Linear Regression as a Neural Network

Linear Regression Model:

$$Y^{(i)} = b + w_1 x_1^{(i)} + \cdots + w_p x_p^{(i)} + \varepsilon^{(i)}$$

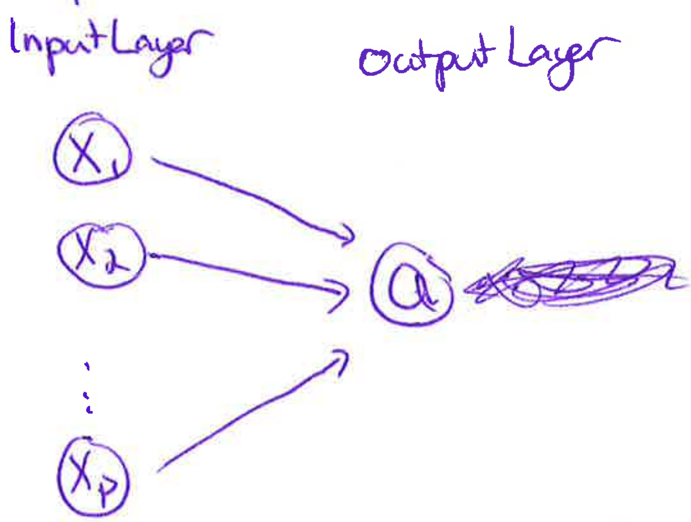
$$\varepsilon^{(i)} \sim \text{Normal}(0, \sigma^2)$$

... or ...

$$Y^{(i)} \sim \text{Normal}(b + w_1 x_1^{(i)} + \cdots + w_p x_p^{(i)}, \sigma^2)$$

... or ...

$$Y^{(i)} \sim \text{Normal}(b + w^T x^{(i)}, \sigma^2)$$



the circle for a represents 2 steps:

1) calculate $z^{(i)} = b + w^T x^{(i)}$

2) calculate $a^{(i)} = g(z^{(i)})$

for linear regression we use an <u>identity activation</u> $g(z) = z$

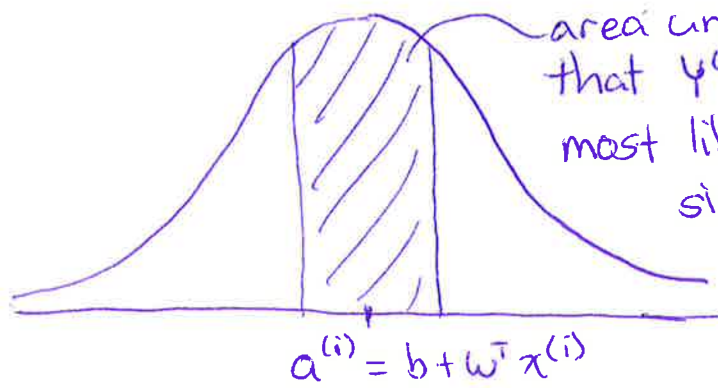
so $a^{(i)} = g(z^{(i)}) = z^{(i)} = b + w^T x^{(i)}$

also sometimes referred to as a linear activation

So $Y^{(i)} \sim \text{Normal}(a^{(i)}, \sigma^2)$

Likelihood for regression:

$$y^{(i)} \sim \text{Normal}(b + w^T x^{(i)}, \sigma^2)$$



area under curve gives probability that $y^{(i)}$ will be in that region

most likely to get $y^{(i)}$ near $a^{(i)}$, since curve is highest there.

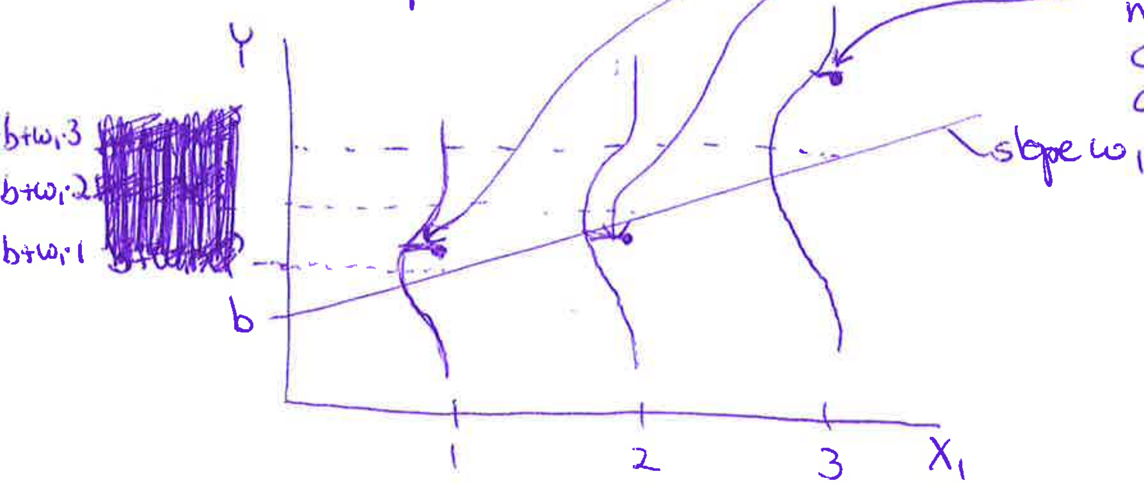
↑ this axis represents values of $y^{(i)}$

The formula for this curve is

$$f(y^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y^{(i)} - a^{(i)})^2}$$

keep (i)

Ex. with $p=1$ feature:



the higher the normal distribution curve is at each observed response value, the better the fit of the line to our data.

Likelihood function is

$$\mathcal{L}(b, w) = \prod_{i=1}^{m} f(y^{(i)})$$

similar intuition to $\prod_{i=1}^{m} P(Y^{(i)} = y^{(i)})$, <u>but</u> $f(y^{(i)})$ is not $P(Y^{(i)} = y^{(i)})$

We choose b and w that maximize $\mathcal{L}(b, w)$

Log-likelihood for regression

Equivalently, we choose b, w to maximize the log-likelihood

$$\ell(b, w) = \log\{\mathcal{L}(b, w)\}$$

... or ....

minimize the negative log-likelihood cost function

$$J(b, w) = -\log\{\mathcal{L}(b, w)\}$$

$$= -\log\left\{\prod_{i=1}^{m} f(y^{(i)})\right\}$$

$$= -\log\left\{\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y^{(i)} - a^{(i)})^2}\right\}$$

$$= -\sum_{i=1}^{m} \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y^{(i)} - a^{(i)})^2}\right\}$$

$$= -\sum_{i=1}^{m}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y^{(i)} - a^{(i)})^2\right]$$

$$= -m \cdot \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^{(i)} - a^{(i)})^2$$

$$= -m \cdot \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left\{y^{(i)} - (b + w^T x^{(i)})\right\}^2$$

Minimizing the above is equivalent to minimizing

$$RSS = \sum_{i=1}^{m}\left\{y^{(i)} - (b + w_1 x_1^{(i)} + \dots + w_p x_p^{(i)})\right\}^2$$