1. Regression with one predictor variable

(a) We will predict the mean of the $y$-values: $\hat{y} = (1 + 3 + 4 + 6)/4 = 3.5$. The MSE of this prediction is exactly the variance of the $y$-values, namely:

$$
\text{MSE} = \frac{(1 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (6 - 3.5)^2}{4} = 3.25.
$$

(b) If we simply predict $x$, the MSE is

$$
\frac{1}{4} \sum_{i=1}^{4} (y^{(i)} - x^{(i)})^2 = \frac{1}{4} ((1 - 1)^2 + (1 - 3)^2 + (4 - 4)^2 + (4 - 6)^2) = 2.
$$

(c) We saw in class that the MSE is minimized by choosing

$$
a = \frac{\sum_i (y^{(i)} - \bar{y})(x^{(i)} - \bar{x})}{\sum_i (x^{(i)} - \bar{x})^2},
b = \bar{y} - ax
$$

where $\bar{x}$ and $\bar{y}$ are the mean values of $x$ and $y$, respectively. This works out to $a = 1, b = 1$; and thus the prediction on $x$ is simply $x + 1$. The MSE of this predictor is:

$$
\frac{1}{4} (1^2 + 1^2 + 1^2 + 1^2) = 1.
$$

2. Lines through the origin

(a) The loss function is

$$
L(a) = \sum_{i=1}^{n} (y^{(i)} - ax^{(i)})^2
$$

(b) The derivative of this function is:

$$
\frac{dL}{da} = -2 \sum_{i=1}^{n} (y^{(i)} - ax^{(i)})x^{(i)}.
$$

Setting this to zero yields

$$
a = \frac{\sum_{i=1}^{n} x^{(i)}y^{(i)}}{\sum_{i=1}^{n} x^{(i)}^2}.
$$

3. (a) The best predictor is $\hat{y} = x_1 + x_2 + x_3 + x_4 + x_5 + 5$: to minimize the fluctuations due to $x_6 + \cdots + x_{10}$, we use its mean.

(b) All errors come from the variance in $x_6 + \cdots + x_{10}$, so

$$
\text{MSE} = \text{var}(x_6 + \cdots + x_{10}) = \text{var}(x_6) + \cdots + \text{var}(x_{10}) = 5.
$$

4. The loss induced by a linear predictor $w \cdot x + b$ is

$$
L(w, b) = \sum_{i=1}^{n} |y^{(i)} - (w \cdot x^{(i)} + b)|.
$$
5. Define

\[ X = \begin{bmatrix}
  x^{(1)} \\
  x^{(2)} \\
  \vdots \\
  x^{(n)}
\end{bmatrix} \]

\[ XX^T = \begin{bmatrix}
  x^{(1)} \cdot x^{(1)} & x^{(1)} \cdot x^{(2)} & \cdots & x^{(1)} \cdot x^{(n)} \\
  x^{(2)} \cdot x^{(1)} & x^{(2)} \cdot x^{(2)} & \cdots & x^{(2)} \cdot x^{(n)} \\
  \vdots & \vdots & \ddots & \vdots \\
  x^{(n)} \cdot x^{(1)} & x^{(n)} \cdot x^{(2)} & \cdots & x^{(n)} \cdot x^{(n)}
\end{bmatrix} \]

6. Discovering relevant features in regression.

(a) A sensible strategy is to do linear regression using the Lasso, and to choose a regularization constant \( \lambda \) that yields roughly 10 non-zero coefficients.

(b) The smallest value of \( \lambda \) we tried that gave nonzero coefficients for 10 features is 0.4. This yielded the following features (numbering starting at 1): 2, 3, 5, 7, 11, 13, 17, 19, 23, 29.

7. Logistic regression. Since

\[ \Pr(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}}, \]

we can rearrange terms to get

\[ w \cdot x + b = \ln \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)} \]

(a) \( w \cdot x + b = \ln 1 = 0 \)

(b) \( w \cdot x + b = \ln 3 \)

(c) \( w \cdot x + b = -\ln 3 \)

8. With vocabulary \( V = \{ \text{is, flower, rose, a, an} \} \), the bag-of-words representation of the sentence “a rose is a rose is a rose” is (2, 0, 3, 3, 0).

9. We want to find the \( z \in \mathbb{R}^d \) that minimizes

\[ L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (x^{(i)}_j - z_j)^2. \]

Taking partial derivatives, we have

\[ \frac{\partial L}{\partial z_j} = \sum_{i=1}^{n} -2(x^{(i)}_j - z_j) = 2nz_j - 2\sum_{i=1}^{n} x^{(i)}_j. \]

Thus

\[ \nabla L(z) = 2nz - 2\sum_{i=1}^{n} x^{(i)}. \]

Setting \( \nabla L(z) = 0 \) and solving for \( z \), gives us

\[ z^* = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}. \]

10. \( L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4 \)
(a) The derivative is
\[ \nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4) \]

(b) The derivative at \( w = (0, 0, 0, 0) \) is \( (2, -4, 0, 0) \). Thus the update at this point is:
\[ w_{\text{new}} = w - \eta \nabla L(w) = (0, 0, 0, 0) - \eta(2, -4, 0, 0) = (-2\eta, 4\eta, 0, 0). \]

(c) To find the minimum value of \( L(w) \), we will equate \( \nabla L(w) \) to zero:
- \( 2w_1 + 2 = 0 \implies w_1 = -1 \)
- \( 4w_2 - 4 = 0 \implies w_2 = 1 \)
- \( 2w_3 - 2w_4 = 0 \implies w_3 = w_4 \)

The function is minimized at any point of the form \((-1, 1, x, x)\).

(d) No, there is not a unique solution.

11. We are interested in analyzing
\[ L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 + \lambda\|w\|^2. \]

(a) To compute \( \nabla L(w) \), we compute partial derivatives.
\[ \frac{\partial L}{\partial w_j} = \left( \sum_{i=1}^{n} -2x_j^{(i)}(y^{(i)} - w \cdot x^{(i)}) \right) + 2\lambda w_j \]

Thus
\[ \nabla L(w) = -2 \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})x^{(i)} + 2\lambda w. \]

(b) The update for gradient descent with step size \( \eta \) looks like
\[ w_{t+1} = w_t - \eta \nabla L(w_t) \]
\[ = w_t(1 - 2\eta \lambda) + 2\eta \sum_{i=1}^{n} (y^{(i)} - w_t \cdot x^{(i)})x^{(i)} \]

(c) The update for stochastic gradient descent looks like the following.
\[ w_{t+1} = w_t(1 - 2\eta \lambda) + 2\eta(y^{(i_t)} - w_t \cdot x^{(i_t)})x^{(i_t)} \]

where \( i_t \) is the index chosen at time \( t \).