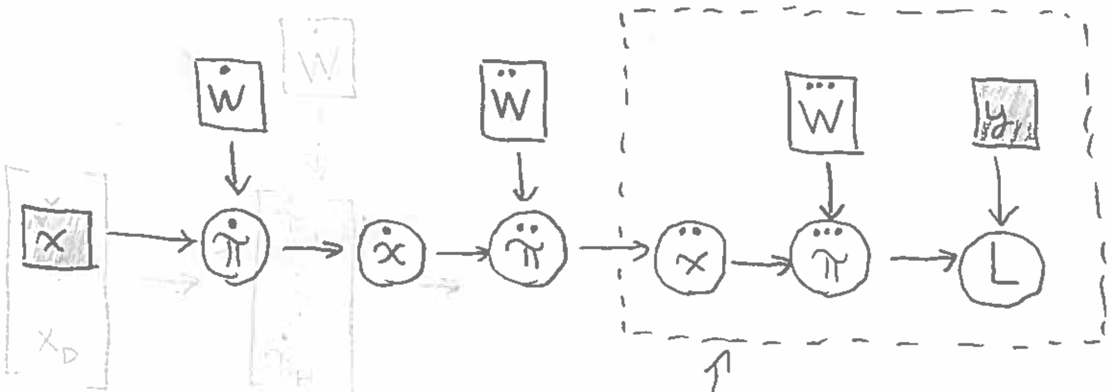
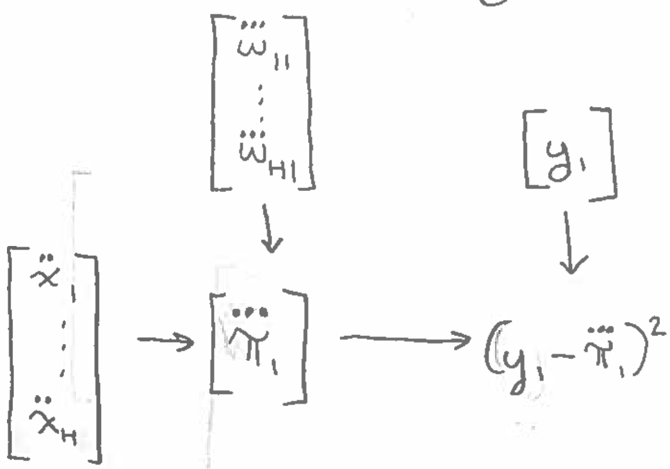
MULTIWAY CLASSIFICATION

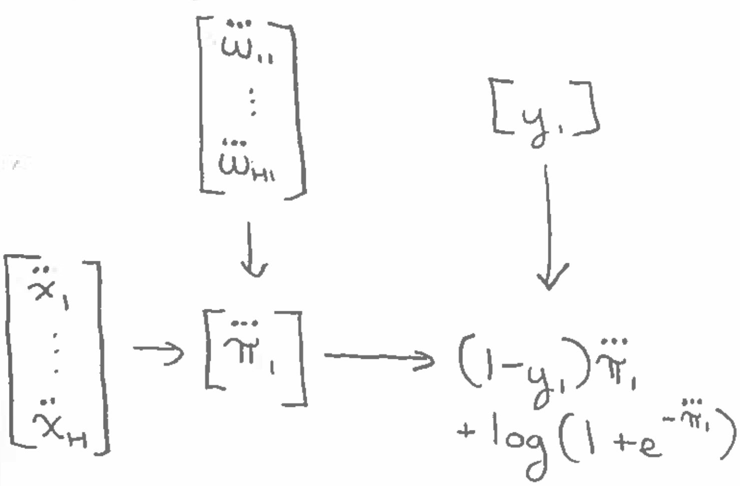
① Recall the architecture for a feedforward neural network (here, using $M=3$ layers):



② We've seen so far a couple different instances of this "output layer":



(linear regression)



(logistic regression)

Multiway Classification

③ In each, we produce a single scalar $\dddot{\pi}_i$ and compare it to scalar response y_i to compute loss L.

e.g.

$$\begin{bmatrix} 1 \\ 24 \\ 150 \end{bmatrix} \xrightarrow{\begin{bmatrix} -50 \\ 2 \\ 1 \end{bmatrix}} [148] \longrightarrow \underset{\uparrow\,[182]}{(182-148)^2}$$

(linear regression)

$$\begin{bmatrix} 1 \\ 24 \\ 150 \end{bmatrix} \xrightarrow{\begin{bmatrix} -150 \\ 0.1 \\ 1 \end{bmatrix}} [2.4] \longrightarrow \underset{\uparrow\,[1]}{\log\left(1+e^{-2.4}\right)}$$

(logistic regression)

---

④ These output layers address two distinct tasks:

- regression: the output variable is a real number (e.g. cholesterol level)
- classification: the output variable is a Boolean value (e.g. whether you have high cholesterol or not)

MULTIWAY CLASSIFICATION

(5) But sometimes you want to classify something into one of several discrete categories. For instance, given an image of an animal, we might want to automatically identify whether it is a horse, a zebra, or a panda.

