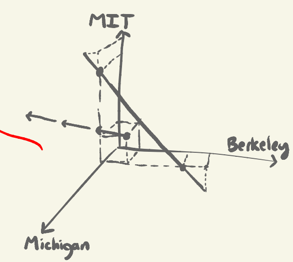
<START> FOUR SCORE AND ? ?



E. researcher

Softmax is ... the canonical e-flat parameterization of the simplex





MIT Mich Berk
(.33 .33 .33)
observe Snow ×(.4 .4 .1)
(.13 .13 .03)
(.44 .44 .11)
observe Snow ×(.4 .4 .1)
(.17 .17 .01)
(.48 .48 .03)
observe Snow ×(.4 .4 .1)
(.19 .19 .003)
(.496 .496 .008)