Multinomial Logistic Regression

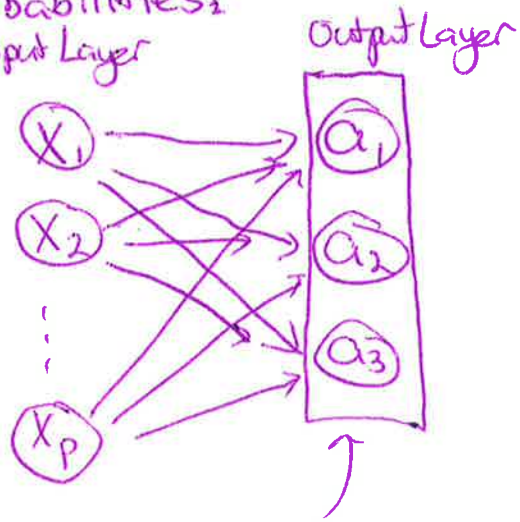
$y^{(i)}$ is categorical with K classes

Ex: Picture is of a dog, a cat, or a bird. K = 3 classes

"One-hot" encoding or "indicator variable":

$$y^{(i)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$ if observation # i is of a cat (second class)

Our network needs to generate a vector of class probabilities:



- each of a_1, a_2, a_3 is between 0 and 1
- $a_1 + a_2 + a_3 = 1$

2 steps:

1) compute $\underset{\sim}{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} b_1 + w_1^T x \\ b_2 + w_2^T x \\ b_3 + w_3^T x \end{bmatrix}$

$$= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} w_1^T \\ w_2^T \\ w_3^T \end{bmatrix} \underset{\sim}{x}$$

w_3^T: a row vector

$$\underset{\sim}{z} = \underset{\sim}{b} + W^T \underset{\sim}{x}$$

2) Compute $\underset{\sim}{a} = g(\underset{\sim}{z})$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = g\left(\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}\right)$$

Activation function $g(\underset{\sim}{z})$ is <u>softmax</u>:

$$\text{softmax}\left(\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}\right) = \begin{bmatrix} \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \\ e^{z_2}/(e^{z_1} + e^{z_2} + e^{z_3}) \\ e^{z_3}/(e^{z_1} + e^{z_2} + e^{z_3}) \end{bmatrix}$$

← each entry is positive ($e^c > 0$)

each entry is < 1

$\left(\frac{e^c}{e^c + \text{stuff}}\right) < 1$

Multinomial Logistic Regression Example:

Suppose $y^{(i)} = \begin{cases} 1 \text{ if dog} \\ 2 \text{ if cat} \\ 3 \text{ if bird} \end{cases}$

Only one feature:

$x_1^{(i)}$ = weight of animal number i in pounds

My first observation is Benedict: $y^{(1)} = 2$, $x_1^{(1)} = 10$

Model has a b and w associated with each class.

Suppose $b_1 = -4$, $w_{11} = 2.3$

$b_2 = 0$, $w_{21} = 2$

$b_3 = 10$, $w_{31} = -5$

bias (intercept) associated with class 3

weight (slope) for first feature, class 3.

For Benedict, we have a z associated with each class:

$z_1^{(1)} = -4 + 2.3 \cdot 10 = 19$

$z_2^{(1)} = 0 + 2 \cdot 10 = 20$

$z_3^{(1)} = 10 - 5 \cdot 10 = -40$

We then have an a for each class:

$$\begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \end{bmatrix} = \text{softmax}\left(\begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ z_3^{(1)} \end{bmatrix}\right) = \begin{bmatrix} \frac{e^{19}}{e^{19}+e^{20}+e^{-40}} \\ \frac{e^{20}}{e^{19}+e^{20}+e^{-40}} \\ \frac{e^{-40}}{e^{19}+e^{20}+e^{-40}} \end{bmatrix}$$

178,483,301 $\approx 1.78 \times 10^8$

4.85×10^8

4.25×10^{-18}

$$= \begin{bmatrix} \frac{178,482,301}{178,483,301 + 485,165,195 + 4.2 \times 10^{-18}} \\ \frac{4.85 \times 10^8}{1.78 \times 10^8 + 4.85 \times 10^8 + 4.25 \times 10^{-18}} \\ \frac{4.25 \times 10^{-18}}{1.78 \times 10^8 + 4.85 \times 10^8 + 4.25 \times 10^{-18}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.269 \\ 0.731 \\ 6.4 \times 10^{-27} \end{bmatrix}$$

model says, probability 0.731 that Benedict is a cat

Since Benedict is a cat, $y^{(1)} = 2$

$$\underbrace{a_{y^{(1)}}^{(1)} = a_2^{(1)} = 0.731}_{\text{contribution to the likelihood function from observation 1 in my data set}}$$