- multiway classification: the output variable is a member of a finite, unordered set (e.g. {horse, zebra, panda}).

(6) How do we represent the response variable $y^{(n)}$ for a multiway classification task? Strings aren't particularly convenient:

X (evidence vector)	y (response)
$x^{(1)}$	"horse"
$x^{(2)}$	"zebra"
$x^{(3)}$	"zebra"
$x^{(4)}$	"panda"
$x^{(5)}$	"horse"

because how do we compare a string with our output vector $\dddot{\pi}$?

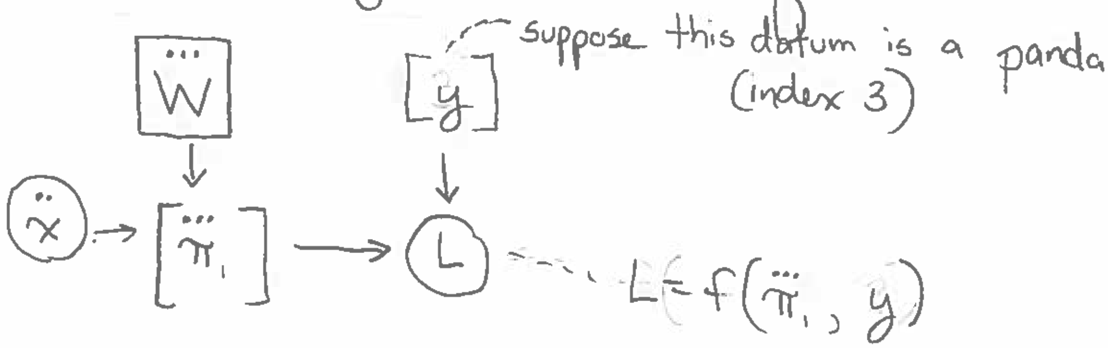
MULTIWAY CLASSIFICATION

⑦ We <u>could</u> represent each response as its index in an ordered list of the possible categories, e.g. for

$\langle \overset{1}{\text{horse}}, \overset{2}{\text{zebra}}, \overset{3}{\text{panda}} \rangle$:

X (evidence vector)	y (response)
$x^{(1)}$	1
$x^{(2)}$	2
$x^{(3)}$	2
$x^{(4)}$	3
$x^{(5)}$	1

⑧ If we do this, then we end up comparing numbers to numbers, but in a way that's kind of weird.



The loss needs to be some differentiable function of response y and output $\dddot{\pi}_1$. If this subject is a panda, then we want to reward "panda predictions" $\dddot{\pi}_1$. Let's say we just use the simple loss:

$$L = (y - \dddot{\pi}_1)^2$$

MULTIWAY CLASSIFICATION

(9) That means we want $\hat{\pi}_1$ to be close to 3 for panda images. That's ok, but it doesn't penalize horses and zebras equally. For a zebra, the loss is:

$$L = (3-2)^2 = 1$$

while for a horse, the loss is:

$$L = (3-1)^2 = 4$$

---

(10) Essentially, if we use the index of an arbitrarily ordered list to represent our response, we are implicitly saying that neighbors in the list (e.g. panda, zebra) are "closer" than elements that are further apart (e.g. panda and horse).

It's hard to imagine a loss function L that doesn't impose this bias, if we need L to be differentiable.

