Linear Regression: Analytic Solution *(find and work with a partner)*

In the case of *multiple* linear regression, each example has p features. We still have an associated response (output) variable y for each example. We typically add a "fake 1" to the features of each example, so that $\vec{x} = [1\ x_1\ x_2 \cdots x_p]^T$. Our model is now the dot product between the weight vector $\vec{w}$ and the features:

$$h_{\vec{w}}(\vec{x}) = w_0 + w_1 x_1 + \cdots w_p x_p = \vec{w} \cdot \vec{x}$$

If we consider $\boldsymbol{X}$ to be the entire matrix of features (with dimensions $n \times (p+1)$) then we can reframe the linear regression problem in terms of vectors and matrices. In this case the analytic solution for the weight vector is:

$$\hat{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$$

1. *Warmup.* Given matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ below, compute both $\boldsymbol{AB}$ and $\boldsymbol{BA}$. Is matrix multiplication commutative?

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 \\ 3 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 0 & 2 \\ -4 & 1 \end{bmatrix}$$

$$\boldsymbol{AB} = \begin{bmatrix} 1(0) + (-1)(-4) & 1(2) + (-1)(1) \\ 3(0) + 2(-4) & 3(2) + 2(1) \end{bmatrix} = \boxed{\begin{bmatrix} 4 & 1 \\ -8 & 8 \end{bmatrix}},$$

$$\boldsymbol{BA} = \begin{bmatrix} 0(1) + 2(3) & 0(-1) + 2(2) \\ -4(1) + 1(3) & -4(-1) + 1(2) \end{bmatrix} = \boxed{\begin{bmatrix} 6 & 4 \\ -1 & 6 \end{bmatrix}}.$$

Since $\boldsymbol{AB} \neq \boldsymbol{BA}$, matrix multiplication is not commutative.

2. Going back to our small example from the last handout, we will use this matrix/vector form to verify our solution for $\vec{w}$ (even though $p = 1$ we can still use the multiple linear regression derivation). We again have the data: $(x_1, y_1) = (1, 0)$ and $(x_2, y_2) = (0, 1)$. First, rewrite $\boldsymbol{X}$ with the added column of 1's. Then write $\vec{y}$ as a vector.

Each row of $\boldsymbol{X}$ is $[1 \quad x_i]$, so

$$\boxed{\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}}.$$

3. Now use the matrix/vector analytic solution to verify the weight vector $\vec{w}$.

First compute

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}, \qquad \boldsymbol{X}^T\vec{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Therefore,

$$\hat{\vec{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}.$$

Thus $w_0 = 1$, $w_1 = -1$, and the fitted line is $\boxed{\hat{y} = 1 - x}$.

Linear Regression: SGD solution *(find and work with a partner)*

In linear regression, we seek to minimize the sum of squared errors between the actual response and our prediction. We often call this RSS (residual sum of squares) or SSE (sum of squared errors). As an objective function, we often call it J and include a $\frac{1}{2}$ in front to make the derivatives work out nicely.

$$J(\vec{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(h_{\vec{w}}(\vec{x}_i) - y_i\right)^2$$

For linear regression in general, one iteration of stochastic gradient descent includes the following updates (usually with the data points shuffled):

```
for i = 1,2,...,n:
```
$$\vec{w} \leftarrow \vec{w} - \alpha(\vec{w}\cdot\vec{x}_i - y_i)\vec{x}_i$$

We will begin with our same data from the previous two handouts: $(x_1, y_1) = (0, 1)$ and $(x_2, y_2) = (1, 0)$, except we will reverse the order of the points to make the progress of gradient descent a bit clearer. So in this case our matrix/vector formulation is:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad \vec{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

1. Assuming $\alpha = 0.1$ and our initial values are $w_0 = 0$ and $w_1 = 0$, what are w_0 and w_1 after the just the first data point is used to update the gradient?

   For the first example,

   $$\vec{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad y_1 = 1, \qquad \vec{w}\cdot\vec{x}_1 = 0.$$

   The update is

   $$\vec{w} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1(0-1)\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \boxed{\begin{bmatrix} 0.1 \\ 0 \end{bmatrix}}.$$

   Thus $w_0 = 0.1$ and $w_1 = 0$.

2. What are w_0 and w_1 after the second data point is used? Since we only have two examples here, your result would be the weight vector after the first iteration of SGD.

   For the second example,

   $$\vec{x}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad y_2 = 0, \qquad \vec{w}\cdot\vec{x}_2 = 0.1.$$

   Therefore,

   $$\begin{aligned}\vec{w} &\leftarrow \begin{bmatrix} 0.1 \\ 0 \end{bmatrix} - 0.1(0.1-0)\begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.01 \\ 0.01 \end{bmatrix} = \boxed{\begin{bmatrix} 0.09 \\ -0.01 \end{bmatrix}}.\end{aligned}$$

3. What is the value of the objective function (cost) after this initial iteration?

   The model is now
   $$h_{\vec{w}}(x) = 0.09 - 0.01x.$$
   Its two residuals are
   $$h_{\vec{w}}(0) - 1 = -0.91, \qquad h_{\vec{w}}(1) - 0 = 0.08.$$
   Hence,
   $$J(\vec{w}) = \frac{1}{2}\left[(-0.91)^2 + (0.08)^2\right] = \frac{1}{2}(0.8281 + 0.0064) = \boxed{0.41725}.$$

4. *(optional)* Sketch the linear model after this initial iteration.

   The line to sketch is
   $$\boxed{\hat{y} = 0.09 - 0.01x}.$$
   It has a y-intercept of 0.09 and a small negative slope of -0.01.