Excercise 2: Why use quadratic loss?

$$\hat{w} = \underset{w}{\operatorname{argmax}}\, L(X, Y, w) = \underset{w}{\operatorname{argmax}}\, \mathbb{P}[y_1 = \hat{y} = (x_i; w), ..., y_n = \hat{y}(x_n; w)|w]$$

$$= \prod_{i=1}^{n} \mathbb{P}[y_i = \hat{y}(x_i; w)|w] = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i - \hat{y}(x_i;w))^2}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i - \hat{y}(x_i;w))^2} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}(x_i; w))^2$$

Taking a partial derivative with respect to w, which is equal to least squares, results in :

$$\frac{\partial \ell(w, \sigma^2)}{\partial w} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial w}((y - x_w)^T(y - x_w)) = -\frac{1}{2\sigma^2}[2X^T X w - 2X^T y]$$

Equating this term to 0 gives:

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y = \hat{w}_{ML}$$

This proves, that minimizing the square loss is equal to maximizing the likelihood.