MULTIWAY CLASSIFICATION

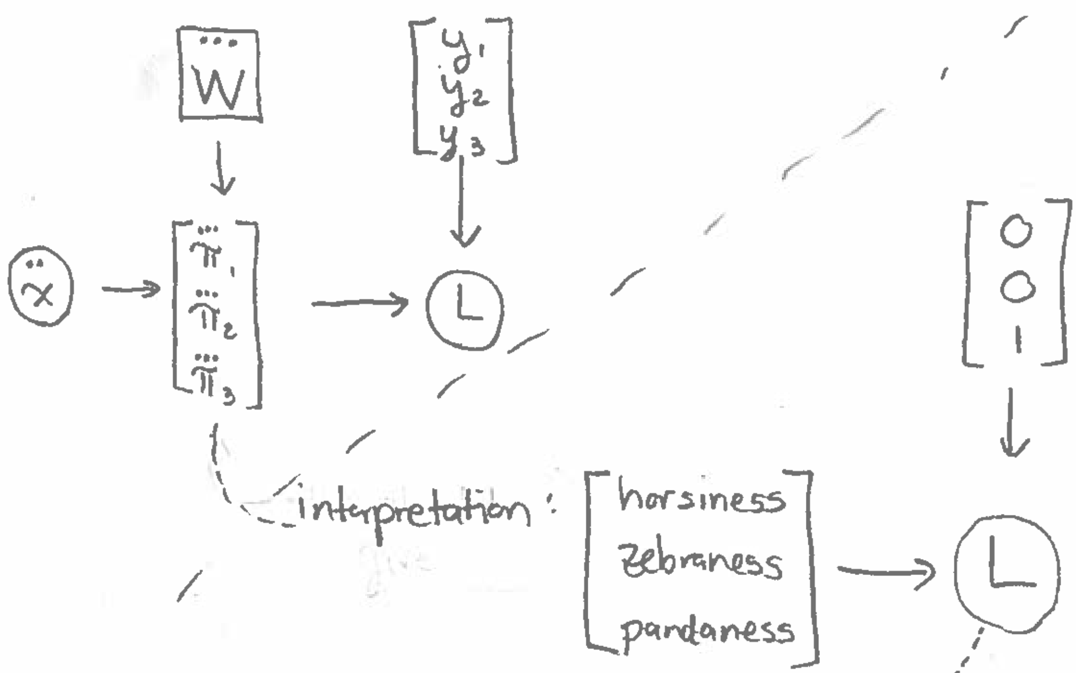
11) Back to the drawing board. How else could we represent the response? What if, again we provide an ordered list of the categories (e.g. <horse, zebra, panda>), but now we represent the response as a <u>vector</u> whose kth element is equal to 1 if the response is the k^{th} element of the list?

X (evidence vector)	y (response)	
$x^{(1)}$	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	← "horse" vector
$x^{(2)}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	← "zebra" vector
$x^{(3)}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	← "zebra" vector
$x^{(4)}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$	← "panda" vector
$x^{(5)}$	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	← "horse" vector

These are referred to as "one-hot" vectors.

Multiway Classification

(12) If we go this route, then we'd want our output $\dddot{\pi}$ to be a vector as well:



loss L should penalize outputs with a high horsiness or zebraness, and reward outputs with high pandaness

(13) One straightforward implementation of this intuition is to take the output vector, replace its max value with 1, and replace the other values w. 0:

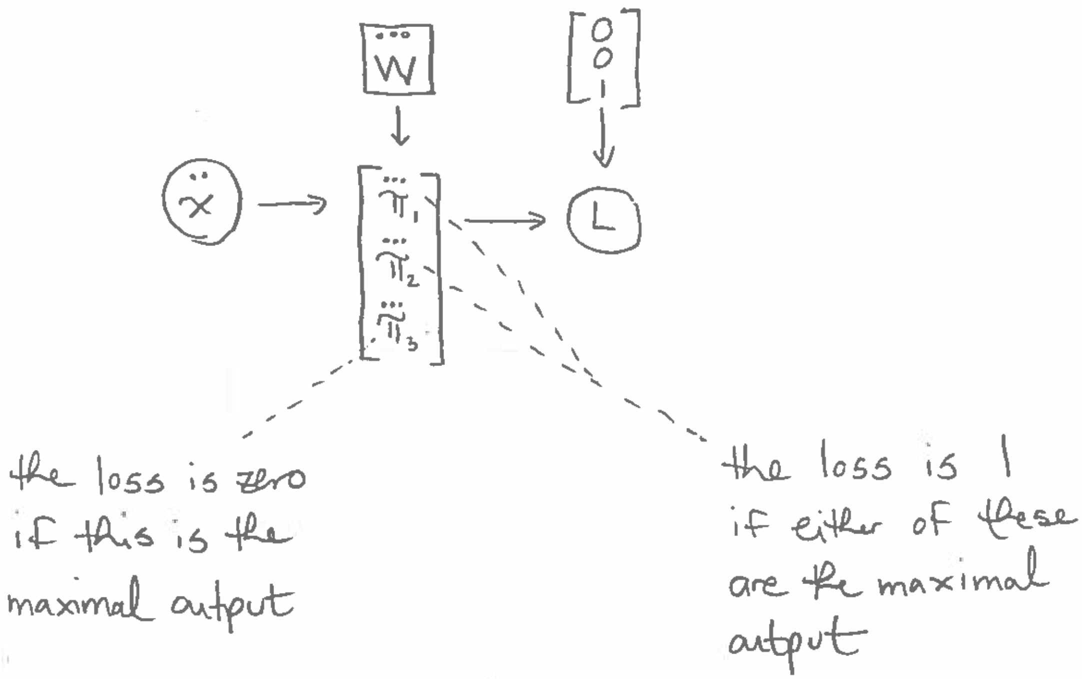
e.g. $\begin{bmatrix} 2.4 \\ 4.2 \\ 1.0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ $\qquad$ $\begin{bmatrix} 1.2 \\ -2.5 \\ 1.5 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

