Consider stochastic gradient descent to minimize the error $E = e(f(x;w) - y)$ on a single training example $x$. Here $w$ is the parameter vector of the model $f$ and $y$ is the true label of the training example. Work out the stochastic gradient update rule as specifically as possible, when the error function $e$ is quadratic: $e(z) = z^2$. Hint: Use the notation $e'(z)$ for the derivative $\frac{d}{dz}e(z)$.

**Answer.** For each component $w_j$ of the parameter vector $w$, the update rule is

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} E$$

where $\alpha$ is the learning rate. By the chain rule of derivatives,

$$\frac{\partial}{\partial w_j} E = e'(f(x;w) - y) \frac{\partial}{\partial w_j} f$$

If $e(z) = z^2$ then $e'(z) = 2z$ so

$$w_j := w_j - \alpha(2(f(x;w) - y)) \frac{\partial}{\partial w_j} f.$$
Additional comments. Intuitively, if the predicted label value \( f(x; w) \) is larger than the true label value \( y \), and increasing \( w_j \) makes \( f \) increase, then \( w_j \) should be decreased.

This question above is very similar to the question in a 2011 quiz in the online lecture notes. It is worth it to study these closely.