b) $\frac{1}{1+e^{-(x_1w_1+x_2w_2+b)}}$

$\frac{d\sigma}{dw_1} =$

$$\frac{(1+e^{-x_1w_1-x_2w_2-b})(0)-(1)(-e^{-x_1w_1-x_2w_2-b})}{(1+e^{-x_1w_1-x_2w_2-b})^2}$$

$$\frac{+e^{-x_1w_1-x_2w_2-b}}{(1+e^{-x_1w_1-x_2w_2-b})^2}$$

$$\frac{x_1 e^{-x_1w_1-x_2w_2-b}}{(1+e^{-x_1w_1-x_2w_2-b})^2}$$

c) $\frac{1}{1+e^{-(x_1w_1+x_2w_2+b)}}$

$$\frac{d\sigma}{db} = \frac{e^{-x_1w_1-x_2w_2-b}}{(1+e^{-x_1w_1-x_2w_2-b})^2}$$

$$= \frac{e^{x_1w_1+x_2w_2+b}}{(1+e^{x_1w_1+x_2w_2+b})^2}$$

6) G) Mario, Aleister, Eris
- 15 hours - 20 hours
- 3.5 - 4
- 4 - 5

1) a) My sister and her boyfriend went on a break.
- either they broke up, or they went on a trip.

b) I ~~saw~~ saw the old woman with the glasses.

c) Sasha and Sophie played cards.

2) Statement 1: No, because for each independently pulled set of data for each model run, the error rate will not be the same across ~~the~~ every run.

Statement 2: Yes, because the tuned model, trained on T and tuned on V, will be a more accurate, in theory, representation of what the data look like. By training on one set then tuning on V, the data are not overfit to the T set but should be a generally good predictor of V.

Statement 3: yes, because if the max depth is chosen randomly then the ~~decision tree won't~~ error rate won't be biased.

3) a) chain rule:

$$\frac{d\sigma}{dz} = \frac{d}{dz}\left[\frac{1}{1+e^{-z}}\right]$$

$$= \frac{d}{dz}(1+e^{-z})^{-1}$$

$$= -(1+e^{-z})^{-2}(-e^{-z})$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-z}} \cdot \frac{(1+e^{-z})-1}{1+e^{-z}}$$

$$= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= \sigma(z) \cdot (1-\sigma(z))$$