Then, take the dot product of the resulting vector with the response:

e.g. $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 0$ $\qquad$ $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 1$

Multiway Classification

(14) This gives us a reward of 1 if the maximal value that is output by the neural network coincides with the "true" category (and a reward of zero otherwise). To convert this into a loss function, we can just take 1 minus the reward.



(15) We can formalize this by defining onehot(k, d) to be the d-dimensional one-hot vector whose k^{th} element is 1, e.g. onehot$(2,3) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, onehot$(1,3) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$.

Then we define the loss as:

$$L = 1 - y^T \cdot \text{onehot}(\operatorname*{argmax}_i \dddot{\pi}_i, C)$$

where C is the number of categories (e.g. $C=3$ in our running example).

MULTIWAY CLASSIFICATION

(16) There's only one problem. This loss function isn't differentiable. We can't compute $\frac{\partial L}{\partial \ddot{\pi}}$, so we can't compute $\frac{\partial L}{\partial \theta}$, and thus we can't use gradient descent to optimize the weights θ.

(17) But can we find an alternative loss function that is similar in spirit, but which is differentiable?

First, observe that our "hard max" function is essentially mapping the vector $\ddot{\pi}$ to a probability distribution:

$$\begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix} \longrightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

only thing is — all the probability mass is concentrated on one value.

So one idea would be to map $\ddot{\pi}$ to a probability distribution for which most of the probability mass is concentrated on one value, e.g.

$$\begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix} \longrightarrow \begin{bmatrix} .130 \\ .003 \\ .867 \end{bmatrix}$$

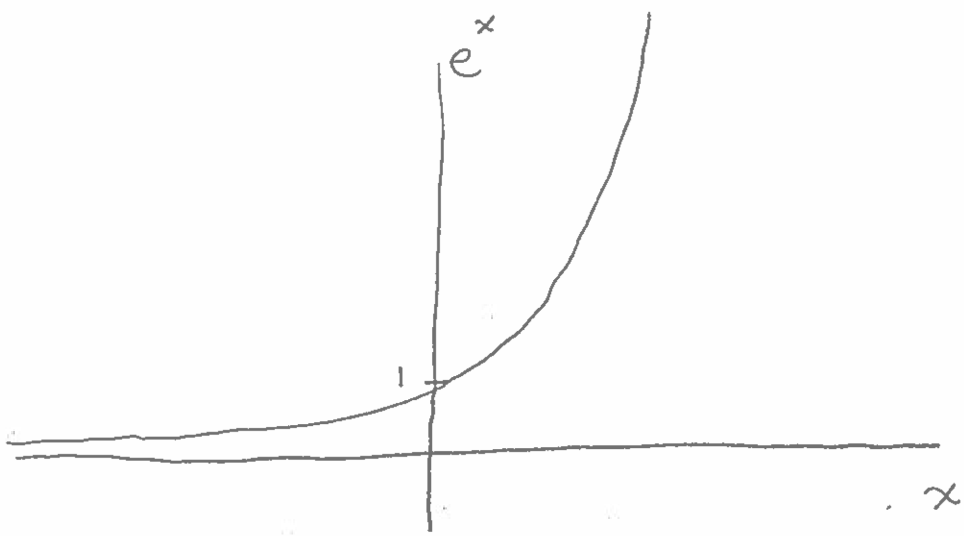
MULTIWAY CLASSIFICATION

(18) How do we map a vector of reals into a probability distribution such that most of the probability mass is concentrated on the highest original value?

$$\begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix}$$

we want most of the probability mass to be concentrated on the maximal original value

(19) One cool trick is to notice that the exponential function maps the real numbers to strictly positive numbers:



It also does so in a way that magnifies the differences between the original numbers.

