Exercise 1

Consider a linear regression problem, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$, with squared loss. The hypothesis set is the set of *constant* functions, that is $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, where $h_a(\mathbf{x}) = a$. Let $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ denote the training set.

- Derive the hypothesis $h \in \mathcal{H}$ that minimizes the training error.
- Use the result above to explain why, for a given hypothesis $\hat{h}$ from the set of all linear models, the coefficient of determination $R^2 = 1 - \frac{\sum_{i=1}^m (\hat{h}(\mathbf{x}_i) - y_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2}$ where $\bar{y}$ is the average of the $y_i, i = 1, \ldots, m$ is a measure of how well $\hat{h}$ performs (on the training set).

$\mathcal{H}$: $h_a \in \mathcal{H}$, $a \in \mathbb{R}$

$h_a(\vec{x}) = a \quad \forall \vec{x} \in \mathcal{X}$

Solution

- Given $h_a \in \mathcal{H}$, the training error for such hypothesis:

$$L_S(h_a) = \frac{1}{m}\sum_{i=1}^{m}\left(h_a(\vec{x}_i) - y_i\right)^2$$

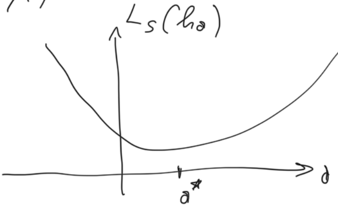
since $h_a(\vec{x}) = a$ $\forall \vec{x} \in \mathcal{X}$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(a - y_i\right)^2$$

Now, finding $h_a \in \mathcal{H}$ that minimizes the training error corresponds to find a that minimizes

$$L_S(h_a) = \frac{1}{m}\sum_{i=1}^{m}\left(a - y_i\right)^2 = (\quad)a^2 + (\quad)a + (\quad)$$

As a function of a



$\Rightarrow$ compute $\dfrac{d\,L_S(h_a)}{da}$ and derive a s.t. $\dfrac{d\,L_S(h_a)}{da} = 0$

$$\frac{d L_S(h_a)}{da} = \frac{d}{da}\left(\frac{1}{m}\sum_{i=1}^{m}(a-y_i)^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\frac{d\left((a-y_i)^2\right)}{da}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 2(a-y_i)$$

$$\frac{d\left((a-y_i)^2\right)}{da}$$

$$= 1 \cdot \frac{d\left((a-y_i)^2\right)}{d(a-y_i)}$$

$$= 1 \cdot 2(a-y_i)$$

$$\cancel{\frac{2}{m}} \sum_{i=1}^{m}(a-y_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{m}(a-y_i) = 0$$

$$\Leftrightarrow \left(\sum_{i=1}^{m} a\right) - \left(\sum_{i=1}^{m} y_i\right) = 0$$

$$\Leftrightarrow \cancel{m} \cdot a = \sum_{i=1}^{m} y_i \Big/ m = \bar{y}$$

- $R^2 = 1 - \left(\sum_{i=1}^{m} \left(\hat{h}(\vec{x}_i) - y_i\right)^2\right) \Big/ \left(\sum_{i=1}^{m} (y_i - \bar{y})^2\right)$

this is the error of $\hat{h}$ (error on the training set) relative to the error of the "best" naive predictor (which predicts a constant, without looking at $\vec{x}$)

Polynomial models

Regression problem ($\mathcal{Y} = \mathbb{R}$), $\mathcal{X} = \mathbb{R}$.

How can we use as hypothesis set $\mathcal{H}$ the set of polynomials of degree r with the machinery we have already developed?

polynomial of degree r : $w_0 \cdot 1 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots + w_{r-1} x^{r-1} + w_r x^r$

Given $x \in \mathbb{R}$, obtain vector: (feature expansion)

$$\vec{x}' = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{r-1} \\ x^r \end{bmatrix}$$

$\Longrightarrow$ the hypothesis class of linear models for $\vec{x}'$ corresponds to polynomials of degree r for x.

Given $\vec{x} \in \mathbb{R}^d$, $\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$, use the following expansion:

$$\vec{x}' = \begin{bmatrix} 1 \\ x_0 \\ x_0^2 \\ \vdots \\ x_0^r \\ x_1 \\ x_1^2 \\ \vdots \\ x_1^r \\ \vdots \\ x_d \\ x_d^2 \\ \vdots \\ x_d^r \end{bmatrix} \Rightarrow \text{use linear models for } \vec{x}'$$

Different expansion: $r=2$, $\vec{x} \in \mathbb{R}^3$ $\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

$\Rightarrow$ obtain

$$\vec{x}' = \begin{bmatrix} 1 \\ x_0 \\ x_1 \\ x_2 \\ x_0^2 \\ x_1^2 \\ x_2^2 \\ x_0 x_1 \\ x_1 x_2 \\ x_0 x_2 \end{bmatrix} \Longrightarrow \text{build linear models on } \vec{x}'$$