MULTIWAY CLASSIFICATION

(20) For instance, applying the exponential function; we get:

$$\begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix} \xrightarrow{\text{exponentiate}} \begin{bmatrix} 3.32 \\ 0.08 \\ 22.2 \end{bmatrix}$$

Now that we have a vector of positive numbers, we can simply normalize to get a probability distribution.

$$\begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix} \xrightarrow{\text{exponentiate}} \begin{bmatrix} 3.32 \\ 0.08 \\ 22.2 \end{bmatrix} \xrightarrow{\text{normalize}} \begin{bmatrix} .130 \\ .003 \\ .867 \end{bmatrix}$$

(21) This function is referred to as <u>softmax</u>:

$$\text{softmax}\left(\begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix}\right) = \begin{bmatrix} \dfrac{e^{z_1}}{\sum_n e^{z_n}} \\ \vdots \\ \dfrac{e^{z_N}}{\sum_n e^{z_n}} \end{bmatrix}$$

MULTIWAY CLASSIFICATION

(22) Retrofitting our loss function from (15) to use <u>softmax</u> instead of <u>hardmax</u>, we get:

$$L = 1 - y^T \cdot \text{softmax}(\dddot{\pi})$$

(23) Examples:

→ if output $\dddot{\pi} = \begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix}$ and response $y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

then: $$L = 1 - \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} .130 \\ .003 \\ .867 \end{bmatrix} = .133$$

→ if output $\dddot{\pi} = \begin{bmatrix} 1.2 \\ -2.5 \\ 3.1 \end{bmatrix}$ and response $y = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

then: $$L = 1 - \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} .130 \\ .003 \\ .867 \end{bmatrix} = .997$$

In other words, the loss is the total probability mass accorded to incorrect responses.

MULTIWAY CLASSIFICATION

(24) Of course, this is all for nothing if the loss isn't differentiable. And it is:

$$\frac{\partial L}{\partial \dddot{\pi}_i} = \frac{\partial}{\partial \dddot{\pi}_i}\left(1 - y^T \cdot \text{softmax}(\dddot{\pi})\right)$$

$$= \frac{\partial}{\partial \dddot{\pi}_i}\left(-y^T \cdot \text{softmax}(\dddot{\pi})\right)$$

$$= \frac{\partial}{\partial \dddot{\pi}_i} - y^T \cdot \begin{bmatrix} \dfrac{e^{\dddot{\pi}_1}}{\sum_n e^{\dddot{\pi}_n}} \\ \vdots \\ \dfrac{e^{\dddot{\pi}_N}}{\sum_n e^{\dddot{\pi}_n}} \end{bmatrix}$$

$$= \frac{\partial}{\partial \dddot{\pi}_i}\left(\frac{-1}{\sum_n e^{\dddot{\pi}_n}}\right) y^T \cdot \begin{bmatrix} e^{\dddot{\pi}_1} \\ \vdots \\ e^{\dddot{\pi}_N} \end{bmatrix}$$

$$= y^T \cdot \begin{bmatrix} e^{\dddot{\pi}_1} \\ \vdots \\ e^{\dddot{\pi}_N} \end{bmatrix} \frac{\partial}{\partial \dddot{\pi}_i}\left(\frac{1}{\sum_n e^{\dddot{\pi}_n}}\right) + \frac{1}{\sum_n e^{\dddot{\pi}_n}} \frac{\partial}{\partial \dddot{\pi}_i}\left(y^T \begin{bmatrix} e^{\dddot{\pi}_1} \\ \vdots \\ e^{\dddot{\pi}_N} \end{bmatrix}\right)$$

$$= y^T \cdot \begin{bmatrix} e^{\dddot{\pi}_1} \\ \vdots \\ e^{\dddot{\pi}_N} \end{bmatrix} \left(\frac{-e^{\dddot{\pi}_i}}{\left(\sum_n e^{\dddot{\pi}_n}\right)^2}\right) + \frac{y_i}{\sum_n e^{\dddot{\pi}_n}}$$